行政院國家科學委員會專題研究計畫　成果報告

應用知識運算技術建構財金知識入口網站之雛形系統

計畫類別：個別型計畫
計畫編號：NSC91-2416-H-002-021-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立臺灣大學工商管理學系

計畫主持人：陳文華

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 2 月 3 日

行政院國家科學委員會補助專題研究計畫 **■ 成 果 報 告**
□期中進度報告

應用知識運算技術建構財金知識入口網站之雛形系統

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 91－2416－H－002－021－
執行期間：　91 年 8 月 1 日至　92 年 7 月 31 日

計畫主持人：陳文華
計畫參與人員： 施人英、徐聖訓

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立臺灣大學工商管理學系

中　華　民　國　九十三　年　一　月　二十八　日

摘要

財金相關議題的研究向來為政府決策單位、學術界及產業界十分重視的焦點。目前研究者在財金研究議題所運用的資料主要分成兩大類，一為數值運算分析的數值資料，一為文字性描述的資料，前者多為預測模式、分類模式、關係探討等數量方法上研究，後者多為事件分析、制度規劃、現象探討等定性研究。在相關研究資料的蒐集整理上，往往須奔波於各個財經網站、資料庫、新聞媒體或是原始資料的產生單位，有時基於資料的正確性及完整性的考量，須再進一步調整處理蒐集來的原始資料，導致資料蒐集處理的程序頗為曠日費時，又未必能完整蒐集到相關資訊，往往迫使研究須作某種程度上的折衷，令研究人員較難充分發揮其專長。在研究方法上，多數採用統計計量方法，近來隨著資訊科技的進步, 人工智慧等相關的資料採掘(data mining)與文字採掘(text mining)工具逐漸廣泛應用於工程領域，並獲得相當不錯的效果，故本研究在知識運算工具的發展上，除了採行統計計量工具外，另外將引用人工智慧相關的方法，提供財金研究領域另一個可行的研究方向。是故，建置一個具備正確且完整資訊的財經資訊中心，以及多樣化的研究運算工具，並將這些資源以便利的方式提供給相關的研究人員，將是每一位金融市場研究學者的期望。

基於以上的背景, 本研究致力於運用各種知識運算(knowledge computing)工具發展財金知識入口網站。首先將針對目前財經知識入口網站的供給面進行調查研究；其次，探討運用各種知識運算工具創造財金知識、發展財金知識地圖以及財金決策支援工具雛形，以解決現有的財金問題；最後，將逐步拓展為一個整合呈現財金資料、資訊及知識的入口網站。

關鍵詞：知識管理、知識入口網站、文件分類、主題地圖、SOM, SVM

Abstract

Many organizations have knowledge repository or data warehouses; however, information or knowledge is scattered everywhere without proper use. Worse, the rapid growth of Internet accelerates the creation of unstructured and unclassified information. It has resulted in information overload. Though we can reach a flood of information through general-purpose search engines, the effort of browsing through them is quite tedious and painstaking. Since most of us are unable to utilize information effectively, there is a need for technology to solve this issue. The purpose of this research is to explore text mining and data mining techniques to address the problem. We discuss the major components of these knowledge computing techniques required for building up a professional financial knowledge portal.

In the practical implementation, the laws and regulations related to securities and futures markets were utilized to demonstrate the usage of SOM for topic categorization (topic map). We also applied support vector machines in credit rating problems and stock market indices prediction, with much better results than the traditionally statistics approaches and neural network.

Keywords: knowledge management, financial knowledge portal, document categorization, topic map, self-organizing map, support vector machines

I. 前言

財金相關議題的研究向來為政府決策單位、學術界及產業界十分重視的焦點。目前研究者在財金研究議題所運用的資料主要分成兩大類，一為數值運算分析的數值資料，一為文字性描述的資料，前者多為預測模式、分類模式、關係探討等數量方法上研究，後者多為事件分析、制度規劃、現象探討等＋定性研究。在相關研究資料的蒐集整理上，往往須奔波於各個財經網站、資料庫、新聞媒體或是原始資料的產生單位，有時基於資料的正確性及完整性的考量，須再進一步調整處理蒐集來的原始資料，導致資料蒐集處理的程序頗為曠日費時，又未必能完整蒐集到相關資訊，往往迫使研究須作某種程度上的折衷，令研究人員較難充分發揮其專長。在研究方法上，多數採用統計計量方法(Tsitsiklis & Van Roy, B., 2001)，近來隨著資訊科技的進步，人工智慧等相關的資料採掘(data mining)與文字採掘(text mining)工具逐漸廣泛應用於工程領域，並獲得相當不錯的效果，故本研究在知識運算工具的發展上，除了採行統計計量工具外，另外將引用人工智慧相關的方法(Gencay & Qi, 2001；Chen et al, 2001)，提供財金研究領域另一個可行的研究方向。是故，建置一個具備正確且完整資訊的財經資訊中心，以及多樣化的研究運算工具，並將這些資源以便利的方式提供給相關的研究人員，將是每一位金融市場研究學者的期望。

綜觀市面上現有的財經研究用途資料庫廠商，目前較為著名的僅有台灣經濟新報社(TEJ)、路透社(Reuters)、彭博 (Bloomberg)等少數幾家，但這些廠商的資料檢索費甚高，並非每位研究學者皆有能力負擔；至於其餘免費的財經網站，由於其開站目的是為了吸引一般市場投資者，故所提供的資料皆屬於即時性的短期資料，功能上著重在證券市場技術面的分析，與研究者的資料需求差異甚大。在分析工具的提供上，大多基於少數幾種統計方法，尚未能有效結合一些更適合某些研究分析議題的資料採掘與文字採掘工具。

由於財金資料內容十分繁多，為有效整合這些原始資料，並進一步產生有價值的資訊與知識，本計畫希望在自動化進行資料索引(index)、資料間之關係探討(association)、資料內容有意義的分類(categorize)、相關資訊摘要整理(summarize)、現象預測(prediction)上，藉由資料採掘與文字採掘等技術有效達成目標，開發一具高品質的財金知識入口網站，以作為學術及產業界於從事研究時可供參考的來源。

II. 研究目的

在知識經濟時代，如何善用資訊產生知識成為企業持續成長的利基。很多企業並非缺乏知識庫或資料倉儲，而是知識庫太擁擠、繁雜，以致在需要的時候無法適當地取得資料。再加上網際網路的興起，網路上龐大的、未經組織的、未經分類的及高重複性的資料特性使得資料擷取變得更加複雜。透過一般目的 (general purpose)的搜尋引擎（如 google）會搜尋到上千筆的資料。對於使用者而言，透過瀏覽超過數百萬個網頁來尋找相關的資料是一沉重的負擔，而目前已開發的搜尋系統並無法正確地滿足使用者的需求。資訊超載的情況，使得人們無法有效地進行資料搜尋，有必要利用資訊技術來尋找相關且高品質的資訊。針對上述問題，衍生出目前所面臨的主要議題：如何透過資訊技術來分析大量的文件，並將其分析結果以有效的視覺化及互動效果，來協助使用者了解其內容。

基於以上的背景，本研究致力於運用各種知識運算(knowledge computing)工具發展財金知識入口網站。首先將針對目前財經知識入口網站的供給面進行調查研究；其次，探討運

用各種知識運算工具創造財金知識、發展財金知識地圖以及財金決策支援工具雛形，以解決現有的財金問題；最後，將逐步拓展為一個整合呈現財金資料、資訊及知識的入口網站(portal)。此一 portal 可產生的研究方向，如以下幾個建議重點：

1. 建立同義字詞彙庫，解決目前類似現金流量表等這一類半結構化的報表資料無法精確取出所需數值的問題，未來可從中自動化或半自動化過濾篩選出可運用於財務分析的資料項目與數據。
2. 可針對網際網路上豐富的新聞、重大事項宣告、分析報告、研究報告等文字資料進行索引建置與產生摘要文字，建立完善的財金知識地圖，使財金知識的 visualization 更佳。
3. 目前我國內公開資訊電子資料網路申報的資料格式目前未如美國 EDGAR 系統一般針對其內容定義標準的 tag，未來可應用 text mining 技術，對資料內容自動進行結構分類(auto tagging)及建置索引(Index)，以利查詢檢索。
4. 從眾多種類的資訊中，自動發掘有價值的訊息，例如公司發佈的各種重大訊息與股價間的關係模式（事件與股價行為的探討)、董監事經理人持股變化與股價的關係等。
5. 檢測公司營運的狀況，例如重大訊息內容（不定期發佈的即時資料）與公司在年報或是公開說明書中所宣稱的營運計畫是否一致？或是偏離？
6. 勾稽公司各種申報資訊的正確性與即時性，各項資料間的關係是否一致？例如剪報資訊內已有的公司重大事件新聞，在重大資訊資料內是否也有？是否一致？
7. 發掘公司資訊揭露不實的情形，例如調降財測的次數與幅度。
8. 證券市場交易異常的監視，運用 data mining 發掘是否有內線交易的情形？股價異常的可能原因分析？
9. 運用 data mining 與 text mining 發展各種投資分析的工具（例如關係模式、股票未來價值計算等）和各種證券市場指標，成為證券市場知識中心。
10. 運用 data mining 與 text mining 建立企業財務危機預警制度。

