圖書資訊學刊　第1卷第1期（2003.6）頁37-58

應用主題地圖於知識整理

Application of Topic Map on Knowledge Organization

陳文華*、徐聖訓**、施人英**、吳壽山***

Wun-Hwa Chen, Sheng-Hsun Hsu, Jen-Ying Shih, Soushan Wu

摘要

事實上，很多企業不是沒有知識庫或資料倉儲，而是知識庫太繁雜，以致在需要時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、未經組織與分類的、及高重複性的資料特性使得資料擷取問題更加複雜。透過一般常用的搜尋引擎（如：google）會搜尋到上千筆的資料。對於使用者而言，瀏覽超過數百萬個網頁來尋找相關的資料是一項沉重的負擔，而目前已開發的搜尋系統並無法確切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。然而，僅藉由搜尋引擎來尋找知識是不足的，因為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。通常使用者還是得透過搜尋引擎尋找數次、瀏覽許多不必要的網頁之後才能找到所需的資料，而非一次就能完成。因此本研究的主要目的，在於介紹如何利用文字探勘來發現蘊藏在大量中文文件中的知識。本文也將深入探討此技術的各項主要元件。透過主題地圖的實證研究，我們將製作兩類的主題地圖，分別是顯性知識（臺灣證券暨期貨法令資料）及隱性知識（王永慶思想哲學）。藉由這兩個地圖的比較來探討顯性知識與隱性知識在主題地圖的呈現上所發現的問題。

Abstract

Knowledge management (KM) has received much attention from both academics and practitioners in the past few years. Following the KM trend, many organizations have built their own knowledge repositories or data warehouses. However, information or knowledge is still scattered everywhere without being properly managed. The rapid growth of the

* 國立臺灣大學商學研究所教授(Professor, Graduate Institute of Business Administration, National Taiwan University)

** 國立臺灣大學商學研究所博士生(Doctoral student, Graduate Institute of Business Administration, National Taiwan University)

*** 長庚大學管理學院教授兼院長(Professor and Dean, College of Management, Chang Gung University)

Internet accelerates the creation of unstructured and unclassified information and causes the explosion of information overload. The effort of browsing information through general-purpose search engines turns out to be tedious and painstaking. Hence, an effective technology to solve this information retrieval problem is much needed. The purpose of this research is to explore the application of text mining technique in organizing knowledge stored in unstructured natural language text documents. Major components of text mining techniques required for topic map in particular will be presented in detail.

Two sets of unstructured documents are utilized to demonstrate the usage of SOM for topic categorization. The first set of documents is a collection of speeches given by Y.C. Wang, Chairman of the Taiwan Plastics Group, and the other is the collection of all laws and regulations related to securities and future markets in Taiwan. We also try to apply text mining to these two sets of documents to generate their respective topic maps, thus revealing the differences between organizing explicit and tacit knowledge as well as the difficulties associated with tacit knowledge.

關鍵詞：知識管理、知識入口網站、文件分類、主題地圖、SOM

Keywords：Knowledge management; Knowledge portal; Document categorization; Topic map; Self-organizing map

一、緒論

在知識經濟世代，如何善用資訊產生知識成為企業持續成長的利基。尤其在這詭譎多變的時代中，競爭的憑藉已由有形資產，如土地、原物料、廠房、資本等轉為無形的知識。有系統的知識及智慧，能提供企業解決問題的能力及達成目標的主要工具。由於其在企業競爭及發展上的重要性，甚至是對一企業價值的評價，也取決於企業是否有能力管理其知識及運用其智慧資本(Bloodgood and Salisbury 2001)。良好的知識管理(knowledge management)能為組織帶來許多效益，如產品創新、品質改善、提升顧客滿意度，以及降低營運成本等。

知識管理具有高度的挑戰性，因為知識通常存在於個人或透過動態、非結構化且通常細緻的程序累積在組織中，並不易透過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; Cho et al. 2000; Pan and Scarbrough 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)、專家網絡 (expertise network)、儲存非結構化的研討報告、技術文件線上查

詢，以及企業外部資料庫等。

很多企業並非缺乏知識庫或資料倉儲(data warehousing)，而是知識庫太繁雜，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、未經組織與分類的、及高重複性的資料特性使得資料擷取的問題更加複雜。透過一般目的搜尋引擎(general purpose search engines) ，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(information overload)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。本研究的主要目的在於探討如何利用文字探勘(text mining)來發現蘊藏在大量中文文件中的知識，並針對文字探勘的各項元件加以深入探討：

1. 文字探勘最重要就是如何將文件適當的以文字表達，以利後續統計分析。而相較於資料探勘(data mining)，文件資料有其特殊意義及結構，因此文字探勘的主要工作包括文件擷取、中文斷詞、及關鍵詞篩選。
2. 利用資料探勘技術來發現新的規則或現象。本研究採用自我組織映射圖（self-organizing map, SOM）來實做主題地圖(topic map)。
3. 視覺化呈現及互動結果介紹。
4. 主題地圖的實證研究：藉由文字探勘及SOM，我們做了兩類的主題地圖，分別是顯性知識 (法律資料)及隱性知識 (王永慶談話錄)。以比較顯性知識與隱性知識在主題地圖的呈現上所發現的差異。

本文的章節架構如下。在第二節中，我們首先藉由文獻探討來了解知識、知識管理等議題，並進一步指出資訊技術在這方面的強處及限制；在第三節中，我們將介紹如何製作主題地圖的整個流程；第四節中，我們將進行主題地圖的實證研究；最後是結論及未來研究方向。本文的重點並不在於設計新的演算法，而是利用現有的軟體系統來展示主題地圖的製作。

