圖書資訊學研究6：2 (June 2012)：37-81

以詞彙表爲基礎的知識本體雛型建構研究——以「公司治理」領域知識爲例

A Research on How to Construct the Prototype of Knowledge Ontology Based on Glossary -- Using the Domain Knowledge of "Corporate Governance" as an Illustration

周濟群

Chi-Chun Chou

國立臺北商業技術學院會計資訊系　副教授
Associate Professor
Department of Accounting Information
National Taipei College of Business

戚玉樑

Yu-Liang Chi

中原大學資訊管理學系　教授
Professor
Department of Information Management
Chung Yuan Christian University

曾建勛

Jian-Shiun Tzeng

國立臺灣大學資訊管理學系　博士候選人
Ph.D. Candidate
Department of Information Management
National Taiwan University

投稿日期：2012.5.11；接受日期：2012.6.25
email: 周濟群ccchou0412@gmail.com；戚玉樑max.chi@msa.hinet.net；
曾建勛tzeng.icr915@gmail.com

【摘要 Abstract】

本研究利用文字探勘技術（text mining），協助專家發展「公司治理」領域的知識本體（ontology）。近年來，電腦系統爲達到資訊共用及資源共享之目的，發展各領域的本體已成爲知識分享的重要方法；然而，建置各專業領域的知識本體，在實務上仍有賴各領域自行籌組，應用效益因而受限。由於本體建置過程是一項知識密集的作業，須依賴專家提供領域內的經驗或知識，再經由分析、歸納、抽象化及修正等冗長程序，因此建置本體是一項費時耗力的工作。爲解決人工作業的瓶頸，本研究以A-G方法爲基礎，提出改良的本體建置程序，稱爲AGOE（A-G ontology engineering）方法，嘗試由萃取文件中的知識元素著手，整合文字探勘、語意分析等技術，快速建置特定領域的「雛型本體」，以利提供人類專家作後續修改。本研究以「公司治理」領域爲例，分別利用文字分析建構詞彙表、建構語意網路、再以語意網路進行分析，逐步完成公司治理領域之本體。由實證的評估顯示，五位領域專家對於本研究產出的雛型本體之同意比例皆甚高，對於需要文字探勘技術輔助來建立本體知識的專業領域，AGOE方法應可作爲本體建置前置處理之參考方法。

This study utilizes text mining to assist domain experts in building a prototype of corporate governance ontology. In recent years, ontology becomes an emerging approach in building expertise to achieve reusable and sharable knowledge. However, developing a specific ontology is uneasy due to the development of labor-intensive and time-consuming issues. The development process is highly knowledge-intensive procedures including analysis, summary, abstraction and revision. To address the above difficulties, this study develops a method, called A-G ontology engineering (AGOE), improved from the A-G method. The AGOE utilizes text mining and related mechanisms for extracting knowledge items and relations into a prototype ontology. Domain experts then take advantages of the prototype to revise for a better ontology. An example of the “corporate governance” domain is implemented for demonstrating the utilizations of the AGOE method. Main works include creating a glossary, building the semantic network and analyzing their semantics for deriving corporate governance

ontology. The empirical evaluation indicates that the proportion of consenting to the prototype ontology from 5 domain experts is very high. Consequently, the AGOE is proved to facilitate users to design ontology in the early development stage, especially for those domains requiring text mining techniques to accomplish the ontology design.

[關鍵字 Keywords]

詞彙表；本體；文字探勘；語意網路；公司治理

Glossary; Ontology; Text mining; Semantic network; Corporate governance

壹、緒論

本體（ontology）是對概念明確且具體的描述，也是將抽象概念轉化爲具體描述的過程，本體常應用於電腦搜尋、資訊交換及語意理解，亦可用於跨領域整合表達共享之概念。本體也是建立知識管理系統的一個重要元件，其主要功能是產生領域的知識結構，用以儲存與使用知識。由於各個領域的專家對於建立該領域的知識管理系統之需求殷切，卻又因爲特定領域本體的發展並不多，因此許多研究在建立一個知識管理系統前，都需先建立該領域之本體。

建置本體通常由知識擷取（knowledge acquisition）開始，它是整個建置過程中最困難的一環，主要原因在於領域專家對於電腦及本體相關技術不熟悉，而負責知識擷取的領域專家或知識工程師，卻對於領域知識不熟悉（Hiroko, Simmons, Newton, & Robert, 1997）。由於建構本體相當耗時費力，且需與領域專家配合，因此許多學者以跨領域的方式，相繼投入有關本體建置方法的研究。在建置時，通常會採取兩種途徑：一、由本體專家配合領域專家，並透過人工擷取知識以建置符合領域的本體，每一步驟需經過本體專家與領域專家之間的討論與修正，投入之溝通成本占整體研究成本之比例甚高；二、爲降低第一種開發方法的時間與人力，本體專家利用半自動化之開發方法，協助領域專家來擷取知識，以更有效率的方式來建置本體。

以本研究探討主題「公司治理」爲例，公司治理資訊揭露向來被認爲是企業資訊透明度的重要指標之一，國內於民國92年起實施上市櫃公司資訊揭露評鑑制度，在此評鑑系統中，公司治理資訊的揭露即爲最主要的資訊項目，除了作爲管理當局自我評估其公司治理情況的參考之

外，更是投資者選擇投資標的時的重要訊息。然而，現行的「公開發行公司年報應行記載事項準則」或「公開發行公司年報應行記載事項準則」均未針對公司治理資訊應揭露事項制訂詳細的內容，因此建構完整的「公司治理」本體結構，將有助於監理機關制訂相關應揭露事項之報告架構。基於公司治理的意義及範圍，在國內外均無統一之標準（財團法人中華民國證券暨期貨市場發展基金會，2011）。若完全由人類專家來建置，恐曠日廢時，發生前段所述第一種本體建置方法的問題，因而促使本研究希望以第二種方法來協助領域專家建置「公司治理」本體的動機。

使用第二種方式建置本體，通常需要提供大量事實資料，並透過文字探勘（text mining）或自然語言處理（natural language processing）等方式，從中擷取建置本體之領域知識。輸入之資料主要分爲三種：分別爲結構化資料、半結構化資料以及非結構化資料（Zhou, 2007）；這類研究早期主要使用結構化或半結構化資料作爲知識擷取之來源，然而在許多專業領域中，有許多資料並非結構化，由於此類非結構化文件的表達原本僅適用於給人類理解，對於系統或機器處理幫助不大。因此，近年來國內外許多研究者開始發展由非結構化資料建置本體之方法，基於非結構化資料之本體建置亦被統稱爲基於文件分析之本體設計方法（ontology design based on corpus analysis; Aussenac-Gilles, Biébow, & Szulman, 2000）。然而，多數研究較偏向於個別技術的改善（Downey, Etzioni, Soderland, & Weld, 2004; Hamon & Nazarenko, 2001; Hindle, 1990; Morgan, Hirschman, Yeh, & Colosimo, 2003），較少將方法整合應用於某特定知識領域上（戚玉樑、蔡明宏，2007；Chi, 2007; Lee, Huang, & Fang, 2010; Liu, Hu, Wang, & Liu, 2011; Rajsiri, Lorré, Bénaben, & Pingaud, 2010; Tang & Cai, 2010），而應用於會計財金等影響經濟市場甚鉅的專業領域之研究更加少見。因此，戚玉樑與蔡明宏及Chi等學者即曾提出基於文件爲建置本體素材的研究，至少應能解決以下問題：一、如何在文件中篩選出有意義的詞彙項，並去除與概念無關的詞彙；二、如何將詞彙匯集爲知識概念，亦即以語意（semantics）的角度，將具有相類似內涵的詞彙匯集至同一概念；及三、如何建立概念之間的上下階層關係，亦即尋找概念之間的“is-a”結構。相對於過去研究或偏重於技術發展、或未能建立專業領域本體，本研究希望能同時針對上述三項問題，發展一套客觀的本體建置方法，並整合目前可行技術，成爲可操作的本體建置流程，協助本體開發者應用於特定專業領域上。