III. 文獻探討

一、知識管理常用的資訊技術

知識管理具有高度的挑戰性，因為知識是透過動態、非結構化和通常細緻的過程存在於個人或累積在組織中，並不容易用正式訓練程序或資訊系統來傳播(Swap et al. 2001)。但知識管理真正的價值是在分享不容易文件化的見解或看法，也就是一般所謂的隱性知識(McDermott 2000)，所以知識管理不能只強調資訊技術，同時還必需兼顧知識創造、傳播與分享的環境或文化，和組織的制度、流程及策略等議題，否則會事倍功半 (Allee 1999)。雖然如此，資訊技術在知識管理上還是扮演著一個非常重要的角色(Tyndale 2002)。企業在引進知識管理資訊技術時，其做法包括建立知識庫 (knowledge repository)、專家網路(expert network)，儲存非結構化的討論文件報告、技術文件線上查詢，以及企業外部資料庫等。一般而言，常用的資訊技術如下：

1、通訊基礎建設 (architecture)：含電訊以及網路的應用建設。
2、資料倉儲：資料倉儲提供了一個電子資料的圖書館，其應包含的功能有存取管理、搜尋功能，因此它能滿足企業存取、清洗 (cleanse)、儲存大量資料及對使用者查詢快速回應 (Nemati et al. 2002)。
3、資料搜尋引擎 (information retrieval engine)：其提供了文件索引 (indexing)、搜尋。使用者可單純藉由索引取得資料或是利用其搜尋功能。

4、群組軟體 (groupware)：群組軟體的主要目的是協助一群人一起工作的，藉由群組軟體使用者可以互相溝通、協調而解決問題，傳遞的內容包含文字、聲音及影像。資訊技術可以打破時空的限制，免去必需面對面才能解決問題的困擾 (Shim et al. 2002)。企業內部員工可以藉由企業內部網路的群組軟體分享資訊；而客戶、供應商及合作夥伴也可以藉由企業間網路達到資訊分享的目的。

5、電子公告欄 (electronic bulletin board)：電子公告欄提供了一個虛擬空間讓有共同專業的團體 (communities of practice)的人在上面交流訊息，通常在組織內這是一種非正式的組織架構，它的形成是自動自發地當有人需要幫忙、或有人提供了新點子等。網路社群吸引人們的地方，是它提供了一個讓人們自由交往的生動環境，雖然有的時候只是萍水相逢，但是更多的時候，人們在社群裡持續性的互動，而從互動中創造出一種互相信賴和彼此了解的氣氛。(Armstrong and Hagel 1996)。

6、智慧型代理人 (intelligent agents)：智慧型代理人可以代表使用者做勞力密集的資訊處理工作，如：從數個資訊來源找到並收集所要的資料、解決資訊矛盾、並過濾不相關資訊且隨著時間過程，自動調整學習使用者需要 (Shaw et al. 2002)。

7、資料探勘：資料探勘在近幾年蓬勃發展的原因在於現代企業經常收集大量資料，如：市場、顧客、競爭對手及未來資訊等重要資訊，但龐大的資料量令許多企業組織遭遇到有效利用資料的障礙，再加上資訊超載及無結構化，使得大量資料無法發揮其價值，甚至使決策行為產生誤導與誤用。因此需要透過資料採掘技術從大量資料中挖掘出有用的資訊、知識，來解決企業所面臨的問題與輔助決策的制定以提昇企業競爭優勢。資料探勘為從資料庫中挖掘出隱藏在大量資料中先前不知道的和有用的資訊與知識，使用者可以利用這資訊或知識做為決策制定與問題解決的依據。

8、文字探勘：文件探勘有別於傳統資料庫探勘。由於傳統上的資料探勘技術主要針對結構化的表格資料，而忽略了非結構化或半結構化的文件資料中隱含的大量資訊。非結構化資料如新聞文件的本文部分，其內容並無一定的格式且通常無法直接取得關鍵資料的屬性。文件探勘具有兩個主要困難點：(1)人工進行多樣且大量的文件特徵選擇，缺乏效率且不符成本。(2)文件資料的內容維度數量過多，即特徵的屬性不易清楚定義或界定。相較於資料探勘，文件探勘需要加上額外的資料選擇處理程序，以及複雜的特徵擷取步驟。

這些資訊技術分別對應到不同的層次，如實體層 (physical layer)、資料層 (data layer)、資訊層 (information layer)、知識層 (knowledge layer)和介面層 (interface layer)，如圖 1。這裡所謂的知識層並不是真正的知識，而是其內容最接近知識的，知識使用者仍必需“了解”其內容，才能將其內化為知識。

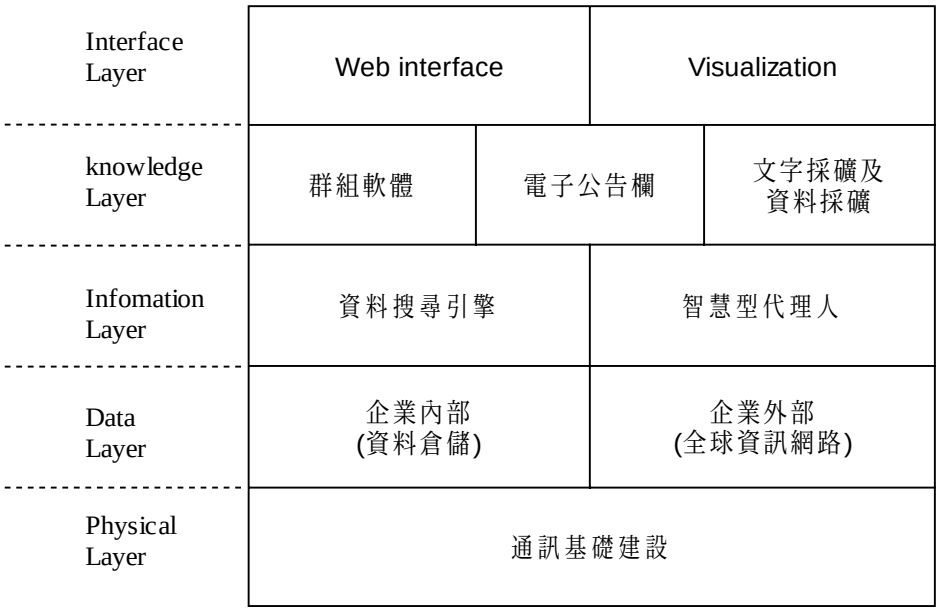

圖 1：資訊技術的分類

二、專業知識入口網站的核心功能

專業知識入口網站提供單一的入口及平台給所有的知識工作者，亦即所有的知識工作者在大部份的情況都能藉由專業知識入口網站找到需要的資料。或者是，透過專業知識入口網站的資訊自動收集功能，獲得競爭對手的最新情報。為協助知識工作者獲取所需的知識，我們也提出了知識入口網站的架構，如

圖 2。此架構分為四層，分別是資料呈現層(presentation layer)、知識創造層 (knowledge creation layer)、處理元件層 (process component layer)及資料來源層 (data source layer)。資料呈現層是與使用者互動的一層並將知識以不同的方式呈現給使用者；知識創造層強調的是各種知識的製作；處理元件層則是知識入口網站主要的核心處理元件，如 Spider、文字處理單元、文字探勘單元及視覺化單元；最後是，資料來源層，資料的來源可能是網際網路、公司內部資料或學術期刊等。

整個處理的流程如下：首先，使用者送了一個查詢字元給 Spider 並選取資料的來源，如透過搜尋引擎、某一網址、公司內部資料庫或研究期刊等。Spider 將所獲取的資料存入當地資料庫以利文字處理單元分析。文字處理單元包括了中文斷詞、詞性分析、詞的標記、關聯性字詞分析、關鍵字篩選及詞典與向量空間展示；處理完後，再交由文字探勘單元，如 Artificial Neural Network (ANN)、Support Vector Machine (SVM)、KNN、SOM 等，來進一步發現知識。最後再將結果送回給使用者。當然，最基本的就是搜尋結果；再來是其他關鍵字建議，由於許多同義詞是用不同的表示方法，藉由相關關鍵字建議，可以協助使用者描述他所想問的問題。Ontology 的製作可分為二類。第一類是主題種類化，藉由專家或使用者事先所定義好的 Ontology，資料可以自動分類到不同的 Ontology；第二類是叢集化，藉由文字探勘單元自動產生 Ontology，並將資料自動分類到不同的 Ontology。當然，由文字探勘單元所產生的 Ontology 精確度一定不如專家來得高，但它的好處就是不用請專家幫你事先定義。最後是主題地圖，所謂「一張圖勝過千言萬語(A picture worth a thousand words)」，藉由視覺化的呈現，讓使用者可以很快的感覺到整個搜尋的結果及大致的分佈情形。