二、文獻探討

由於資訊技術在知識管理上有其限制，若不澄清這些限制而誤認為資訊技術就代表知識管理，則可能會導致意想不到的錯誤結果。因此，我們有必要先探討資訊技術在知識管理上的強處及限制；其次，再探討目前資訊技術在知識管理上的發展。

(一)知識及知識管理

Davenport and Prusak(1998)認為知識是一種流動性的綜合體，其中包括結構化的經驗、價值，及經過文字化的資訊。此外，也包含專家獨特的見解，為新經驗的評估、整合與資訊等提供架構。知識起源於智者的思

想。在組織中，知識不僅存在於文件與儲存系統中，也蘊涵在日常例行工作、程序、執行與規範當中。

Nonaka and Takeuchi (1995)提到知識創造可分為本體論(ontological dimension)與認識論(epistemological dimension)兩個構面來看。首先討論本體論，知識來自於個人的思想，而組織知識也必須由個人所創造；因此，知識的創造過程可以視為發生在一個擴大的、跨組織內部和組織之間的互動結果。而由認識論的構面來看，知識分為內隱知識與外顯知識，內隱知識是個人的，與特別情境有關，同時較難以形式化和溝通；外顯知識則指可以形式化、制度化語言傳遞的知識(Polanyi 1966)。Nonaka and Takeuchi歸納，知識和資訊主要有三個差異，其一，「知識牽涉到信仰與承諾」，也就是說知識關係著某一特定立場、看法或是意圖；其二「知識牽涉到行動」，因此知識通常含有某種目的；最後「知識牽涉到意義」，亦即它和特定情境相互呼應。知識比資訊重要，通常組織裡四處充斥著資訊，但是直到這些資訊被人們應用，這些資訊都不算是知識。就這個觀點來看，資料(data)和資訊 (information)都不算是知識，唯有在分析過資料，了解所獲得之資訊後採取行動，所獲得的才是知識 (Davenport and Prusak 1998; McDermott 2000)。

知識管理指的是以有系統、有組織的方式來改善公司的核心能力，藉由知識的利用來改善決策品質、採取行動並支持公司策略(Horwitch and Armacost 2002; KPMG 2003)。它強調組織知識而非個人知識，以及如何利用組織知識來協助企業策略。良好的知識管理能為組織帶來競爭優勢，除了其本身的不易模仿及不易取代之外，知識往往也是有效利用資源的重要因素。除此之外，知識在使用過程往往能激發新知識，而有報酬遞增的效果 (increasing return)。

知識管理既然這麼重要，為什麼成功的例子卻不多呢 (Arora 2002)？從KPMG (2003)的統計資料來看，80%的受訪者認為知識是公司的策略資產，然而78%的受訪者卻也認為，他們並沒有充份利用知識這項資產。理想與現實之間主要差距的原因如下：

1. 將知識視為傳統資產，如土地、勞力及資產，來管理。而事實上，知識是在人的頭腦中、是看不見的，因此，組織並無法強迫員工貢獻知識。知識的分享與創造只能在員工願意自動合作時才會發生 (Kim and Mauborgne 1997)。
2. 認為知識可以獨立於個人之外 (Quintas et al. 1997)。即使員工由知識庫搜尋，這並不代表他就能夠獲得知識，除非他能夠了解所獲得的知識 (Lueg 2002)。

事實上，許多的知識管理專案充其量只能說是資訊專案；更糟的是，在未能認清失敗的主因之前，有些企業就加倍地投資於管理顯性知識及資訊技術 (Fahey and Prusak 1998)。雖然，資訊科技可以協助知識的傳播，但往往由於人的私心或沒有分享的制

度，而使得個人知識只是個人所有，不願意分享。因此，資訊技術只是知識管理成功的要素之一，而非全部。知識管理要能成功必需同時考量組織設計、組織文化、績效衡量、資源提供、與策略上的結合及領導者的堅持等 (Choi and Lee 2002; Hlupic et al. 2002; Kakabadse et al. 2001; KPMG 2003; O'Dell and Grayson 1998; Quintas et al. 1997)。而一般認為，組織文化是目前知識管理最大的關鍵及障礙，而非技術方面的議題(Alavi and Leidner 2001; Davenport and Prusak 1998)。

(二)知識管理上常用的資訊技術

在本節中，我們將探討目前在知識管理上常用的資訊技術。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. 資訊搜尋引擎 (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具共同專業的團體 (communities of practice)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，尤其當有人需要幫忙或有人提供新點子時(McDermott 2000)。網路社群吸引人的地方，是在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。而互動的基礎主要是基於人類的四種需求：興趣、關係、交易、與幻想 (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執行一些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、學習使用者的需要 (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、競爭對手及未來商機等重要資訊，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，再加上資訊超載及非結構化，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。因此需要透過資料探勘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：文字探勘有別於傳統資料探勘。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。結構化資料如關聯資料庫中定義明確的表格與欄位，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。文字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，文字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。

而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)，如圖一。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，知識使用者仍須“了解”其內容，才能將其內化為知識。