基於文件分析的本體設計，目前尚未有一般公認的方法論，以本研

究主要參考之Aussenac-Gilles等人（2000）方法（以下簡稱A-G方法）爲例，雖然該方法已提供原則性的框架，但僅以闡述概念爲主（例如「標準化」及「正規化」等），並未具體說明各原則之間應如何達成及銜接，亦未提供標準技術或工具。爲改善本體建置的方法，本研究提出基於A-G方法的改良版，除修正部分流程外，也將各原則結合可執行的技術，成爲本體建置的工程方法，本研究將它稱爲AGOE方法（A-G ontology engineering）。A-G方法是以詞典建置開始，因爲「詞典」（或索引典；thesaurus）具有詞彙明確之定義與說明（Brewster, Ciravegna, & Wilks, 2003），但實務上，並非各領域均存在索引典。因此本研究提出的AGOE方法考量實用性，將自行發展結構上較爲簡單的詞彙表（glossary）。此外，AGOE方法也針對A-G方法僅具備原則性框架而無標準技術工具的問題，自行發展出結合詞彙關係計算、協助語意網路建構的分群演算法、以及發展概念階層關係的詞彙共現法與句型分析法等。本研究儘可能以客觀方式發展上述流程與方法，以期提供其他領域本體在建置時的使用或參考。最後，本研究將以「公司治理」領域爲例，以驗證本研究的本體建置方法。

本研究之內容安排如下：第貳節蒐集並比較、分析、應用於各領域的文件分析本體設計之相關研究。第參節介紹AGOE方法的六個步驟，並說明各流程所使用的技術與方法，包括文字探勘、分群演算法以及本研究自行發展的語意結構分析法等。第肆節依AGOE方法的六個步驟，以會計領域中公司治理相關書籍與論文，作爲建構本體之輸入素材，實際建置本體雛型並評估其結果。最後，第伍節總結本研究所提出之本體建置方法，並提供研究者後續研究之建議。

貳、應用於各領域的文件分析本體設計研究

依賴專家或透過知識工程師收集專家知識以建立本體的方法，需要耗費相當多的人力與時間，於是研究者開始思考是否可透過半自動化或自動化之方式建立本體，而此類的研究通常被稱爲本體學習（Cimiano, 2006; Maedche & Staab, 2001）或文件分析本體設計（Aussenac-Gilles, et al., 2000）。本體學習（或文件分析本體設計）主要希望能透過半自動化之方式，從資料中擷取對於建構領域模型有用之知識。本體學習主要透過資訊檢索、文字探勘或自然語言處理等技術，由輸入資料擷取建置本體之領域知識，然而，輸入資料除了必須與欲建置本體有關以外，其資

料之表達方式也會影響建置方法的形成。一般來說，輸入資料主要可分爲三種：第一種爲結構化資料，例如XML綱目（schema）、UML圖形或資料庫綱目等綱目型資料；第二種爲半結構化資料，例如XML、HTML或具備表格結構之資料；第三類則爲非結構化資料，通常指無任何標記或格式之資料案例文件。早期本體學習之研究主要是使用結構化或半結構化資料作爲知識擷取之來源，然而，由於有許多資料並沒有結構化之形式，近年來許多研究者開始發展由非結構化資料建置本體之方法。

國內外應用於其他領域的文件分析本體設計研究，可依其目的概分爲本體建置的應用研究與方法改善兩類，分述如下：

應用研究類主要目的是針對不同領域，採用適合的文件分析本體設計方法，以建置各領域的知識本體，甚至應用該知識本體於決策輔助。例如：Chi（2007）、戚玉樑與蔡明宏（2007）均採用文字探勘、語言學分析、凱利方格與正規概念分析法（Formal Concept Analysis, FCA）等綜合性方法，自行發展出一套文件分析本體設計方法，針對主題爲「本體」之電子期刊論文進行分析，以建立「本體」相關詞彙之雛型本體，並以問卷調查方式進行「主觀接受率」之評估。Tang與Cai（2010）則同樣使用文字探勘與正規概念分析法建立旅遊資訊的知識本體，該研究並未進行本體之量化評估，僅以雛形本體展示代替量化評估。Liu等人（2011）則應用文字分析本體設計於資訊科學領域知識的建構，該研究採用詞彙共現性、階層群集分析、構詞分析以及相關性矩陣等不同技術，並應用社會網絡理論，自行發展其整合性文件分析本體設計方法，但亦未進行量化評估。Rajsiri等人（2010）則根據MIT Process Handbook企業流程知識儲存庫的企業流程知識，建構企業間流程協作的本體知識，並連結至其知識法則庫，以協助企業供應鏈間的協同作業。Lee等人（2010）則以臺灣數位典藏國家型科技計畫的知識平臺爲基礎，建構符合臺灣國稅局的稅務本體知識，並依據本體知識建立稅務查核規則。

方法改善類的研究則著重於單一技術或方法論的改善，並不嘗試建構領域本體，例如Hamon與Nazarenko（2001）著重於複合詞分析方法的改善，使用索引典找出組合詞之同義詞，該研究提出三種可能改善複合詞分析的方法：一、主要詞爲同義詞，修飾詞相同；二、主要詞相同，修飾詞爲同義詞；三、主要詞與修飾詞皆爲同義詞，比較效率的方法則是將研究計算結果與現有詞典進行比對。Lin（1998）則著重於向量空間模型的改善，發展新的相似度分群方法，其作法是在特定領域的文件中，將出現於概念附近的詞彙均表示爲可以代表此概念之向量，透過此向量可以改善計算概念與概念間相似度的計算精準度。Morgan