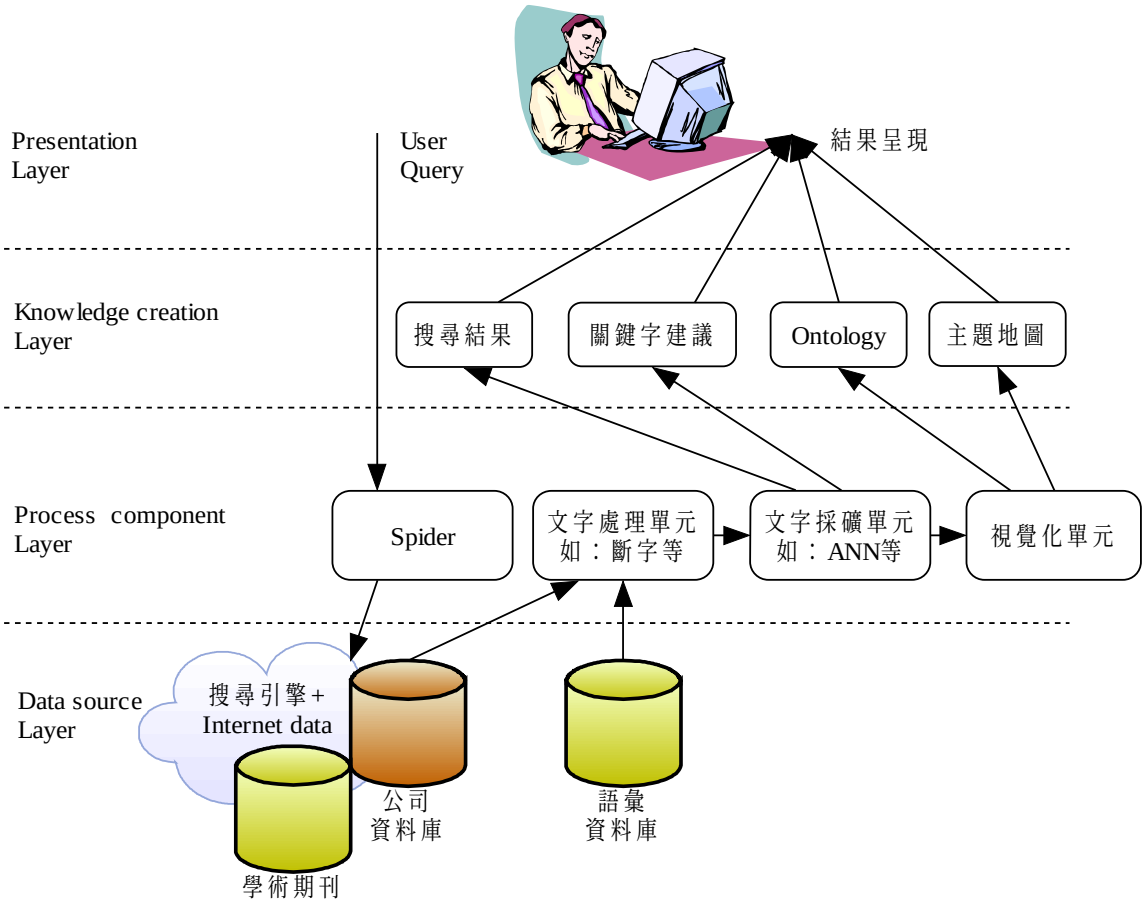

圖 2：專業知識入口網站架構

IV. 研究方法

一、研究架構

本計畫的目的在於提供高品質且豐富之財金知識入口網站，結合數值型與文字型財金相關資料源，並運用知識運算技術，以提供財金相關研究人員一高價值之知識入口網站。本計畫分三個階段來進行：

第一階段：財金資料之取得與整理

此部份在於蒐集與整理台灣地區的財金資料資源，蒐錄的範圍概述如下：

	資料庫項目	詳細內容	資料型態
	上市、上櫃公司資訊	基本資料、股利分派資訊、內部人持股異動資料、轉投資事業、每月營收、背書保證、資產負債損益表….等資訊	結構化的數字資料
	未上市(櫃)公司資訊	基本資料、資產負債及損益表…等	
	證券商概況	基本資料、資產負債及損益表、違約彙總、交易月累計資料、證券商損益分析及承銷商資料…等	

	資料庫項目	詳細內容	資料型態
數值性資料	集中市場證券交易概況	每分鐘指數、以及個股、各類股之每日、週交易狀況、零股交易、巨額成交、融資融券、集保庫存…等	
	店頭市場證券交易概況	每分鐘指數、個股及各類股之每日、週交易狀況及三大法人買賣金額等	
	國內外投資法人顧問資訊	基金經理人、基金淨值、基金名稱、投信公司股東代表人資料以及投顧與外國專業投資機構、保管銀行等資料	
	證券市場統計概要	八個主要國家證券市場上市家數、成交金額、本益比及週轉率等資本市場變動情形	
文字性資料	網際網路資源	國內外各網站資源	半結構化的文字資料、非結構化的各式文字資料
	財務預測	上市(櫃)公司財務預測 未上市(櫃)公司財務預測	
	財務報告書	上市(櫃)公司每季財務報告 未上市(櫃)公司每年財務報告	
	公開說明書	上市(櫃)公司公開說明書 未上市(櫃)公司公開說明書	
	年報	上市(櫃)公司每年年報 未上市(櫃)公司每年年報	
	重大訊息公告	上市(櫃)公司發布之重大訊息公告	

第二階段：建構財金資訊搜索知識入口網站雛形系統

財金資訊搜索知識入口網站的系統架構圖如圖 3 所示：

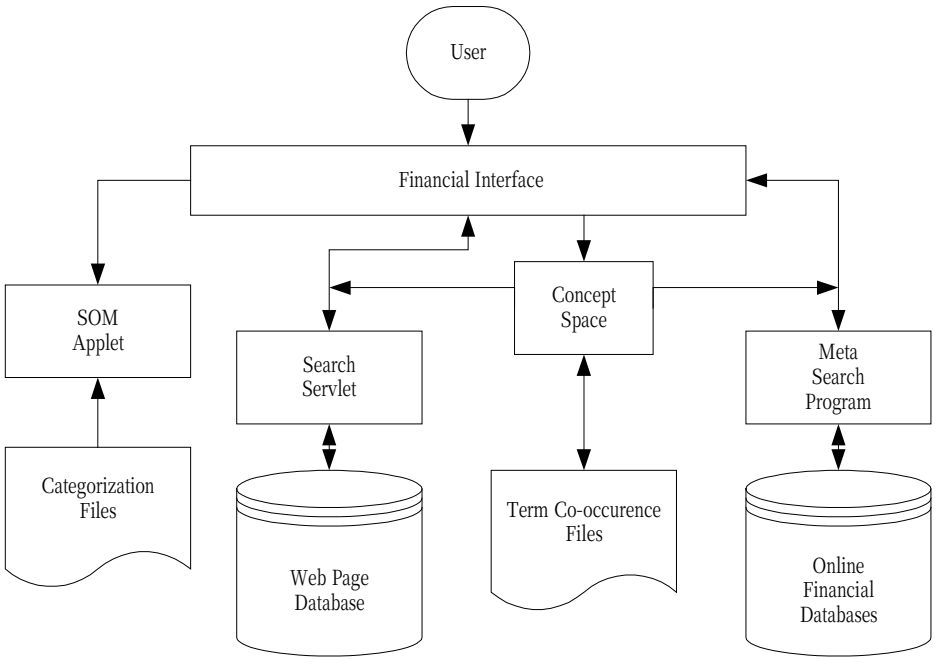

圖 3 財金資訊搜索知識入口網站的系統功能

財金資訊搜索知識入口網站主要包含下列 5 個功能：

(1)智慧型知識入口網站(Intelligent Knowledge portal)：財金資訊搜索知識入口網站的使用者界面。

使用者可以透過單一的搜尋界面，整合網頁搜尋、相關財金用語與資料庫搜尋等功能。以往使用者為了能夠獲取所有的資訊，必須進行多重版本(multi-format)的搜尋，不僅費時且十分不便。運用智慧型知識入口網站則可以節省搜尋時間並提高使用者的滿意度與工作效率。