(三)專業知識入口網站的核心功能

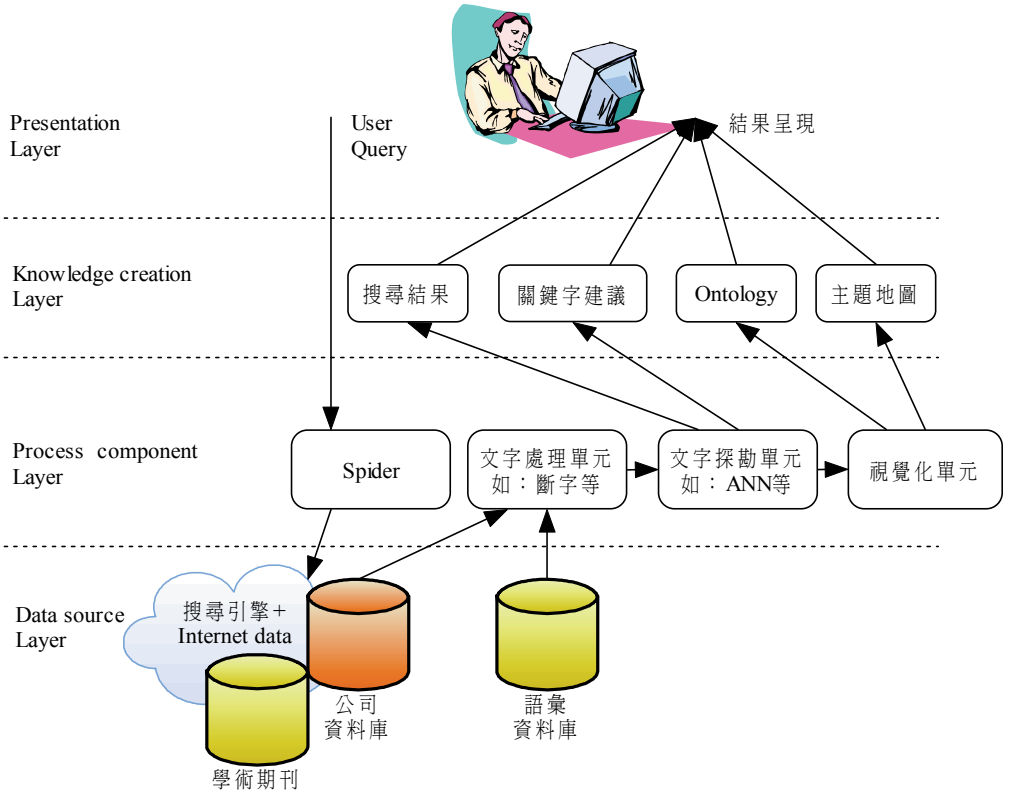
專業知識入口網站提供單一的入口及平台給所有的知識工作者，亦即所有的知

<table>
<tr><td>介面層</td><td colspan="2">Web interface</td><td>Visualization</td></tr>
<tr><td>知識層</td><td>群組軟體</td><td>電子公告欄</td><td>文字探勘及
資料探勘</td></tr>
<tr><td>資訊層</td><td colspan="2">資料搜尋引擎</td><td>智慧型代理人</td></tr>
<tr><td>實料層</td><td colspan="2">企業內部
(資料倉儲)</td><td>企業外部
(全球資訊網路)</td></tr>
<tr><td>實體層</td><td colspan="3">通訊基礎建設</td></tr>
</table>

圖一：知識管理資訊技術的分類

識工作者在大部份的情況都能藉由知識入口網站找到他要的資料。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。專業知識入口網站可以為企業帶來以下的優勢：提供一個整合的環境來分享專業的資料和知識；對跨越區域的企業據點提供相關資訊的驅動、管理、和整合；使入口網站與外部服務具有高度互動性；有一個整合的工作流程可以使入口網站的內容達到智慧化；快速容易的找尋相關的資訊；持續而可靠的效能；可延展的和富彈性的開放式服務；全球化服務。它不僅提供了對內的一般性資訊及個人協助，它也提供企業間商業資訊的流通及競爭情報的收集。

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，如圖二。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介面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。



圖二：專業知識入口網站架構

而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)、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。知識分類學(ontology)的製作可分為二類。第一類是主題分類，藉由專家或使用者事先所定義好的知識分類學，資料可以自動被分類到不同的知識分類學；第二類是叢集化，藉由文字探勘單元自動產生知識分類學，並將資料自動分類到不同的知識分類學。當然，由文字探勘單元所產生的知識分類學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 is worth 1000 words)」，藉由視覺化的呈現，讓使用者可以很快的瞭解整個搜尋的結果及大致的分布情形。

三、文字探勘

由於網際網路的興起，大量的文件提供了更多知識探索的機會。廣義來說，文字探勘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. 萃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、決策樹(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。

(一)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。由於中文的詞與詞之間並不

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。然而，中文方面的研究則困難許多(許中川和陳景揆, 2001)，直到近年才有人開始研究 (Wong and Li 1998)。前置處理語言、文字的第一個步驟就是斷詞 (word segmentation)。斷詞方法主要有下列三種分類：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。此種做法的優點是可處理大量資料和執行速度較快，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，常見的比對方法是"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。此種做法的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。斷詞的品質和詞庫中詞彙的多寡有關，且詞庫的內容必需時常更新。

良好的斷詞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。但若將這些詞都加入詞庫，詞庫的大小可能就會增加好幾倍。因此在片語或是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，這種做法所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。

2. **中文詞性**

由於語言詞性太多，在此僅介紹幾個重要的詞性。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」！

3. **詞性的標記**

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。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的刪除而不做考慮，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的專精領域與研究方向，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。

(二)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。

(三)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知識分類學(ontology)。知識分類學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，及釐清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。它可以提供文件探勘在文件分析時的重要參考架構，特別是針對眾多專業領域的特徵擷取及知識探索。藉由各領域專家所建構的特定領域知識分類，文字探勘系統可以從大量文件中找出概念上與知識分類模型相符的樣式，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，知識分類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，如表 一。