等人（2003）以現有模型資料庫爲基礎，使用簡單的樣式規則，比對現有模型資料庫中的基因名稱與摘要，以找出相關詞彙，並依據此相關詞彙結果作爲訓練資料，建立一個基於隱馬可夫模型（hidden Markov model）的基因新詞彙辨識與標記工具，該研究提出的名稱與摘要比對方式雖然不能保證產出絕對正確的結果，但可以協助研究者快速建立新詞彙與現有詞彙間的關係，並節省研究者手動建立時所需耗費的大量時間，主要評估方法爲F統計量。Downey等人（2004） 則使用句型樣式（syntactic pattern）找出同義詞，此方法使用一組與領域無關之句型結構以及特定領域之詞彙進行觀察，找出常出現在同樣句型結構中之詞彙作爲同義詞，評估方法是以領域無關之句型結構作爲基準以進行比較。Hindle（1990）同時使用句型樣式與互信息兩方法針對詞彙進行分群（clustering）以找出相似詞。由於特定的詞彙只會出現在某些結構之中，例如wine可以被“drunk”，“produced”或“sold”，但是不能被“pruned”，所以如果處於同樣結構中的詞彙，其相似度就會越高，該研究以“subject-verb-object”句型樣式爲例，計算該結構中詞彙間的交互信息，找出許多相似度高的詞彙，並以人工方式比較評估。Church與Hanks（1989）則提出以交互信息方式衡量詞彙間關係強度的方法，交互信息越高，表示兩詞彙在文件中同時出現的機率越高，兩詞彙之間的關係也越強。

本研究亦屬於應用研究類，與其他領域本體建置研究之主要不同處，在於本研究乃是依據A-G方法改良的AGOE方法，並以「公司治理」專業詞彙表爲主要素材，此外，本研究亦採用文字探勘方法，結合分群演算法與詞彙句型樣式分析，並自行發展詞彙共現性之語意結構分析方法等，於領域本體的發展、方法論改善、技術工具的應用與本體評估方面，本研究的考量均較爲全面。

除了上述各領域的本體建置相關研究外，在與公司治理相關度較高的會計、財務領域中，雖然較少有利用文字探勘建立知識本體方面的研究，但近來確實亦有若干學者運用文字探勘技術於財務會計研究中，例如Engelberg（2008）即曾定義所謂硬資訊（hard information）與軟資訊（soft information）之差異，利用自然語言處理方法，針對企業法說會內容進行語意分析，將法說會報告內容區分爲以文字表達的質性資訊（軟資訊，例如：管理當局討論與分析）與可數量化的資訊（硬資訊，例如：財務報表資訊），該研究發現軟資訊確實具有增額資訊內涵，但其反應時間較長，顯示資訊處理成本確實會影響資本市場效率。Antweiler與Frank（2004）則使用簡單貝氏分類法（naïve Bayes classifier），分

類分析Yahoo Finance上討論區的文字性資料，發現這些討論區文章確實會影響股價與成交量，因此建議分析師應偵知這些網路上的情緒性資訊（sentiment detection）。Tetlock（2007）則利用General Inquirer文字分析軟體，針對S&P 500的企業盈餘發布前的華爾街日報報導內容進行內容分析，以偵知投資人的情緒，並依哈佛心理字典（Harvard IV-4 Psychological Dictionary）進行內容的分類，結果發現負面報導確實具有異常報酬的解釋力。而國內學者周濟群與連子杰（2011）亦曾針對國內年報電子書等公開非財務資訊的資訊擷取問題，運用文字探勘技術建立「企業年報非結構化文字資訊分類系統」，以年報資料中的「句」爲單位，並應用企業策略分析中常用的分類方式，輔助使用者更有效率地分類擷取、閱讀、分析、整合這些非結構化資訊。

參、本體建置方法——AGOE

本體的發展需要領域工程師與知識工程師的相互配合；由知識工程師進行知識擷取工作，再由設計工程師以知識表達法和邏輯推理，建立系統推理引擎。Waterson與Preece（1999）提出本體建置的困難點主要是太多不同專業領域的領域本體需要被定義。爲了加強建置效率，需要依賴系統化的方法論來進行其本體工程，並驗證其品質。目前已發展出有許多本體建置及評估本體的本體工程，發展比較完備及著名的方法論有：Cyc Method（Reed & Lenat, 2002）、Uschold and King's Method（Uschold & King, 1995）、TOVE Method（Grüninger & Fox, 1995）、KACTUS Approach（Fernández-López & Gómez-Pérez, 2002）、SENSUS Method（Swartout, Ramesh, Knight, & Russ, 1997）及On-to-Knowledge Methodology（Maedche & Staab, 2001）等。上述本體工程方法常需依賴專家由人工建立，因此，在建立的過程裡，概念與關係的判斷基準常有不一致的狀況。運用文件分析、自然語言分析方法來擷取知識，除了能有一致的知識表達以取得具有共通性知識外，也可以讓本體建立的過程更爲明確。

鑒於本研究目的乃是探討如何利用詞彙表與文字探勘技術，來協助專家使用非結構化文件來發展「公司治理」領域的知識本體的方法，前段所提及之本體工程方法多數並非以文件資料（text-based corpus）作爲知識的素材來源，因而不適合作爲本研究的知識本體建置。至於文件分析本體設計至今則尚未有一般公認的方法論（詳見第貳節），故本研究乃參考Aussenac-Gilles等人（2000）所率先提出之文件分析本體設計方法

來進行，基於此方法並未具體說明各步驟原則之間如何達成及銜接，亦未提供標準技術或工具，因此本研究乃提出基於A-G方法的改良版，除依需求修正部分流程外，也將各原則結合可執行的技術，成爲本體建置的工程方法。採用A-G方法的另一原因，則是由於Aussenac-Gilles等人曾說明，運用此法進行本體學習前，須符合兩個前提假設：其一爲本體的建構者須爲該領域的專家，以便判斷概念的重要程度；其二爲該領域專家知道如何藉由語意網路建構本體，本研究之情況皆符合此二條件。

A-G方法提出藉由自然語言處理工具擷取詞彙並分析大量詞彙的含意，進而將大量詞彙轉換成特定領域的語意網路，並利用語意網路判斷概念與概念間及概念屬性間的階層關係，以下說明此法所建議進行的四大活動（圖1）。

一、語料建構（corpus constitution）：建議在選取文件上應由領域專家進行，語料內容應包含整個領域的重要字詞，此步驟可幫助專家對領域知識的重要詞彙有初步的瞭解。

二、語言學分析（linguistic study）：重點在選取適當的自然語言處理工具，並應用此工具分析詞彙間的關係。

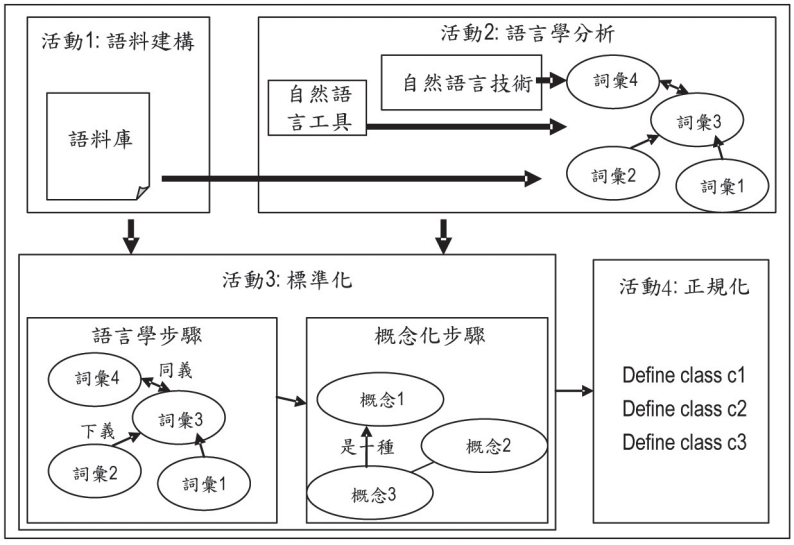

圖1　A-G方法建議的本體設計四大活動

資料來源：*Ontological engineering: With examples from the areas of knowledge management, e-commerce and the semantic web*, by A. Gómez-Pérez, M. Fernández-López, & O. Corcho, 2004, New York: Springer.