(2)搜尋財金網頁(Search Financial Web Pages)：包括搜尋服務(servlet)與網頁資料庫。

網頁搜尋引擎的設計不僅是搜尋與主題相關的網頁，並且將網頁加以過濾以提供高品質的財金資訊。搜尋引擎的架構如圖 4 所示：

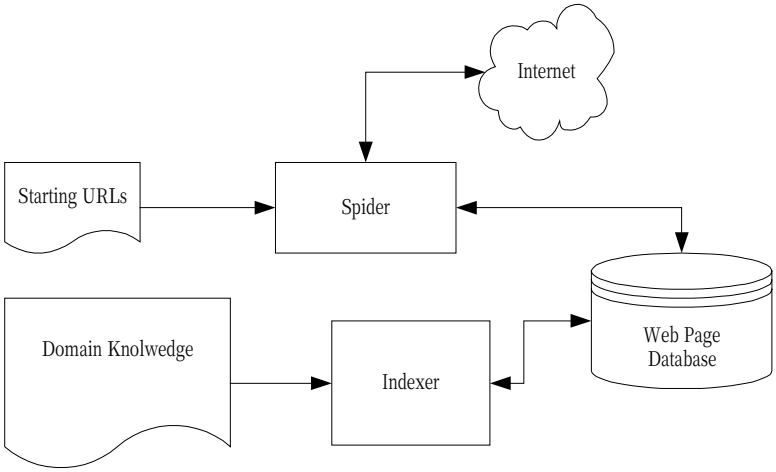

圖 4 搜尋引擎的架構

(3)搜尋財金資料庫(Search Financial Database)：包括 meta 搜尋程式與線上財金資料庫。

目前的搜尋引擎只能搜尋到所謂的”surface web”，然而網路上存在著更多的”deep web”，而這些 deep web 必須透過特定的查詢才能取得。雖然資料庫的 deep web 可能會出現在某些網頁的連結上，但是使用者並無法直接取得資料庫的資料。使用者可以透過財金資訊搜索知識入口網站，選擇所感興趣的資料庫來搜尋財金資料庫裡的資訊，或是選擇搜尋所有的資料庫，並且可以設定每個資料庫搜尋結果所呈現的數目，以避免資訊過多的情況發生。財金資訊搜索知識入口網站並會在所搜尋到的文件中，將使用者所搜尋的字彙特別標示出來，讓使用者可以很快地過濾該份文件是否重要，對其是否有幫助。

(4)相關財金用語(Related Financial Terms)：包括概念空間(concept space)與術語共同出現文件檔(term co-occurrence files)。

相較於一般採用主題標題列表或是採用由專家所發展出來的詞典的方法，財金資訊搜索知識入口網站採用亞歷桑那大學人工智慧實驗室所開發之概念空間(concept spaces)技術。概念空間技術採用先進的索引與分析技術，透過電腦計算二個用語間的關係來建議其他相關的搜尋用語。建立一個概念空間包含下列四個步驟：(1)首先要先定義收集的文件(identifying the document collection)；(2)自動產生索引(automatic indexing)；(3)共同出現分析(co-occurrence analysis)；(4)組合式搜尋(associative retrieval)。

(5)視覺化的瀏覽器(Visual Site Browser)：包含自我組織地圖(self-organizing map, SOM)程式與分類文件夾。

視覺化的瀏覽器是為了加速使用者瀏覽其所找到之財金相關研究所設計的圖形系統。除了以圖形的方式呈現主題類別外，另外還將找到的資訊以文字的方式，依字母順序排列的方式呈現。自動產生主題類別的過程包含三個步驟，其中用到的二種資訊分析技術為自動產生索引及集群分析，第三個部份則是將專家的專業知識與資訊分析技術整合。

第三階段：建置財金知識入口網站

針對某些特定財金主題(如財務危機預警、內線交易、股票未來價值估算…等)進行財金內容知識管理，開發各種財金知識運算工具庫，提供財金決策支援工具，探討財金知識市場的運作機制，建置財金知識入口網站，以提供更正確且攸關的資訊予研究及決策之用。

為了解決必須將網頁一頁一頁往下尋找導致浪費時間且無效率的問題，許多研究開始著重在將所收集的大量網頁加以分析、分類及視覺化。為了使文件能夠被取得，首先必須將文件編成索引。傳統的方法乃是一個由上往下建立分類系統與知識來源，並且事先由專家事先定義知識的展現與格式並且知識創造的過程也是結構化且定義清楚的。而透過 machine learning、統計分析與類神經網路的方法則是知識創造由下往上的互補方法。自動索引演算法(automatic indexing algorithms)已經廣泛地用於從文字中找出主要的概念，而研究亦指出此方法較人工索引更具功效(Salton,1986)。在由下往上的方法中，通常是從文字中找出片語而非僅僅是找出單字。

分類是提供使用者一個文件以更豐富的方式展示的另一個方法。已分類的文件可以讓使用者根據文件叢集的情況來了解文件間關聯性的脈絡情形。文件的叢集是依據集族的假設：關聯性較緊密的文件對於相同的指令(requests)較具相關性(van Rijsbergen,1979)。文件分類有二種方法，第一種方法是依據個別文件的屬性(如：關鍵字、作者、檔案大小…等)加以分類，透過此種技術可以呈現整個文件關聯性的藍圖；第二種方法是透過文件間的相似程度加以分類，此種方法通常包含一些 machine learning 演算法在內，例如 SOM 方法。此方法即是透過類神經網路的演算法，根據文件間的相似程度自動地將文件分類到不同的領域。

二、文字探勘

由於網際網路的興起，大量的文字提供了更多的發現知識的機會，因此，廣義來說，文字探勘也包括了智慧代理人的功能，如從數個資訊來源找到並收集所要的資料、解決資訊矛盾、並過濾不相關的資訊。文字探勘的主要工作如下(Mack et al. 2001)：

1、 將知識或資訊分類到不同群聚 (categorization 或 clustering)，來導覽(navigate)使用者找到他要的資訊。
2、 將資訊或文章做摘要 (summarize)。
3、 粹取文字中隱含的關聯性 (association)。
4、 將一大群的文章提供鳥瞰般的呈現，以期發現新知識；又稱為“主題地圖”。或是提供不同視覺 (visualization)呈現效果。

這些功能的實現必需依賴不同的機器學習或統計方法，例如，SVMs、ANNs、Decision Tree、SOM 等。我們先介紹文字處理，再介紹叢集化、主題分類、與主題地圖的建構部份將在之後討論。

1. 文字處理

自然語言的文件雖然包含豐富的描述性資料，但也因其文字的豐富性及複雜性，要直接對非結構化的自然語言文件作分析就有了許多的限制和困難。一般資料探勘的方法，只適用於結構化的關聯表格資料，無法直接運用到非結構化的文件資料上。而文字處理的目的就是在將文件中的文字或資料轉換成適合後續處理的格式，或是先將文件整理出一些初步的資訊，再從這些資訊建構之後的分析，讓進行主要步驟的時候能有更適切的參考資訊。由於中文詞與詞之間並不像印歐語系具有間隔，故在中文處理上往往需要考慮到斷詞問題。

(1) 中文斷詞

西方語言的資訊擷取技術已經發展多年，且有相當的成果，在中文方面的研究則較為困難 (許中川 和 陳景揆, 2001)，故近年來才有人開始研究 (Wong and Li 1998)。前置處理

語言、文字的第一個步驟就是斷字 (word segmentation)。主要的斷字方法有下列三種分類：字典法或稱詞庫式斷詞法 (dictionary approach) (Chien 1997; Li and Xing 1998)、語言學法 (linguistic approach) (Wu and Tseng 1993)，以及統計法 (statistical approach) (Chien 1997; Yang et al. 1998)。統計式斷詞主要是依機率統計值，訂出一組數學模式來決定斷詞的位置。此種做法的優點是大量資料處理和執行速度較快，缺點是大量的資料取之不易且統計資料會相當佔空間和詞頻會因詞典的建構者而異；「詞庫式斷詞」則根據事先建立的詞彙庫，比對的方法常見的是"長詞優先法"，逐步排除不可能的詞語組合，以達到較好的斷詞結果。此種做法優點是演算法相當直覺且實作容易。基本上將文件和詞庫中收集的詞彙比對，進行斷詞。斷詞的品質和詞庫的詞彙多寡有關，詞庫的內容必需時常更新。

良好的斷字方法對後續的步驟有著莫大的影響。如：「社會問題、國家問題」，若斷成「社會」、「國家」、「問題」，而非「社會問題」及「國家問題」。那後續的步驟就無法了解到底是什麼「問題」了。因此片語或複合的詞彙也是個重要議題。