表一：知識分類學與文件分類

		文件分類	
		人工	自動
知識分類	人工	如：Yahoo	如：文件分類
	自動	如：叢集化	如：SOM

當知識分類與文件分類都是靠人工進行時，是最耗時的，但相對精確度也較高。當知識分類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。當知識分類是人工進行而文件分類是自動時，就如同是文件分類化，當然也可以用關鍵字直接進行文件分

類。當知識分類與文件分類都是自動時，是最省時，但相對精確度也較低。

四、主題地圖的建構

建構主題地圖的核心元件分別為：web spider、文字處理單元、SOM及視覺化功能。Web spider通常又稱為Web robots、Web wanderers、或Web crawlers，以下簡稱spider。而視覺化就是把數據、資訊和知識化為可視的表示形式的過程，視覺化的基本目的是要方便使用者對訊息進行觀察、操作、檢索、瀏覽、發掘和理解。

(一) 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，使得字數較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。

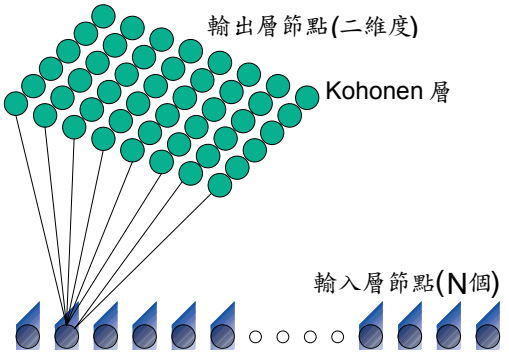
$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times I_j\right)$$

*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。

(二) 自我組織映射圖

自我組織映射圖(SOM)是在1980年由Kohonen(1995)所提出，它是一種無監督式學習網路模式。自我組織映射圖最大的目的，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，我們希望透過權鍵值 (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，如圖三，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。



圖三：SOM網路連結圖

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，這是與其它群集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)，因此，SOM是發展資料探勘技術的良好工具。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。詳細的演算法如圖四。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，本研究使用Growing Hierarchical Self-Organizing Map (GHSOM)(Dittenbach et al. 2002; Rauber 1999)，其地圖大小可以自動調整。

(三)運用SOM來製作主題地圖

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"。早期的主題地圖只是單層平面，並無法階層式顯示，且在地圖標記上的彈性較小。爾後，有許多研究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)。

Begin
Set neighborhood parameters
Set learning rate parameters
Initialize weights
While
For each input factor $\mathrm{x}_k = [x_{k1}, x_{k2}, ..., x_{km}]^{\mathrm{T}}$
For each node, compute the distance:
$d_j \leftarrow \left\| \mathrm{x}_k - \mathrm{w}_j \right\|,\ j = 1, 2, ..., n$
Find index j such that d_j is a minimum
For unit j and its neighborhoods, updates according to
$\varsigma \leftarrow \exp(-R^2 / 2\sigma^2)$
$\mathrm{w}_j \leftarrow \mathrm{w}_j + \varsigma \times \eta \times (\mathrm{x}_k - \mathrm{w}_j)$
Reduce learning rate η and radius of R of neighborhood
Until (Convergence or maximum no. of iterations is exceeded)
End

圖四：SOM演算法

建立主題地圖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輸入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。本研究採GHSOM，相較於傳統的SOM，GHSOM加強了三個部份。

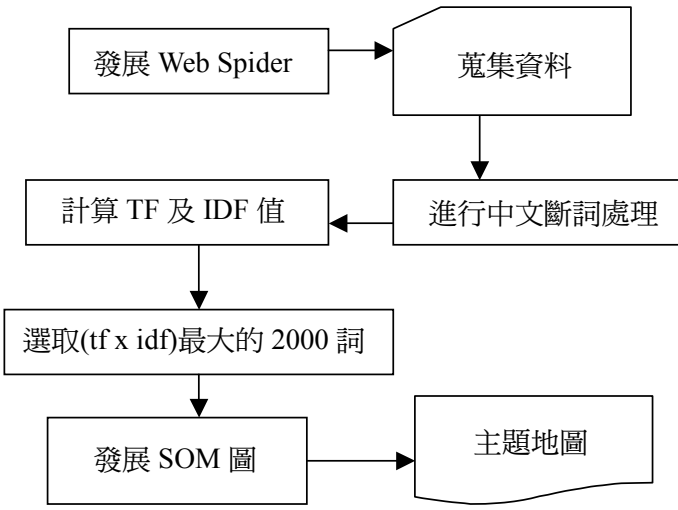
一、 地圖的大小可以由演算法自行決定，而不需要事先指定。

二、 傳統SOM的地圖是單層平面，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，並可減少運算時間及視覺負擔 (visual load) (Yang et al. 2003)。

三、 在標記上，傳統的SOM對每一群集都只標記一個特徵值，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可以選出多個具代表性的特徵值以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，將分群後的集群標記出主要的特徵屬性。LabelSOM (Rauber 1999)的概念如下：

$$q_{ik} = \sum_{x_j \in C_i} \left\| m_{ik} - x_{jk} \right\|, \qquad k = 1, ..., n$$

q_{ik}表示節點i在第k個屬性的量化誤差向量(quantization error vector)值。C_i是所有輸入樣本x_j對應到節點i的集合，m_{ik}表示權重向量(weight vector)的第k個屬性值，x_{jk}則為輸入向量的第k個屬性值。利用此公式來計算權重向量與輸入向量各特徵的距離，距離越小顯示該特徵與群集越接近，越能夠表現出此群集的特徵，藉由此算法，可挑出數個具