三、標準化（normalization）：本活動的結果是藉由語意網路表達概念間的關係，此時的概念模型雖非格式化的概念模型，但概念間的關係直接而清楚。此活動階段分爲兩步驟，語意分析步驟和概念化步驟；語意分析步驟主要在找出詞彙與詞彙間的關係，並根據原始文件的描述對詞彙定義做資訊上的補充，概念化步驟則是將字詞和詞庫轉換成概念和網路關係。

四、正規化（formalization）：將語意網路的概念及關聯轉換成正規化的本體知識。

A-G方法的優勢在於其觀念性架構中導入了自然語言處理的技術，而雖然某些工具已發展支援A-G方法，例如Géditerm（Aussenac-Gilles, et al., 2000）可用於將標準化階段以選擇適當字詞，Terminae（Biébow & Szulman, 1999）則可用於標準化和正規化階段。但上述各項工具均僅支援英文語系文件，並不適用於中文文件之處理，且在A-G方法流程中仍存在可改善其自動化程度的空間，因此本研究擬進一步提出工具的改善方法，以提升A-G方法應用於本體學習之效度。此外，A-G方法對於語料的選擇並未提出明確的建議，本研究則依據Brewster等人（2003）的意見，爲了提升詞彙句型樣式之關係擷取召回率，乃選擇由領域文件建構詞彙表，以作爲擷取關係之語料文本。

本研究提出的AGOE方法之六個步驟與活動簡述如表1所示，該表並將AGOE各步驟如何對照於原A-G方法的何項活動註明於括弧內，各步驟中所應用到的技術則詳述於後：

一、詞彙表建置

㈠文本蒐集與處理

由於公司治理一詞概念相當抽象，且其涵蓋的範圍甚廣，因此，本研究收集公司治理、內部控制及企業風險管理爲主題之書籍、期刊及博碩士論文作爲詞彙表建置語料來源。其中，期刊與碩博士論文多已爲電子檔，書籍部分則透過掃描方式進行電子化，最後再將所有電子文本轉存爲一致的電子文件格式。

㈡關鍵詞篩選——詞頻與交互信息

爲找出代表公司治理領域之關鍵詞，本研究利用電子文件內容產生N元詞（N-gram），並透過詞頻（term frequency）與交互信息（Mutual Information, MI）進行關鍵詞篩選。交互信息的使用方式如下：

爲建立領域關鍵詞詞庫，我們將利用電子文件內容產生N元詞。由

表1　AGOE方法之六個步驟與活動簡述

AGOE方法		對應至A-G方法
步驟一至六	主要活動	
一、詞彙表建置	1.透過中文電子期刊資料庫與電子掃描技術，蒐集並處理領域（本研究爲「公司治理」）相關議題文章。 2.輔以文字探勘技術，擷取出若干共現性高的詞彙。 3.再以詞彙爲本，摘錄出文章中之重要關鍵詞及其釋義，建構出領域詞彙表。詞彙表中，具有解釋之詞彙稱爲詞項。	A-G方法之語料建構
二、文本處理並產生詞庫	1.針對前一步驟所建構之詞彙表解釋內容，利用中央研究院（2009）中文斷詞系統，進行中文斷詞。 2.對斷詞結果作適當的修正，產生本研究之公司治理詞庫。	A-G方法之語言學分析
三、語意網路建構	1.基於前一步驟詞庫所建立之詞彙向量空間，利用TFIDF爲詞項建立特徵向量。 2.進行詞項間的相似度計算，計算出詞項間之關聯程度，以找出關聯性較高之詞項集合。 3.並依據所篩選出之詞項關聯程度，進行詞彙間聯結，以網絡的型態呈現，形成語意網路。	A-G方法之標準化
四、語意結構分析	1.透過分群（clustering）演算法，以協助人類專家進行後續的語意結構分析，將互有緊密關聯的詞彙聚集「群集」，並建議出「群集中心」。 2.人類專家則以「群集」爲基礎，輔以詞彙共現性、同義、反義分析與句型分析等方法，以建構出公司治理本體知識之概念階層。	A-G方法之正規化
五、雛型本體產出	1.依據步驟三至五進一步將語意網路轉換爲以OWL（web ontology language）語言所表達之領域本體知識。	A-G方法之正規化
六、雛型本體評估	1.設計適當問卷，並尋求五名具有公司治理知識之專家進行問卷調查，針對雛型本體的詞彙、概念、階層結構、及建置方法等，進行妥適程度之評估，以檢驗「雛型本體」可否提供後續再利用。	無對應步驟

於並非所有N元詞皆正確，例如「風險評估屬於內部控制的組成要素之一」這句話前面數個字便可產生「風險」、「險評」、「評估」、「估屬」、「屬於」等二元詞，但其中僅有「風險」與「評估」、「屬於」等三個二元詞正確。構詞強度可協助篩選正確的N元詞，而交互信息可作爲測量構詞強度之參考（Yang, Luk, Yung, & Yen, 2000）。因此，本研究透過詞頻與交互信息等方式，選出高構詞強度之N元詞。本研究採用之交互信息（MI）計算公式如下：

$$MI(w) = \log_2 \frac{(N_c)^n \times f(w)}{N_w \times f(c_1) \times f(c_2) \times \ldots \times f(c_n)}$$

-- c_i：字元i

-- w：詞，$w = c_1, c_2, \cdots, c_i, \cdots, c_n$

-- N_c：語料庫中之所有字元數；N_w：語料庫中之所有詞數

-- $f(.)$：頻率

本研究採用中央研究院詞庫小組所開發之中文斷詞系統進行文件前置處理[1]，由於斷詞系統通常只能處理一般詞彙，在處理特定領域詞彙時，可能會出現不適當的斷詞結果（說明：此處所謂的不適當，可能是斷詞錯誤或是斷詞結果所產出之詞彙，無法代表此領域的完整概念），需採用某些輔助方法做進一步的處理。交互信息常用來計算詞彙間之關聯強度，是自然語言處理領域中相當成熟的技術（Church & Hanks, 1990），同樣的概念也被應用於中文斷詞中（Yang, et al., 2000）。本研究透過計算文件中N元詞之交互信息找出公司治理領域之重要詞彙列表，參考此列表配合斷詞系統所產出之斷詞結果，可以協助我們決定是否針對斷詞結果進行後續的處理（例如：修正某些錯誤斷詞，或是將某些詞彙結合爲一個更完整的概念詞）。