要從文件中將片語或是複合的詞彙標示出來，一般而言有兩種方式。一種是先將所有重要項目的詞彙和它們的同義詞定義在一個語彙典 (lexicon)之中，以比對的方式將文章中有出現在語彙典的詞標示出來，這種做法標示出的項目正確性較高，也較能切合分析的需求，但是如果有詞彙或是詞彙的同義詞沒有被列在語彙典裡面，那麼它在分析中就會被忽略了。另一種方法是經由一些設定好的規則去將文件中的單字加以組合，在文件經由詞性標記後，我們就可以依據詞性的規則將單字組合成片語來處理（例如：名詞片語可以由「名詞＋名詞」、「形容詞＋名詞」等形式組成），最後再以統計詞頻等方式來作為選取的考量。

第二個部分是對文件作「詞性標記」，在傳統資訊擷取和文件分析的領域中為求過程的簡化和執行的迅速，文件常會被當成一袋的字來處理，這樣一來就完全忽略了自然語言文件所提供語義上的資訊，然而要讓電腦能理解文字的內容是件非常困難的工作，在一般文件分析中要作到完全的自然語言理解似乎也沒有其必要性，在效益的衡量之下，取而代之的便是較初步的自然語言理解，詞性標記是近年來常被應用在文件分析的自然語言處理技術，在把文件經過詞性標記之後，文件中的字不再是同樣的型態，我們可以依據自己的需要選擇不同的詞性作處理，對於文件內容的分析就有了更多的資訊做參考。詞性的簡單介紹如下：

(2) 中文詞性

由於語言詞性太多，在此僅介紹幾個重要詞性。名詞：人、事等。形容詞：凡表示實物的特徵、屬性等稱之，如：大、小等。動詞：凡指稱行為或事件的詞稱之，如：吃、喝等。副詞：又稱為「限制詞」，凡只能表示程度、範圍、時間、判斷、否定等作用，不能單獨指稱實物或實事的詞稱之，如：很、甚等。指稱詞：你、我、他。介詞：凡是能夠介繫或引進名詞、代詞或是名詞性單位到句子裡，表示時間、對象、處所、方向、範圍、原因、目的、工具和比較等各種關係的詞稱為介詞，如阿扁站「在」總統府前「向」群眾揮手。連詞：凡是用來連接兩個以上的詞、句子、甚至段落的詞稱為連詞。例如：阿扁「和」連戰攜手創造新台灣。助詞：凡是附著在句子前後或中間，表示各種語氣，或是附著在語句的中間，表示它們某種結構上的關係的詞稱為助詞。例如：呼乾「啦」！

藉由詞性分析，可以挑選出關鍵詞，以利下一步驟分析。當然，若能夠將這些詞做進一步的標記，對於文字探勘的精確度就能再進一步提高。

(3) 詞性的標記

例如：要能將這些詞標記為人名(李登輝、陳水扁)、公司名(華碩、技嘉)地點(台北、新竹)等。當然，阿扁與陳水扁應該辨識為同一人。除此之外，還要考慮的問題是有關「數值及時間資訊」的擷取問題；事實上，以關鍵字表達的文件其所描述的概念通常是各個獨立概念的集合；以往在文件關鍵字的擷取過程中，我們都會將數值直接刪除而不做考慮，然而事實上，在人類現實生活中，數值資訊所代表的概念通常是具有一定程度的連續性資訊。

(4) 特定語彙典

「詞庫式斷詞」是根據事先建立的詞彙庫，因此，對於不同領域就必需有特定語彙典，才能斷出好的詞彙。如：生物醫學用語上，基因名稱事先的訂定就非常重要了。此外，每個醫學研究人員可能有不同專精的領域與研究的方向，故在「特定語彙典」的內容上則可能因為使用者的不同而不同，或是使用者在對不同的疾病做研究時而需要有不同的「特定語彙典」；因此，在特定語彙典的介面必需能讓使用者能夠透過此一介面做語彙典的載入、編輯與儲存。

(5) 關聯性字詞 (Relational Keyword)

在醫學文件中，一個描述基因與基因間關聯性的語句，在闡述有關〝正向〞、〝合作〞或是〝負向〞的關聯性時，通常會以某些特定的詞彙來敘述關聯性，舉個例子來說，在描述有關〝正向〞的關聯性時，語句中可能會出現如〝activate〞、〝stimulate〞或是〝regulate〞等的詞彙，在描述有關〝合作〞的關聯性時，會有〝binding〞或是〝cooperate〞等的詞彙，在描述有關〝負向〞的關聯性時，則有〝inhibit〞、〝suppress〞或是〝degrade〞等的詞彙，但並非句子中出現何種類別之詞彙即代表句子含有此類別的關聯性語意，在這裡我們將這樣的詞彙稱為〝關聯性字詞(Relational Keyword)〞。這方面的研究對醫學研文件的分析有很大的幫助。

2. 叢集化

叢集化是用來將一龐大的文件集合自動切分成數個小叢集，並找出每一個叢集的主題。從整個文件集合為一個叢集開始切分，將相似的文件聚集，不同主題的文件另外再歸類。直到將某個叢集內的文件相似程度最大化，而不同叢集間的文件相似程度最小化為止。換句話說，每一個叢集內的文件都含有類似的特徵而被歸在同一類，而不同叢集間的文件主題則差異較大。叢集化適合用在下列應用：協助從集合中移除重複或幾乎重複的文件、指出集合中含有不同於其它文件主題的例外、提供大型文件集合的概觀、指出文件群組之間的隱藏結構、簡化找出類似或相關資訊的瀏覽程序。

3. 主題分類

主題分類一直是資訊擷取領域上的一項很重要的研究。且隨著現今數位資訊，如網頁、電子郵件，數量呈等比級數般的增長，文件自動分類技術的研究越顯得有其必要性與實用性。傳統以人工來進行過濾分類文件將越來越不可行。「文件分類」是提供使用者一個文件以更豐富的方式展示的另一個方法。已分類的文件可以讓使用者根據文件叢集的情況來了解文件間關聯性的脈絡情形。而與叢集化一樣，種類化會使用從文字中擷取出來的特性和統計來執行作業。它和叢集化的不同在於分類架構並非自動產生，而是以預先定義的架構

為基礎。故可透過訓練的方式，來改進分類結果，使更接近使用者所想要的目標。

定義分類架構的步驟如下；一、先定義有那些類別。可以藉由專家來定義專業領域的 Ontology。Ontology 能直接且結構性地描繪出人類的知識並明確地表現出其專業領域的知識結構。Ontology 釐清了在特定領域中有關知識內容組織、知識呈現、及知識交換等重要的觀念及作業。Ontology 可以提供作為文件探勘在文件分析的重要參考架構，特別是針對眾多不同專業領域的特徵擷取及知識探索。藉由各領域專家所建立的 domain-specific ontology，文作探勘系統可以從的大量未知文件中找出概念上與 ontology 模型中相符的 patterns 並從中探勘出有用的知識。二、在每個類別中先放置一些樣本文件。三、執行訓練工具來建立分類原則索引。因此，ontology 的製作可以是人工或自動，文件分類的過程也可以是人工或自動。當 ontology 與文件分類都是靠人工進行時，是最耗時地，但相對精確度也較高。當 ontology 是自動進行而文件分類都是靠人工進行時，可以藉由叢集化先將文件分成若干群，再針對每一群命名。當 ontology 是人工進行而文件分類是自動時，就如同是文件種類化，當然也可以用關鍵字直接進行文件分類。當 ontology 與文件分類都是自動時，是最省時地，但相對精確度也較低。

三、建構主題地圖

1. 文字處理單元：詞典與「向量空間展示」及關鍵詞篩選

通常在斷詞後，有數千個關鍵字可能會從文章中被萃取出來。一般多採用 Salton (1989) 所發展的詞典與向量空間展示(vector space representation)，其主要是利用詞彙頻率 (term frequency, tf_{ij})與文章頻率 (document frequency, df_j)的計算來代表文章。詞彙頻率 tf_{ij} 是指詞彙 j 在文章 i 中出現的頻率；文章頻率 df_j 則是資料庫中有多少文章包含詞彙 j 乘以字數的數目。篩選關鍵詞所用步驟如下：

(1) 決定文章頻率 (df_j)的臨界值 (threshold)，來刪除一些出現過少的詞彙。藉由臨界值可以刪除一些雜訊 (noisy)詞彙並增加分類的效率，但也可能造成一些資訊的流失。而文章頻率的計算會依照字數的多寡來加權，如：會計學，是一個三個字的詞彙，因此文章頻率為原本的文章頻率乘 3。如此，是希望字數多的詞彙能夠留下來；因為，字數越多的詞彙通常所表示的意思也越清晰。