圖五：系統發展流程

代表性的特徵值。

五、主題地圖的實證研究

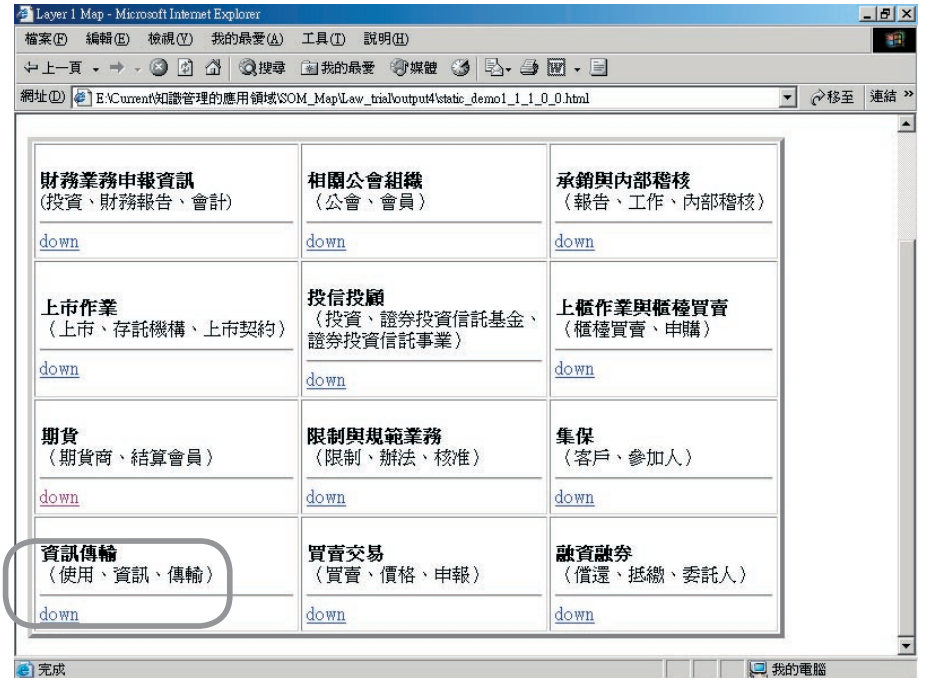
本研究期望藉由知識管理的相關技術以發展出可讀性高且具有導覽功用的知識表達方式-主題地圖。主題地圖可將相關文件經過主題分類，以協助讀者透過層級導引以瞭解該領域的相關知識。以下，我們將以具有顯性知識特徵的臺灣證券暨期貨法令，及具有隱性知識特徵的王永慶管理思想文集為例，運用文字探勘技術來建構其主題地圖，並探討此兩個主題地圖的差異。我們將蒐尋來的文件集，運用中文斷詞軟體，以長詞優先的規則，將這些文件檔案進行斷詞處理，並統計相關的詞彙頻率及文章頻率值。在特徵 (features)的選取上，我們以出現在這些文件內所有詞彙之tf x idf值前2000大為選取標準，以作為發展SOM的輸入值。最後，以這些文件在2000個特徵的tf x idf值作為輸入向量，運用GHSOM技術運算繪製主題地圖。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，實例一的起始地圖大小設為3x2（實例二則設為2x2）。

實例一：臺灣證券暨期貨法令主題地圖

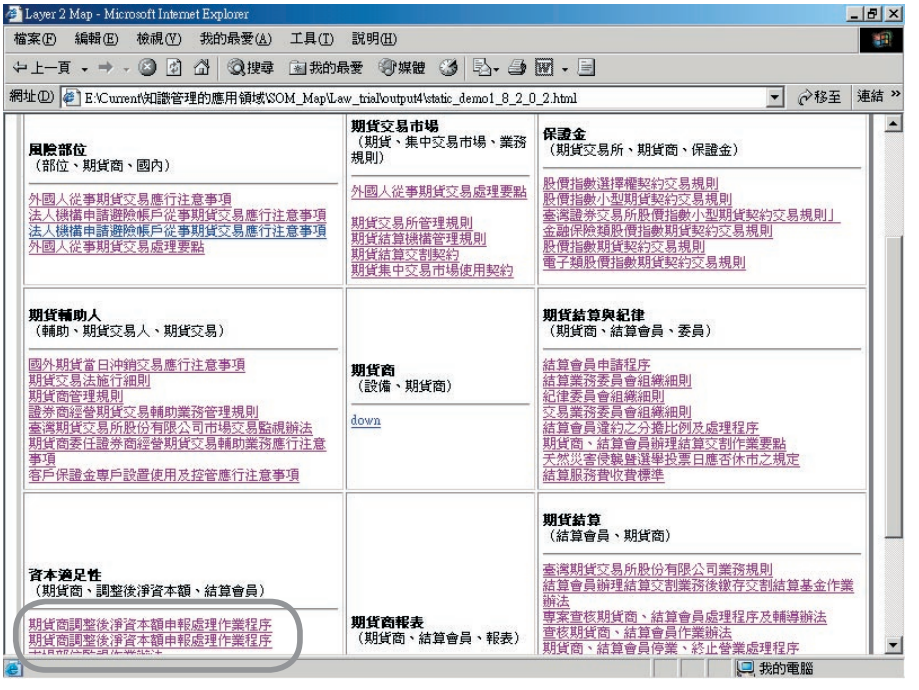
臺灣的證券暨期貨市場為一高度管制的資本市場，政府主管機關主要為財政部證券暨期貨管理委員會。除了官方管制外，這些市場往往仍須受相關民間管理機構及自律組織的約束。這些機構相關的法令規章數量龐大，除非專業人士，否則一般人往往難以對其有一清楚的認知，在認知不清下，往往容易造成誤觸法規的情事。在此，我們運用spider彙整相關機構的法令規章，共計832則，包括證券交易法、臺灣證券交易所股份有限公司有價證券上市審查準則等。運用上述文字探勘方法來發展臺灣證券暨期貨法令主題地圖如圖六～八所示。依圖六所示，這些法規大致上涵蓋十二個主題，包括財務業務申報資訊、相關公會組織、承銷與內部稽核、上市作業、投信投顧、上櫃作業與櫃檯買賣、期貨、限制與規範業務、集保、資訊傳輸、買賣交易和融資融券等主題。每一主題的次主題，以「期貨」主題為例（圖七），可再細分為九個次主題，分別為風險部位、期貨交易市場、保證金、期貨輔助人、期貨商、期貨結算與紀律、資本適足性、期貨商報表和期貨結算。同樣以一到三個關鍵字來提示使用者。以「資本適足性」為例，若使用者點選「期貨商調整後淨資本額申報處理作業程序」超連結後，即可閱讀其這一則法規的內文（圖八）。