首先，我們可以從語料庫中計算出所有字元的數目，表示爲N_c，因此可將交互信息計算公式中對數的部分重寫爲$\frac{f(w)}{N_w} \Big/ \frac{f(c_1)}{N_c} \times \frac{f(c_2)}{N_c} \times \ldots \times \frac{f(c_n)}{N_c}$。由於特定詞$w$是由$c_1, c_2, \cdots, c_n$字元所構成（例如：$w$爲「公司治理」，則$c_1$、$c_2$、$c_3$、$c_4$分別爲「公」、「司」、

1. 英文的每個單字都可以成為詞，具有自己的意義，且每個詞間都有明顯的空白作為分隔，因此沒有所謂斷詞的困擾。相反的，中文在書寫時，詞與詞之間無空白作為區分，因此需要先做斷詞處理。

「治」、「理」），$c_1, c_2, \cdots\cdots, c_n$等字元之出現頻率分別爲$f(c_1)$、$f(c_2)\cdots\cdots f(c_n)$，因此特定詞$w$中各字元的發生機率可表示爲$\frac{f(c_1)}{N_c}$、$\frac{f(c_2)}{N_c}\cdots\cdots\frac{f(c_n)}{N_c}$。假設各字元彼此間相互獨立（彼此間無任何關聯），其發生機率爲$\frac{f(c_1)}{N_c}\times\frac{f(c_2)}{N_c}\times\ldots\times\frac{f(c_n)}{N_c}$。同樣的，我們可以從語料庫中計算出所有詞的數目，表示爲N_w。$f(w)$代表特定詞w於語料庫中的發生頻率，因此，特定詞w的機率就可寫爲。若此對數值越大，表示此特定詞w的出現機率大於構成特定詞的字元彼此獨立時的機率，因此，特定詞的字元間並非完全獨立，而是彼此間互有關聯。依此對數值搭配詞頻排序，將可進行關鍵詞之篩選。

㈢建立領域詞彙表

基於篩選出之關鍵詞，再利用全文搜尋方式，先找到關鍵詞出現之文本位置，再進一步地由領域專家以人工方式找出關鍵詞之定義與解釋。建置詞彙表所儲存的格式主要參考Tzeng、Liou與Sun（2007）的詞彙表架構，將關鍵詞定義與釋義本身做切割。

二、文本處理並產生詞庫

針對前一步驟所建構之詞彙表解釋內容，利用中央研究院中文斷詞系統，進行中文斷詞，並對斷詞結果做適當的修正，產生本研究之公司治理詞庫。進行中文文本處理的重要工作，主要有詞法分析與句法分析。詞法分析針對文件內容做詞與詞間的切割（斷詞），句法分析的作用則爲瞭解詞與詞之間的關係爲何。目前自然語言處理的研究在近幾年來有許多重要的突破，已可自動化的針對上述兩項工作進行剖析。利用自動化解析方式建置相關性詞庫，具有快速與節省精力之優點，可有效率地建置各領域之相關性詞庫，並分析句子中各詞彙間之關係。本研究使用中研院中文斷詞與剖析系統，進行詞法分析與句法分析。以「風險評估屬於內部控制的組成要素之一」一句爲例，詞法分析可將原句切分組成此句之個別詞彙「風險（Na） 評估（Nv） 屬於（VG） 內部（Ncd） 控制（VC） 的（DE） 組成（Nv） 要素（Na） 之一（Nc）」，而句法分析則可進一步展示該句之句型結構以利後續分析（如圖2）。

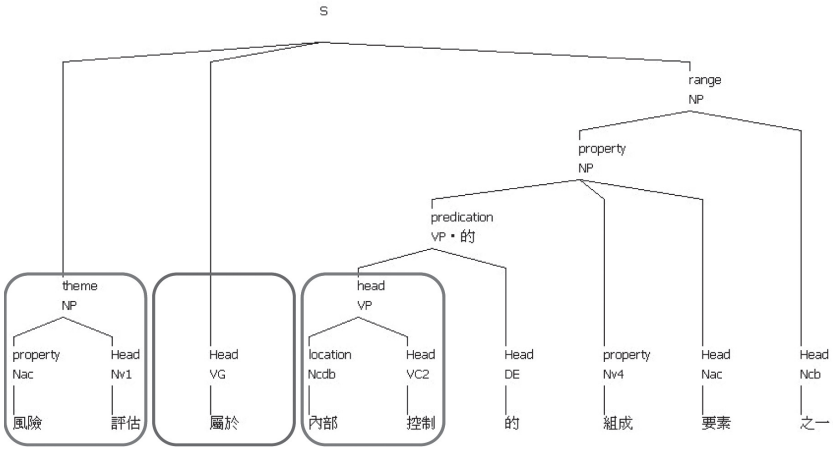

圖2　句法分析結果（部分）

三、語意網路建構

利用斷詞系統產生詞庫後，本研究基於詞庫所建立之向量空間模型（Vector Space Model, VSM; Salton, Fox, & Wu, 1983），使用TFIDF（term frequency inverse document frequency）模式，計算詞項定義中特徵詞權重，以產生詞項特徵向量。TFIDF中之TF（term frequency）代表特徵詞出現之頻率，特徵詞出現頻率越高一般會認爲該特徵詞越重要，因此權重也較高；而IDF（inverse document frequency）與特徵詞出現於不同文件（或定義）之數量成反比，特徵詞越普遍出現於文件中時，表示該特徵詞越不具備區別力。因此，IDF越高表示特徵詞可能僅出現於特定之文件中，透過此特徵詞可區別出特定類型之文件，權重亦較高。本研究爲詞項建立特徵向量，進而計算詞項間之相似度，此三項步驟分述如下：

（一）建立詞彙向量空間：詞彙向量空間由詞庫中之所有詞彙集合所構成，爲一多維度空間。

（二）計算各詞項之特徵向量：所謂的詞項，指詞彙表中具備解釋之詞彙，因此，詞項包含了詞彙及該詞彙之解釋。計算各詞項權重值之計算方式有許多種，本研究採用TFIDF加權模式。透過此模式可計算出各詞項於向量空間中各維度之權重，進而形成代表該詞項之特徵向量，該向量代表詞項於向量空間中所處之位置。

（三）計算詞項間之相似度：本研究利用cosine係數計算特徵向量間之夾角，以代表特徵向量間之相似度。Cosine係數值介於0至1之間，當

Cosine係數值越接近1，代表兩向量夾角越小，兩詞項之相似度越高；反之，則兩詞項的相似度越低。

經過語意相似度計算後，系統會產生上述之語意網路儲存結構，但本步驟尚未擷取兩詞彙關係之標示符號，關係標示符號包含從屬、同義、反義等，將會於語意結構分析步驟擷取出來。

四、語意結構分析

完成了語意網路後，將基於詞項之特徵向量進行分群（clustering）。分群演算法將文件集合中相似的文件區分爲數個子集合，透過該演算法，可得到各群內部相似性高，但群間之相似度低之結果。分群爲無監督式學習（unsupervised learning），因此，計算過程不需要專家參與。