(2) tf x idf (term frequency and inverse document frequency)的計算[1]。

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times I_j\right)$$

---

[1]過濾一般性詞彙：此步驟為過濾一般性詞彙留下關鍵詞彙。在資料檢索領域，最常使用的方法是逆向文件頻率(Inverse Document Frequency, IDF)，反映詞彙在文件集合中的分佈情形[Spark Jones 1972]：

$$IDF(w)=\log_2(n)-\log_2(O(w))+1$$

其中n 是文件集合的文件總數，O(w)是包含詞彙w 的文件總數。當w 出現在一半以上的文件，則其IDF 小於等於0，我們可以認為這個詞彙出現在大部分文件中，因而對於文件集合中的文件較不具有鑑別性。例如有一文件集內含1000 個文件數，有兩詞A和B在文件集都分別出現了2000次，詞頻皆為2000，無法由此區分兩詞彙的重要性。以文件數的角度來看，若A 與B 分別出現在100 篇及900 篇文件之中，則A 及B 兩詞的逆向文件頻率分別為4.3

N 代表文章的總篇數；I 是關鍵詞的長度（字數）

這個式子的意義是詞彙出現越多次、出現在較少的文章中(代表這個詞彙比較特殊)以及字數越多會給予較大的權重。藉由 tf x idf 的計算結果加以排序，再選出最重要的關鍵詞。

在中文斷詞與關鍵詞篩選後，就可以進行文件分類、分群或主題地圖的實作。

2. 利用 SOM 來製作主題地圖

自我組織映射圖網路(SOM)所提出，它是一種無監督式學習網路模式(Kohonen, 1995)。自我組織映射圖網路最大的目的，就是要將高維度的特徵，映射至一維或二維的輸出神經元陣列。換句話說，當特徵之間存在某種測量或拓撲上的關係，即使在高維度，我們希望透過權值 (weights)的學習，使得輸出神經單元之間保持一種拓撲上的關係，而這種陣列的拓撲關係，可以用來了解特徵之間的關係。SOM 為兩層式且完全連接的類神經網路，透過神經單元分佈的自我組織過程 (self-organizing process)，可以將相似的神經單元分在同一類。其主要優點為將高維度資訊視覺化呈現於二維度上，它將相似的資料聚集在最接近它節點群上 (node)，用來分類多維度的資料。

SOM 的基本精神為，輸出層在與輸入資料比對之後，除了最贏向量(winner vector)會調整外，其附近之向量也會隨之調整，如此便能讓鄰近集群相似，這是與其它群聚演算法最大的不同處。使用 SOM 演算法後，越相近的分群將會越來越接近，最後，所呈現的分群結果會變成越相近的分群會排的越鄰近 (Kohonen et al. 2000; Merkl and Rauber 1999)。它能夠將高維度的輸入資料轉換成一個有規則的低維度矩陣方格。SOM 主要參數有學習速率(learning rate)、鄰近距離(neighborhood)與地圖大小(map size)。學習速率是用來控制權重調整的參數，鄰近距離指的是最贏向量影響範圍，由於本研究使用 GHSOM(於之後介紹)，所以地圖大小可以自動調整。

SOM 在主題地圖的建立扮演了核心關鍵，Lin et al. (1991)首先提出了如何利用 SOM 製作“主題地圖”。而早期的主題地圖只是一層平面，並無法階層式顯示，而且在地圖標記上的彈性較小。之後，有許多研究在探討如何精煉其視覺呈現效果 (Yang et al. 2003; Yang and Lee 1999)或是加強地圖的標記 (Dittenbach et al. 2002; Rauber 1999)。

主題地圖建立的第一步就是先將所收集到的文章以詞典與向量空間展示法來表示。換言之，每一篇文章都是一個向量，而向量的組成就是經由中文斷詞與關鍵詞篩選後的詞彙。第二步就是將這些向量送進 SOM 演算法中。將這些文章依相似性排在 SOM 的地圖之後，再由這群向量中，由 SOM 中權重的大小，挑出合適的詞彙，以代表這群文章所代表的含意。在下一節，我們將進行主題地圖的實證研究。本研究採用 Growing Hierarchical Self-Organizing Map (GHSOM)[2] (Dittenbach et al. 2002; Rauber 1999)，相較於傳統的 SOM 製作，GHSOM 加強了三個部份。

一、地圖的大小可以由演算法自行決定，而不需要事先指定。

二、傳統 SOM 的地圖是一層平面，而 GHSOM 可以由演算法決定階層式的地圖深度。這是一個兩階段的分群方式，首先採用 GHSOM 產生一個雛形 (prototype)來當作下一階段分類的資料。除了呈現上能有階層效果，並可減少計算時間，及視覺負擔 (visual load) (Yang et al. 2003)。

[2] http://www.ifs.tuwien.ac.at/~andi/ghsom/

在標記上，傳統的 SOM 對每一群集只標記一個特徵值，但如果這個特徵值意義不大，那就無法了解這集群所代表的意義。而 GHSOM 可以在特徵值中選出多個具代表性的特徵值，幫助使用者解讀群集的意義。傳統的 SOM 雖然有視覺化的功能，但卻無法自動偵測出各群集之間的界限，因此自動標記 (automatic labeling)的目的就是找出具代表性的特徵屬性，將分群後的點標記出主要的特徵屬性，LabelSOM (Rauber 1999)。

V. 結果與討論

一、研究結果

臺灣的證券暨期貨市場為一高度管制的資本市場，政府主管機關主要為財政部證券暨期貨管理委員會。除了官方的管制外，這些市場往往仍須受各個民間管理機構及自律組織的約束。以上管制及約束機構相關的法令規章數量龐大，除非專業人士，否則一般人往往難以對證券及期貨交易市場相關的法規有一清楚的認知。

首先，我們運用 web spider 彙整這些相關機構的法令規章，共計有 832 則，包括證券交易法、臺灣證券交易所股份有限公司有價證券上市審查準則等。其次，運用中文斷詞軟體，以長詞優先的規則，將這些法規檔案進行斷詞處理，共計斷出 5571 個詞彙，並統計相關的詞彙頻率及文章頻率值。接著，在特徵的選取上，我們以出現在這 832 則法規文章內所有詞彙之 tf x idf 值前 2000 大為選取標準，作為發展 SOM 的輸入值。最後，以這些文章在 2000 個特徵的 tf x idf 值作為輸入向量，運用 GHSOM 技術繪製證券暨期貨法令規章主題地圖。本研究設定 GHSOM 中的標籤閥值（label threshold）大於等於 0.35 以上的詞彙作為關鍵詞彙，最多選取三個詞彙作為地圖標籤，故可在圖上顯示一至三個關鍵字來提示使用者。在 SOM 的參數設定上，起始的學習速率設為 0.5，起始鄰近距離設為 3，起始的地圖大小設為 3x2，並依據 GHSOM 發展臺灣證券暨期貨法令主題地圖(參考圖 5~7)。