實例二：王永慶思想哲學主題地圖

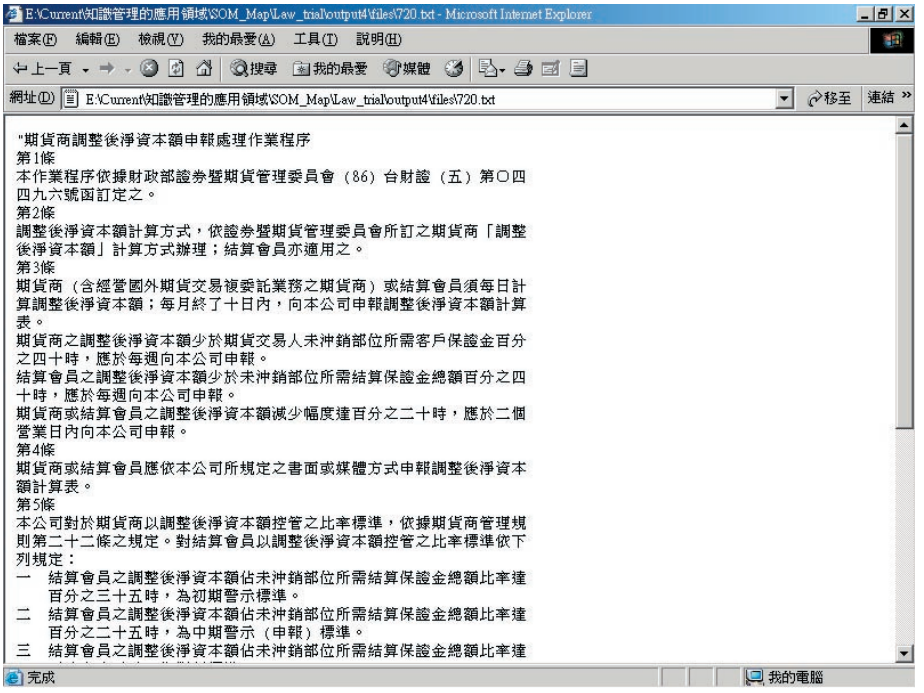
人物的思想哲學脈絡往往需由一專業作家或團隊來蒐集與採訪當事人及其相關著作，經過適當整理後，才能撰寫出人物的思想哲學史或回憶錄以傳承他人。這過程往往耗時且耗成本，除非有足夠的資源支應，否則難以達成。在知識經濟的世代中，人物的