本研究用於計算群間距離之測度爲歐氏距離。首先，如圖3所示，我們採用K-means演算法針對所有詞項進行平面分群，形成多個詞項子集合，並計算出各詞項子集合之重心（centroid）作爲代表該詞項子集合之特徵向量。透過計算各子群集中心之相似度，可進一步實行階層分群，將最相似的詞項子集合以樹狀圖（dendrogram）的階層型式逐步聚合在一起。分群演算法的導入，將可降低人類專家進行語意結構分析時的資訊負載量，俟階層大致分群後，再由人類專家參考知網關係義原中的概念階層關係、同義關係及反義關係來建立詞彙間之關係。

本研究亦藉由半自動化之詞彙共現性分析及句型分析，以協助確立概念階層。以下針對三種分析模式進行討論：

（一）詞彙共現性分析：尋找概念階層關係（亦即A是一種B），本研究根據Sanderson與Croft（1999）提出詞彙共現性之看法，認爲有兩個詞彙t_1與t_2，t_1若沒有出現在所有t_2出現的內容中時，則詞彙t_1較詞彙t_2特

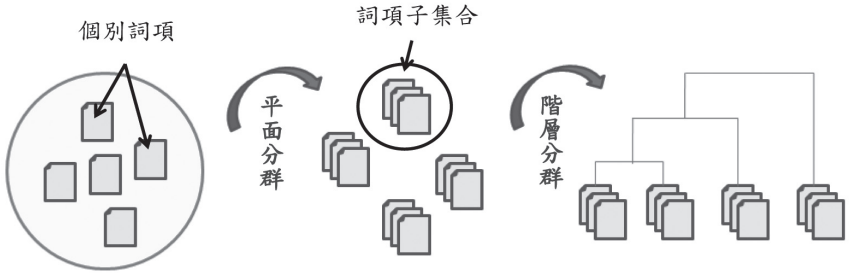

圖3　分群演算法實施步驟

殊。基於此概念，本研究將前述之「內容」限制在句子，並應用於詞彙進行概念階層分析，以下提出兩例：企業目標與營運目標並無出現在所有「目標」出現之句子中，亦即當提到營運目標時，並無出現在另一個目標出現之句子中（企業目標），而當提到企業目標時，並無出現在另一個目標出現之句子中（營運目標），但兩詞彙出現之句子中同時出現「目標」一詞。因此，可歸納出企業目標與營運目標較「目標」一詞特殊，其間可建立概念階層關係。而風險環境、風險分攤、風險規避，我們也可找出與「風險」之概念階層。我們必須強調，利用詞彙共現性分析並非絕對可以找出概念階層關係，該分析僅提供專家快速篩選概念階層關係，因此，我們仍須透過觀察詞彙間之相似度或透過專家檢核才可確立其間之關係。

（二）相似度計算：在此我們必須強調，共有詞的分析結果並非絕對正確，例如，當我們將此方法應用到投資人、股東及經理人上。由於投資人與經理人具備共有詞「人」，依照上述詞彙共現性分析應分爲同一類，然而，投資人與股東雖不具共有詞，依據定義卻更應該分屬同一類之下。此時，我們可以透過比較各詞間之相似度，再進一步判斷其關係。

（三）句型分析：除上述詞彙共現性分析外，我們發現某些句型確實能協助我們確定詞彙間之關係，因此，本研究亦透過系統，針對詞彙之定義進行句型分析。透過句型分析，我們可以找出從屬、同義、反義等關係（詳見第肆節）。

本階段將會利用語意結構分析之方法，由詞彙定義終結取出概念間之關係，而概念階層關係將成爲建構本體概念分類之主要依據。

五、雛型本體產出

任何本體學習方法，最終目標均爲產出正規的本體。因此在本步驟中，本研究將進一步地把語意網路轉換爲以OWL語言所表達之公司治理本體知識。由於本研究的主要目的是提出本體學習的方法，並非建置完整的領域本體，因此在本階段並不嘗試去詳細建置本體中的概念屬性，僅以概念階層的表達爲主。

六、雛型本體評估

本體建置主要是將領域知識予以模型化，通常須經由開發者不斷的修正，才能達到完善且可應用的本體，Noy等人（2001）及Chi（2007）都曾指出：本體建置是一種藝術更甚於科學的工作（“Art more than science”），而現行的知識塑模方法也僅是提供原則性步驟，並沒有客

觀或制式的績效指標；另外，本體建置通常以「問題解決」爲塑模的目標，績效評估也以達成目標的程度來衡量（Corcho, Fernández-López, & Gómez-Pérez, 2003），但因爲本研究的結果是產出可供後續再修正的「雛型本體」，並非最終的應用本體，因此在研究結果的評量上，我們設計適當問卷，針對雛型本體的詞彙、概念、階層結構、及建置方法等，進行妥適程度之評估，並尋求五名具有公司治理知識之專家進行問卷調查，以提供本研究在建置方法之回饋。

肆、建構公司治理本體之應用

由於公司治理研究在會計領域占有一席之地，且公司治理爲近來相當熱門之研究議題，無論是財金或管理領域都對公司治理有許多討論，也都以各自領域的角度對公司治理範圍做出具差異性的定義與解釋，使具相同意涵之關鍵字可能有不同的名稱，或對關鍵字之解釋範圍有所差異，而這些差異乃源於公司治理爲一門由人所闡釋的社會科學。因此，本文以公司治理議題作爲研究基礎，期望本文所建置的公司治理本體能達到知識分享與共用的效果。首先對於公司治理領域做一介紹，隨後則依研究方法表1中提出的AGOE六個步驟，整理說明本研究實作之結果。

公司治理領域分析如下（行政院經濟建設委員會，2003）：

一、公司治理的定義：根據經濟合作暨發展組織（OECD）之定義，公司治理是泛指規範企業、其管理階層、董事（監）會、股東與其他利害關係人（如勞工、債權人、客戶、社區與政府）之間關係的架構，並可透過這種機制釐定公司的營運目標，以及落實該等目標的達成與營運績效的監測。我國中華公司治理協會亦提出公司治理之定義爲：一種指導及管理並落實公司經營者責任的機制與過程，在兼顧其他利害關係人利益下，藉由加強公司績效，以保障股東權益。

二、公司治理涵蓋的範疇：而公司治理涵蓋之範疇，狹義來說，公司治理的範疇係指「公司監理」，尤其上市、上櫃公司的監理，重點涵蓋公司經營者之責任，公司股東之權利義務，以及公司營運之防弊措施、內部控制與企業風險管理等；涉及之規範包括公司法、證券交易法、會計準則等。廣義來說，公司治理的範疇則除公司監理之外，還包括相關之市場機制、企業併購、特定組織（如管制機關、公營事業等）之治理、機構投資人機能、破產與重整機制、財經法之執行與改革等。