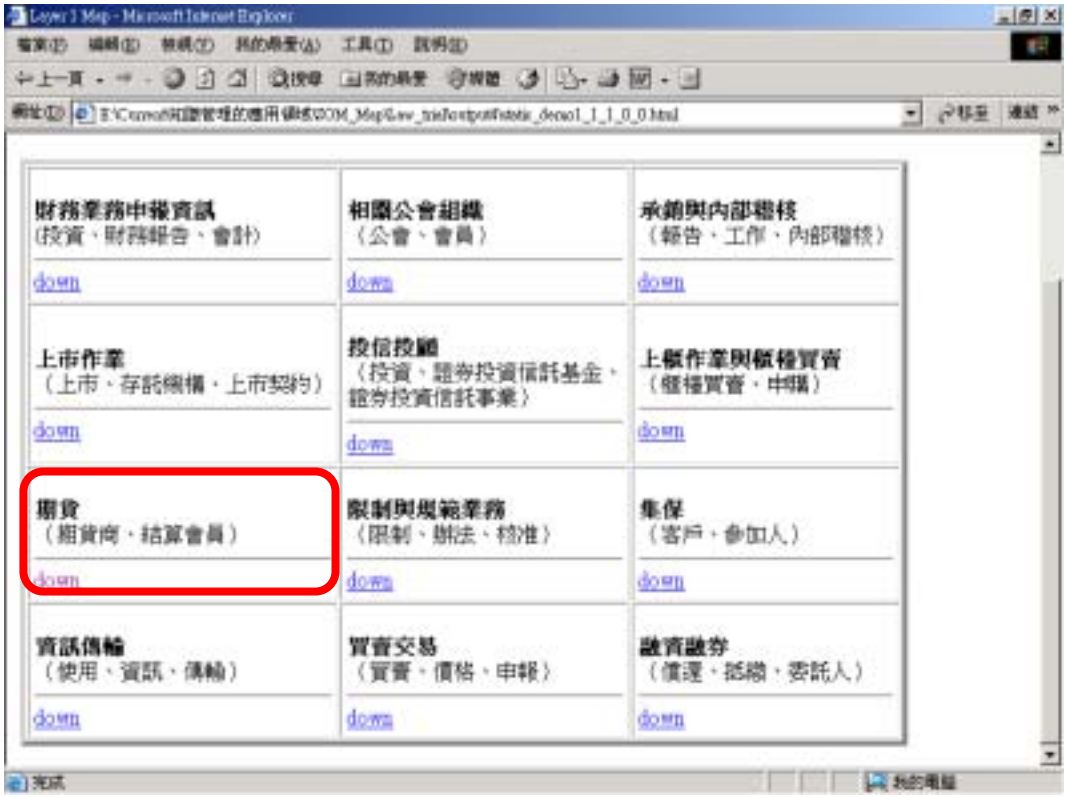

圖 5：臺灣證券暨期貨法令主題地圖—第一層（12 個主題）

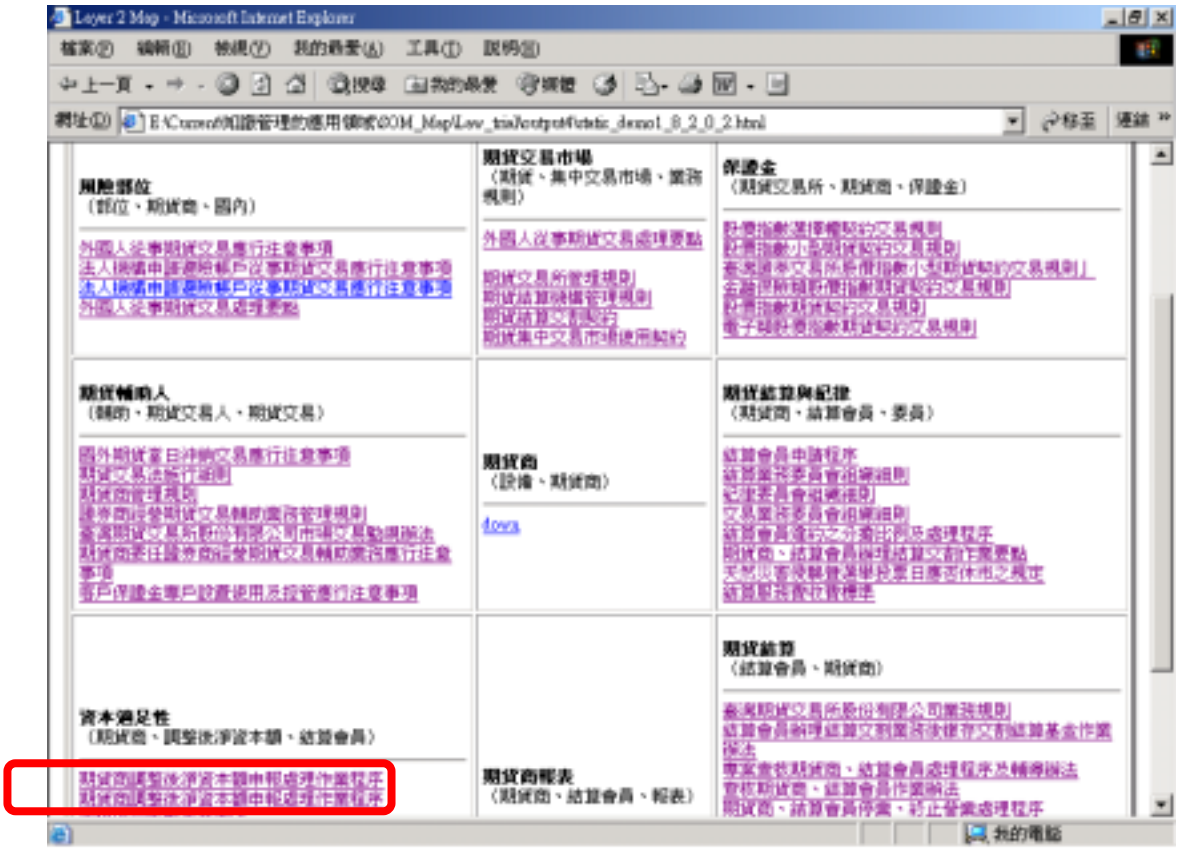

圖 6：臺灣證券暨期貨法令主題地圖—第二層（期貨）

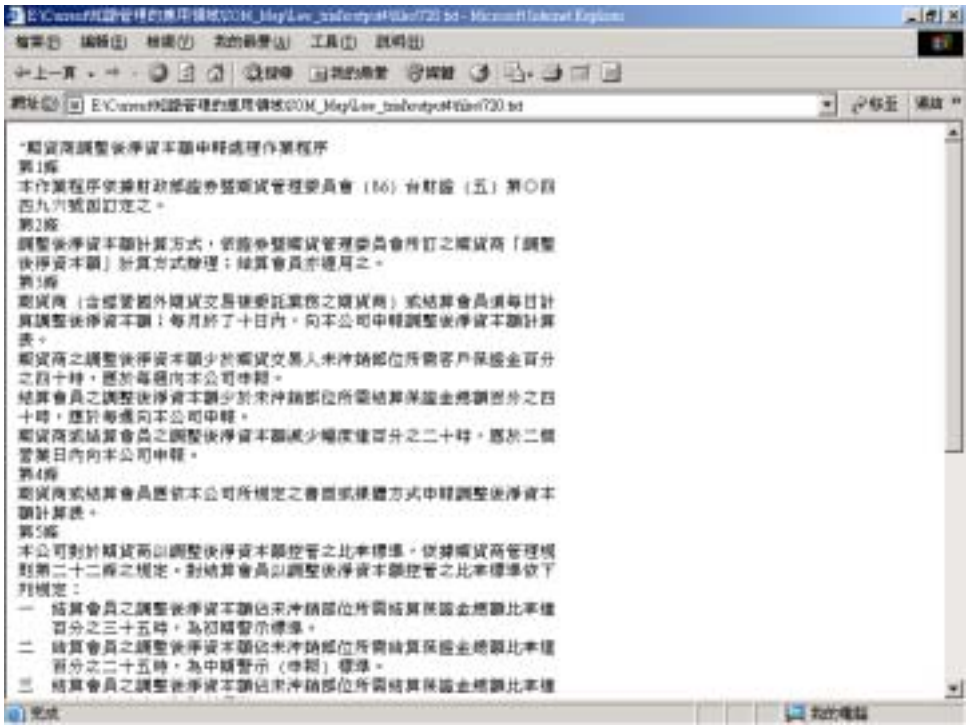

圖 7：臺灣證券暨期貨法令主題地圖—第三層（資本適足性相關法規--期貨商調整後淨資本額申報處理作業程序）

臺灣證券暨期貨市場法規大致上呈現：財務業務申報資訊、相關公會組織、承銷與內部稽核等十二個主題（參見圖 6），其下再發展每一主題的次主題，以「期貨」主題為例（參

見圖 7)，其下可再細分為：風險部位、期貨交易市場、等九個次主題。以「資本適足性」為例，使用者點選「期貨商調整後淨資本額申報處理作業程序」超連結後，可閱讀「期貨商調整後淨資本額申報處理作業程序」這一則法規（參見圖 8)。

對於任何一個臺灣證券暨期貨市場參與者而言，往往須受八百多個法令規章的規範，但學習與瞭解八百多種法令規章本身即是需要相當長的學習時間。本研究期望藉由知識管理的相關技術以發展出可讀性高且具有導覽功用的知識表達方式—主題地圖，以協助市場參與者透過層層導引以瞭解市場相關規範。

二、結論與建議

由於資訊的快速累積，各行各業都亟需較佳的資訊技術來協助他們。例如在法律業務的處理上，如何從繁多的案例、法規中，找出相關的文件，協助律師、法官辦案；醫生如何從過去的診斷記錄，找出相關資訊，作為判斷病情的依據；新聞從業人員如何從過去眾多的新聞報導中，搜尋某一相關主題，作為專題報導或歷史回顧。然而，僅藉由搜尋引擎來尋找知識是不足的，通常透過搜尋引擎來尋找相關的資料並不能一次就協助使用者找到他所想要的資料，使用者必需去瀏覽許多不必要的網頁，即使目前大部份的搜尋引擎都有提供依相關性排序及本文摘要的功能。因此，我們希望“主題地圖”能為他們提供部份解決答案。

相較於傳統的分類方式，主題地圖除了能將文件分類，並自動將每一群集“命名”。藉由不同群集的距離遠近, 也能了解相關群集的差異性。除此之外, 使用者更可藉由 hyper-link 的方式，更進一步了解各群集中精確的含意。在本研究中，我們挑選了「證券暨期貨專業領域」.來進行主題地圖的實作。研究發現顯示證券暨期貨專業領域的用語一致性頗高，故 832 個法令所採用的詞彙總數僅有 5571 個，專業領域的主題地圖易於發展，結構與階層明確，便利萃取與組織知識。

未來在這方面的研究可以著重於以下幾個方面進行。其一、SOM 演算法本身的改良。其二、不同的視覺化呈現，會給使用者不同的感覺，如何以適當的顏色或互動性來幫助使用者快速發現知識，也是值得努力的方向。最後、專業詞彙庫的建立，以斷出有意義的關鍵詞。