圖六：臺灣證券暨期貨法令主題地圖-第一層（12個主題）



圖七：臺灣證券暨期貨法令主題地圖-第二層（期貨）

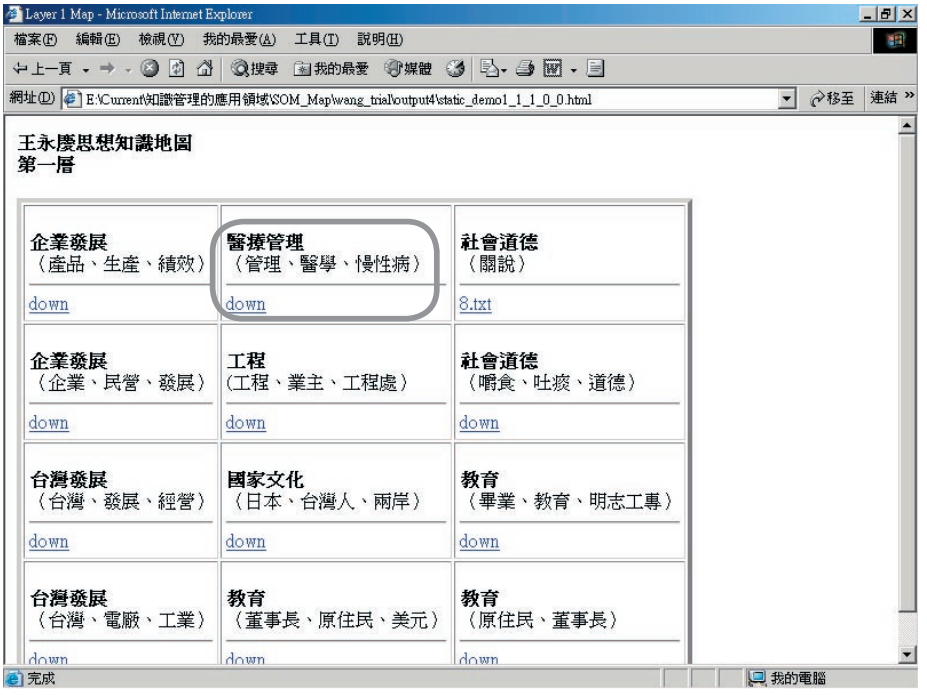


圖八：臺灣證券暨期貨法令主題地圖-第三層
（資本適足性相關法規--期貨商調整後淨資本額申報處理作業程序）

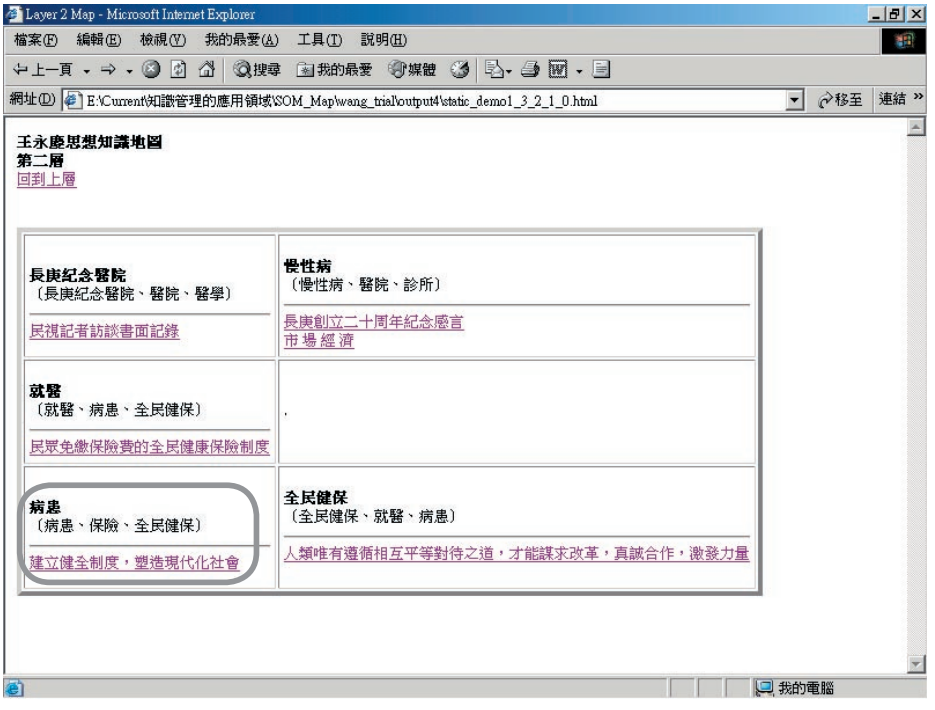
一言一行經過數位化的紀錄後，可以經由前述知識管理相關的技術，達成將如思想哲學等隱性的知識轉化為視覺化的主題地圖，以利傳承及分享。以下我們將以企業界經營之神王永慶先生為例，發展王永慶先生的思想哲學主題地圖。

本研究蒐集王永慶先生歷次的演講稿及發表的文章，共計62篇文章，包括談企業永續經營之道、對長庚大學第三屆畢業生訓勉詞...等談話性文章。發展結果（王永慶思想主題地圖）請參考圖九～十一。據圖九所呈現的主題地圖，王永慶先生的思想大致上可分為企業發展、醫療管理、工程、社會道德、臺灣發展、國家文化及教育七個主題，以醫療管理為例，可再推展其下的相關見解，包括長庚醫院、慢性病、就醫、病患及全民健保五個次主題（請參考圖十）。以「病患」主題為例，可找出「建立健全制度，塑造現代化社會」文 章（請參考圖十一）。

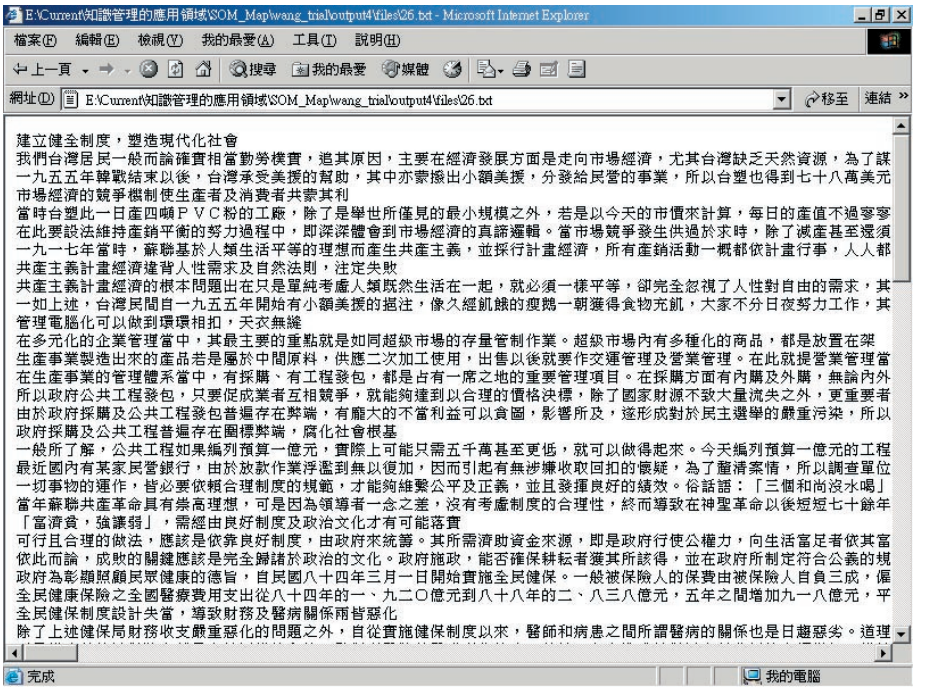
研究發現顯性知識（證券暨期貨專業領域）的用語較一致，故832個法令所採用的詞彙總數僅有5571個，反觀隱性知識（王永慶的思想），由於王永慶先生的思想涵蓋企業管理、醫療、塑化產業、教育等方面，即使僅有62篇談話性文章，但詞彙總數仍有7783個，高於前述832篇證券暨期貨法令文章。相形之下，專業領域的主題地圖較易發展，結構與階層較明確，便利萃取與組織知識。在可讀性上，王永慶思想主題地圖的發