爲使公司治理的改革能充分發揮成效，我國政府推動改革公司治理係以廣義範疇爲著眼。

一、詞彙表建置——公司治理

㈠文本蒐集與處理

要建置完整的公司治理領域詞彙表，首先要取得具代表性的素材，國內專門探討公司治理的中文專書並不多見，目前各大圖書館之館藏多數均爲應用方面的書籍或是翻譯書，依本研究取得公司治理領域詞彙定義之需求，應以中文教科書或相關期刊論文爲主要選擇，經搜尋後，本文進行關鍵詞擷取的書籍包含公司治理（易明秋，2003）及企業風險管理－整合架構（馬秀如，2006），期刊則以會計研究月刊中有關公司治理等議題的文章共4篇爲主，並於全國博碩士論文網下載公司治理議題論文摘要共150篇。

㈡關鍵詞篩選——詞頻與交互信息

接著，利用詞頻與交互信息等文字探勘方法，從文獻中找出233個重要關鍵詞。

㈢建立領域詞彙表

最後，基於篩選出之關鍵詞，研究者會再進行以下步驟，並參考現有詞典之儲存架構，以建立公司治理詞彙表：

1. 全文搜尋詞彙定義：利用全文搜尋方式，找到關鍵詞出現之文本位置，再以人工方式找出關鍵詞之定義與解釋，例如，從以企業風險管理爲主題的文章中，找出其重要組成要素風險評估及風險回應，再取得原始文件中兩詞彙的特性描述，找出兩者的關鍵詞釋義。所謂的定義釋義，並非只摘錄該關鍵詞的解釋，尚包括了該關鍵詞之組成要素及基本要件等敘述。
2. 同詞異義的處理：爲了避免在建置詞彙表時參雜主觀意識造成文件呈現的偏差，本文並不對各關鍵字的定義加以修改，當某重要關鍵詞出現不同定義時（同詞異義），研究者會將不同釋義併成該詞彙的共同解釋。
3. 異詞同義的處理：另外，部分不同的關鍵詞常有相同的概念及相近的釋義（異詞同義），主要是因爲公司治理相關議題是由人所進行闡釋的一門社會科學。因此，在不同的發展階段或在不同的國家常用不同的詞語形容相同的概念。爲了減低主觀上之判斷問

題，本文在進行詞彙定義時，乃是以上述文獻中各研究者最常使用的詞語來描述。圖4即爲詞彙表建置結果之一例。

二、文本處理並產生詞庫

本研究之詞庫建置過程主要是將建置完成的詞彙表（包含重要關鍵詞及其釋義），輸入中研院所開發的中文斷詞系統，協助建置公司治理領域詞庫。但由於中研院斷詞系統尚未能對部分領域詞彙做出正確的斷詞處理，使斷詞結果可能會與詞彙表原先定義不一致。因此，本研究在進行中文斷詞後，會將斷詞結果與本研究所制訂之詞彙表中的詞項進行比對，例如若中文斷詞結果將「風險」與「目標」分開，而詞項中包含「風險目標」此詞彙，本研究會將「風險」與「目標」合併爲「風險目標」，以期修正中文斷詞系統斷詞之結果。中研院斷詞系統在本研究中主要針對詞項之解釋部分進行斷詞，以找出具備定義之詞項，以及隱藏在詞彙釋義中的重要詞彙。

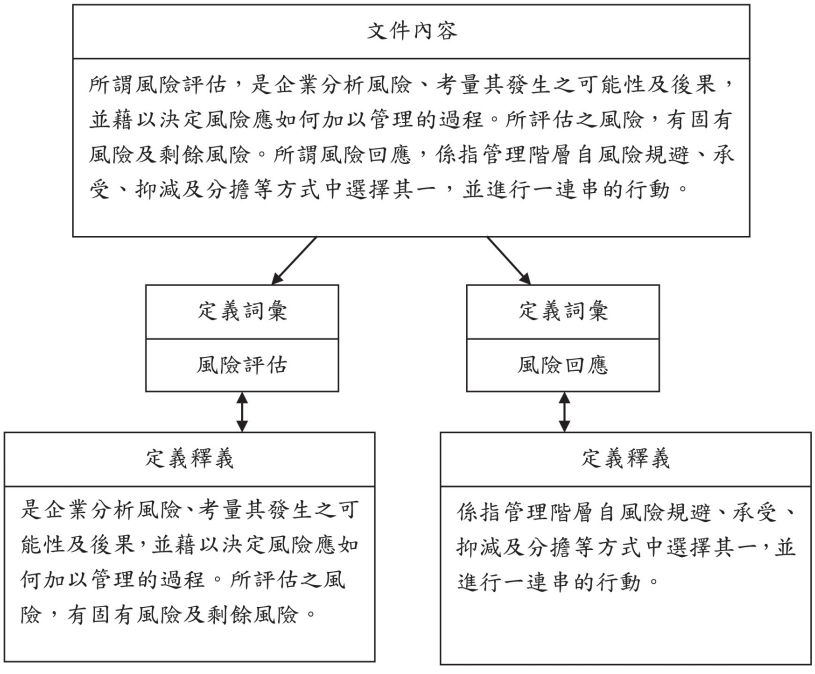

圖4　定義詞彙與釋義

資料來源：「企業風險管理」，馬秀如、賴森本、阮中祺、李美雀，2005，*會計研究月刊*，*238*，頁28-78。

三、語意網路建構

將詞彙表輸入斷詞系統產生詞庫後，即可使用TFIDF建立詞項特徵向量，並計算詞項間之相似度，結果分述如下：

（一）計算各詞項之特徵向量：本研究之TFIDF權重計算定義如下：

$tfidf_{ij} = tf_{ij} * idf_j = tf_{ij} * \log_2 (n / df_j)$

-- $tfidf_{ij}$：詞j在詞彙定義i裡的權重

-- tf_{ij}：詞j在詞彙定義i裡的出現頻率

-- df_j：包含詞j的詞彙定義數

-- idf_{ij}：與df_j成反比，idf_{ij}越大表示該詞越能代表該詞彙定義

-- n：詞彙定義總數

經過運算後，本研究相似度計算總連結筆數n爲20,301筆。表2則列出「內部控制」特徵詞向量部分維度。

表2　特徵詞向量（部分）

維度	TFIDF
資產	15.8462
合理	12.4442
記錄	11.2934
會計	10.1427
確信	9.23024
內部	9.10654

（二）計算詞項間之相似度：詞彙關係計算完成後，將可依詞項間之相似度建構語意網路；本研究將語意網路定義爲如下之結構（Tzeng, et al., 2007）：

$N = (T, R, Rel, M)$

此處

T：詞彙集合

$R：T \times T$

I：關係標示符號（identifier）之集合

$Rel：T \times T \rightarrow I$，將兩詞彙對應至其關係標示符號之函數

$M：T \times T \rightarrow [0, 1]$，將兩詞彙對應至其語意相似度之cosine函數

此處$T = \{t_1, t_2, \cdots t_i, \cdots, t_n\}$，其中$t_i$表單一詞彙，T爲詞彙所構成之集

合；而T × T爲兩詞彙集合之卡氏積（cartesian product），即T × T={(t_1, t_1), (t_1, t_2), ⋯, (t_1, t_n),⋯,(t_n, t_n)}；I表示關係標示符號集合，I = {i_1, i_2, ⋯ i_k, ⋯, i_m }；Rel：T × T → I說明函數Rel之定義域（domain）爲T × T，值域爲I，透過函數Rel，可將兩詞彙序對(t_i, t_j)對應至關係標示符號i_k，M：T × T → [0, 1]亦同。