參考文獻

1. 許中川, 陳景揆, 2001, 探勘中文新聞文件, 資訊管理學報, 第 7 卷, 第 2 期, pp. 103-122。
2. Alavi, M. and Leidner, D.E. "Review: knowledge management and knowledge management systems: conceptual foundations and research issues," MIS Quarterly (25:1), 2001, pp. 107-136.
3. Allee, V. "The art and practice of being a revolutionary," Journal of Knowledge Management (3:2), 1999, pp. 121-131.
4. Armstrong, A.G. and Hagel, J.I. "The Real Value of On-Line Communities," Harvard Business Review, 1996.
5. Arora, R. "Implementing KM - a balanced score card approach," Journal of Knowledge

Management (6:3), 2002, pp. 240-249.
6. Bloodgood, J.M. and Salisbury, W.D. "Understanding the influence of organizational change strategies on information technology and knowledge management strategies," Decision Support Systems (31), 2001, pp. 55-69.
7. Chien, L.F. "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," Proceedings of the 1997 ACM SIGIR, 1997, pp. 50-58.
8. Cho, C.G., Jerrell, C.H., and Landay, C.W. Program Management 2000: Know the Way - How Knowledge Management Can Improve DoD Acquisition, Defense Systems Management College, Virginia.
9. Choi, B., and Lee, H. "Knowledge management strategy and its link to knowledge creation process," Expert Systems with Applications (23), 2002, pp. 173-187.
10. Dittenbach, M., Rauber, A., and Merkl, D. "The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data," Neurocomputing (48), 2002, pp. 199-216.
11. Hlupic, V., Pouloudi, A., and Rzevski, G. "Towards an integrated approach to knowledge management: 'hard', 'soft' and 'abstract' issues," Knowledge and Process Management (9:2), 2002, pp. 90-102.
12. Kohonen, T., Self-Organizing Maps Springer-Verlag, Berlin, 1995.
13. Kohonen, T., Kaski, S., Lagus, K., Salojvi, J., Paatero, V., and Sarela, A. "Self Organization of a Massive Document Collection," IEEE Transactions on Neural Networks (11:3), 2000, pp. 574-585.
14. KPMG "Insights from KPMG's European Knowledge Management Survey 2002/2003," 2003.
15. Li, Z., and Xing, L. "Search the Chinese Web - Design and the Operation of Net-Compass," Proceedings of the First Asia Digital Library Workshop, 1998, pp. 42-46.
16. Lin, X., Soergel, D., and Marchionini, G. "A self-organizing semantic map for information retrieval," Proc. of 14 th ACM/SIGIR Conf. Research and Development in Information Retrieval, 1991.
17. Mack, R., Ravin, Y., and Byrd, R.J. "Knowledge portals and the emerging digital knowledge workplace," IBM Systems Journal (40:4), 2001, pp. 925-955.
18. McDermott, R. "Knowing in community: 10 critical success factors in building communities of practice," IHRIM Journal (March), 2000, pp. 1-12.
19. Merkl, D., and Rauber, A. "Automatic Labeling of Self-organizing Maps for Information Retrieval," Proceedings of ICONIP '99. 6th International Conference, 1999, pp. 37-42.
20. Nemati, H.R., Steiger, D.M., Iyer, L.S., and Herschel, R.T. "Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing," Decision Support Systems (33), 2002, pp. 143-161.
21. Rauber, A. "LabelSOM: On the Labeling of Self-Organizing Maps," Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), Washington, DC, 1999.
22. Salton, G. Automatic Text Processing Addison-Wesley, MA, 1989.
23. Shaw, N. G., Mian, A., and Yadav, S. B. "A comprehensive agent-based architecture for intelligent information retrieval in a distributed heterogeneous environment," Decision

Support Systems (32), 2002, pp. 401-415.
24. Shim, J. P., Warkentin, M., Courtney, J. F., Power, D.J., Sharda, R., and Carlsson, C. "Past, present, and future of decision support technology," Decision Support Systems (33), 2002, pp. 111-126.
25. Swap, W., Leonard, D., Shields, M., and Abrams, A.L. "Using mentoring and storytelling to transfer knowledge in workplace," Journal of Management Information Systems (18:1), 2001, pp. 95-144.
26. Tyndale, P. "A taxonomy of knowledge management software tools: origins and applications," Evaluation and Program Planning (25), 2002, pp. 183-190.
27. Wong, K.F., and Li, W.J. "Intelligent Chinese Information Retrieval - Why Is It So Difficult?," Proceedings of the First Asia Digital Library Workshop, 1998.
28. Wu, Z., and Tseng, G. "Chinese Text Segmentation for Text Retrieval: Achievements and Problems," Journal of the American Society for Information Sciences (44), 1993, pp. 532-542.
29. Yang, C., Yen, J., and Yung, S. "Chinese Indexing using Mutual Information," Proceedings of the First Asia Digital Library Workshop, 1998, pp. 57-64.
30. Yang, C.C., Chen, H., and Hong, K. "Visualization of Large Category Map for Internet Browsing," Decision Support Systems (35), 2003, pp. 89-102.
31. Yang, H., and Lee, C. "A Text Data Mining Approach Using a Chinese Corpus Based on Self-Organizing Map," The Fourth International Workshop on Information Retrieval with Asian Languages, 1999.

附錄

A. 已發表或接受

1. 陳文華、徐聖訓、施人英、吳壽山，(2002)， “應用主題地圖於知識整理”，圖書資訊學刊。1(1)，pp.37-58
2. Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2003),”Credit Rating Analysis with Support Vector Machines and Neural Network: A Market Comparative Study”, Decision Support Systems.(Accepted)
3. Huang, Z., Chen, H., Guo, F., Xu, J.J., Wu, S., and Chen, W.-H. (2004), “Knowledge Mapping: Visualizing the Expertise Space”, Decision Support Systems.(Conditionally Accepted)

B. 送審中

1. 陳文華、施人英、徐聖訓 (2003), “發行人信用評等分類模式之研究”管理學報
2. Chen,H.-W. S.-H. Hsu, H.-P. Shen (2003), “Application of SVM and ANN for Intrusion Detection,” Computers and Operations Research.
3. Chen, W. -H.., J.-Y. Shih and S.-H. Hsu, (2003) “Application of Support Vector Machines in Forecasting Major Asian Stock Market Indices” European Journal of Operations Research.

計畫成果自評部份

一、研究內容與原計畫相符程度

本計畫原先預定完成之工作項目包括下列幾項：

1. 以知識運算(knowledge computing)技術為基礎，進行財金相關知識工具之開發及應用。
2. 進行財金知識市場(knowledge market)的理論架構及機制之探索。
3. 建構一財金知識入口網站雛型系統，結合資料及文件檔案來共構一具高價值、高品質的研究參考網站。
4. 建構一財金知識圖(knowledge map)，協助彙總完整的相關詞彙及其相關性，若佐以各項數據資料，則為一重要的學術成果。
5. 針對所建議的財金重要研究議題，可運用本知識庫得到一些對台灣財金政策及市場發展重要的建議方案。

就本報告而言，目前與原計畫高度相符。

二、達成預期目標情況

本計畫由於原先規劃之系統範圍頗為龐大且所需應用之資訊技術包含甚多，故於一年期限內僅完成主題地圖之建構及應用 SVM 於亞洲主要證券市場指數之預測及信用評等之分類模式。目前正在朝向完整系統之建構方向努力，希望能產生更多學術論文及一可供其他研究人員使用之專業知識網站。

三、研究成果之學術或應用價值

我們希望藉由此一雛型系統之開發，能更進一步來開發一完整的財金知識入口網站。財金資訊搜索知識入口網站雛型系統之建立，使財金研究者、財金政府決策單位、財金業界的從業人員可以更省時且更有效率地獲取重要且相關的資訊，以幫助其進行研究、制定財金政策、投資決策、授信分析等用途。

本計畫實施後，預期可為學術界、政府部門和產業界帶來以下效果：

1. 提供研究者豐富且有效的財金知識運算工具，使其能專注於創造各種財金核心理論與解決實務問題，省去繁瑣的資訊處理作業。
2. 各個學術研究單位的資料維護單一化，可節省資料維護成本。
3. 提供龐大的財經資料庫支援，提升財經政策的決策品質。
4. 運用財經研究成果知識管理系統，可作為政府部門的施政參考。
5. 提供產業界多元化的財經資料庫產品。
6. 提供產業界人士財金知識運算工具，協助其投資決策或授信決策過程。
7. 提供學術界研究成果之知識管理系統，供產業界人士參考。

四、學術期刊發表

本研究已有三篇文章為國際及國內期刊所出版或接受，另有三篇文章已送審。