圖九：王永慶思想主題地圖-第一層



圖十：王永慶思想主題地圖-第二層（醫療管理）



圖十一：王永慶思想主題地圖-第三層（建立健全制度，塑造現代化社會）

展受限於資料蒐集有限，再加上許多篇談話性文章所討論的主題頗多，即一篇文章包含許多主題，並非針對單一主題進行談話，因此，分群效果相對較差，導致地圖主題分佈不明顯。因此，本研究建議未來可將談話性的文章再進一步細分為數篇文章，使每篇文章的主題更明確。

六、結論與未來研究方向

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，我們刻意挑選了不同性質的文章來進行主題地圖的實作，分別是代表顯性知識的「證券暨期貨專業領域」

及隱性知識類的「王永慶的思想」。

未來在這方面的研究可以著重於以下幾個方面進行。第一、SOM演算法本身的改良。第二、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。第三、專業詞彙庫的建立，以斷出有意義的關鍵詞。

參考文獻：

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，頁103-122。

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.

Allee, V., "The Art and Practice of Being a Revolutionary," *Journal of Knowledge Management* (3:2), 1999, pp. 121-131.

Armstrong, A.G. and Hagel, J.I., "The Real Value of On-Line Communities," *Harvard Business Review*, 1996.

Arora, R., "Implementing KM - A Balanced Score Card Approach," *Journal of Knowledge Management* (6:3), 2002, pp. 240-249.

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.

Chien, L.F., "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," *Proceedings of the 1997 ACM SIGIR*, 1997, pp. 50-58.

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.

Choi, B., and Lee, H., "Knowledge Management Strategy and Its Link to Knowledge Creation Process," *Expert Systems with Applications* (23), 2002, pp. 173-187.

Davenport, T. H., and Prusak, L., *Working Knowledge*, Harvard Business School Press, Boston, 1998.

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.

Fahey, L. and Prusak, L., "The Eleven Deadliest Sins of Knowledge Management," *California Management Review* (40:3), 1998, pp. 265-276.

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.

Horwitch, M., and Armacost, R., "Knowledge Management: Helping Knowledge Management Be All It Can Be," *Journal of Business Strategy*, 2002, pp. 26-31.

Kakabadse, M.K., Kouzmin, A., and Kakabadse, A., "From Tacit Knowledge to Knowledge Management: Leveraging Invisible Assets," *Knowledge and Process Management* (8:3), 2001, pp. 137-154.

Kim, W.C., and Mauborgne, R. "Fair Process: Managing in Knowledge Economy," *Harvard Business Review*, 1997, pp. 65-75.

Kohonen, T., *Self-Organizing Maps*, Springer-Verlag, Berlin, 1995.

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.

KPMG, "Insights from KPMG's European Knowledge Management Survey 2002/2003," 2003.

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.

Lin, X., Soergel, D., and Marchionini, G., "A Self-Organizing Semantic Map for Information Retrieval," *Proc. of 14th ACM/SIGIR Conf. Research and Development in Information Retrieval*, 1991.

Lueg, C., "Knowledge Management And Information Technology: Relationship And Perspectives," *Upgrade* (III:1), 2002, pp. 4-7.

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.

McDermott, R., "Knowing in Community: 10 Critical Success Factors in Building Communities of Practice," IHRIM Journal (March), 2000, pp. 1-12.

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.

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.

Nonaka, I. and Takeuchi, H., *The Knowledge-Creating Company*, Oxford, New York, 1995.

O'Dell, C. and Grayson, C.J., "If Only We Knew What We Know: Identification and Transfer of Internal Best Practices," *California Management Review* (40:3), 1998, pp. 154-174.

Pan, S.L. and Scarbrough, H., "Knowledge Management in Practice: An Exploratory

Case Study," *Technology Analysis & Strategic Management* (11:3), 1999, pp. 359-374.

Polanyi, M., "The Logic of Tacit Inference," *Philosophy* (41), 1966, pp. 1-18.

Quintas, P., Lefrere, P., and Jones, G., "Knowledge Management: A Strategic Agenda," *Long Range Planning* (30:3), 1997, pp. 385-391.

Rauber, A. "LabelSOM: On the Labeling of Self-Organizing Maps," *Proceedings of the International Joint Conference on Neural Networks* (IJCNN'99), Washington, DC, 1999.

Salton, G., *Automatic Text Processing*, Addison-Wesley, MA, 1989.

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.

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.

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.

Tyndale, P., "A Taxonomy of Knowledge Management Software Tools: Origins and Applications," *Evaluation and Program Planning* (25), 2002, pp. 183-190.

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.

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.

Yang, C., Yen, J., and Yung, S., "Chinese Indexing Using Mutual Information," *Proceedings of the First Asia Digital Library Workshop*, 1998, pp. 57-64.

Yang, C.C., Chen, H., and Hong, K., "Visualization of Large Category Map for Internet Browsing," *Decision Support Systems* (35), 2003, pp. 89-102.

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.