本研究所訂之相似度篩選條件，設定門檻值爲大於0.2者；原本總連結筆數爲20,301筆，經篩選後僅餘375筆。本文選取門檻值須大於0.2的主要原因爲：當門檻值小於0.2時，具相關性的詞彙重複出現的頻率偏高，若將之進行相似度的連結，反而與本研究以提高本體建置效率之目的相違背。因此，在針對詞彙關係計算結果進行解讀及判斷後，本研究將門檻值設定爲大於0.2以上者。

將233個詞彙（具有375筆關聯）做連結後，並標示出詞彙間的關聯程度，形成語意網路。由於詞彙筆數數量龐大，若以單一主題爲核心呈現，會顯得過於複雜，且因版面限制，因此將詞彙網路關係圖分別以公司治理、內部控制、企業風險管理爲中心分別做呈現，如圖5所示。

四、語意結構分析

本研究的實作目的主要爲呈現利用文字探勘技術所建置的公司治理議題之本體知識，因此，須針對複雜的語意網路進行分析及簡化工作，以便從中找出本體階層關係。由於將特定領域予以知識化（模型化），是一項耗時費力的工作，因此藉由文字分析技術的協助（例如Machine Learning、Text Mining等），可由大量的文件中萃取領域的可用資訊。但這類稱爲KDT（knowledge discovery from text）的應用，著重於分析上的時效，例如快速提供文字分群，但分群後的架構、從屬關係或與解題的關聯等，仍須仰賴專家依據設定的「問題解決」加以人工調整，領域專家將各分群視爲描述「公司治理」的子概念，並進行概念性架構的發展建置。

從圖5中可發現，此圖以公司治理爲核心向外做延伸，且可發現結構相似之概念群，甚至有些概念會在各圖中重複的出現，這些特徵皆與本研究所進行的本體知識的建置具有重要的關聯性，將在進行本體知識建置時再詳細說明。至於概念間的關聯程度值，經由本研究檢視關聯程度後發現，利用文字探勘技術計算出的相似度越高者，事實上其關聯程度未必相對較高，而關聯程度相對較低者，也並不表示實際上的關聯程度較低。關聯程度不一致的原因有二：（一）本研究透過詞彙表中概念定義（或解釋） 作爲概念內涵，並透過概念定義相似程度之比較作爲關聯

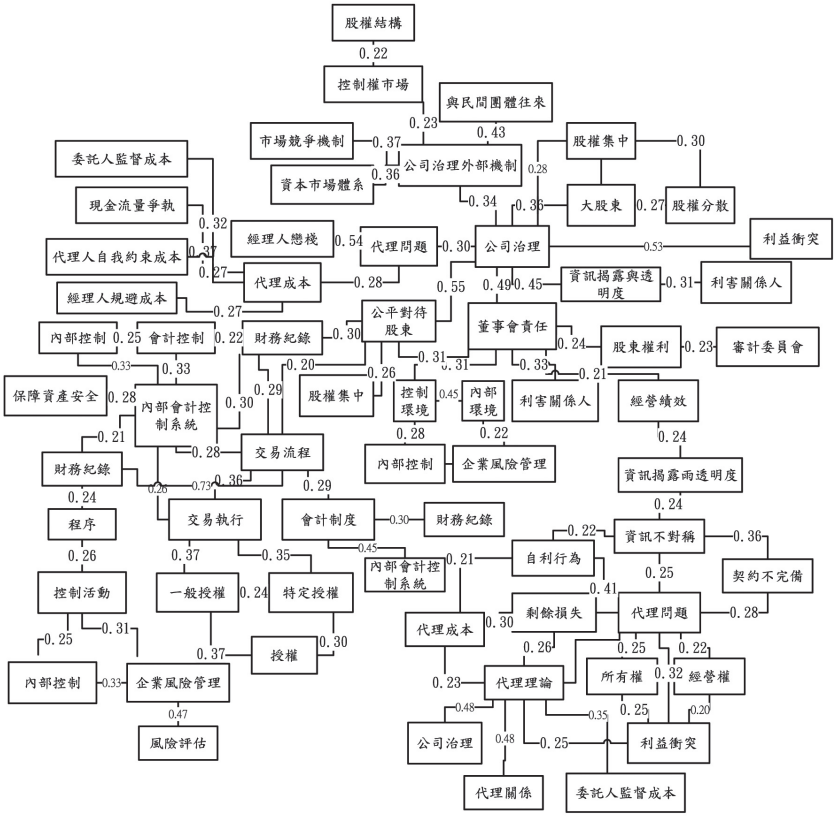

圖5　公司治理語意網路（部分）

程度計算之基礎。概念之定義是由公司治理相關文獻或論文整理而成，其中可能有看法或用語不一致或不完整之情形；（二）本研究單純由詞頻向量計算相似度，此計算方法忽視詞彙間之文法關係、實務習慣等，因此發生關聯程度不一致。在建構公司治理本體之應用中，本研究並未全然接受關聯程度值的結果，僅將之視爲建構本體時的輔助工具，作爲連結的參考，而非本研究在連結時的重要依據。因此，以下展示之本體階層關係，將不顯示相似度值，避免概念連結時造成資訊的錯誤傳達。

爲了進行分析及簡化工作，本研究所提出的創新方法爲：利用平面分群從語意網路中辨識結構相似的概念群，再針對各概念群進行階層分群。在上述分群的過程中，本研究針對平面分群後各群之中心概念，提出幾個重要問題，以便瞭解概念群之所以互相連結的始末，進而利用階

層分群以及語意結構分析中的上下關係、同義關係與反義關係，進行概念的整合並找出概念間的基本階層關係，最後組織各概念群以呈現簡單結構的公司治理本體知識。將建置過程具體化後，提出四個方法步驟協助建構本體知識，如圖6所示，並針對各步驟做詳細的說明如下：

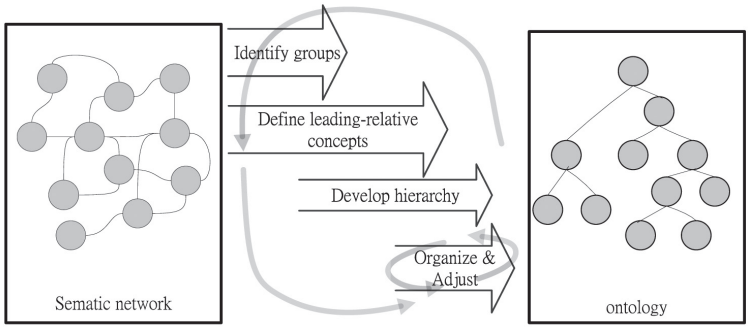

圖6　本體知識建置步驟

資料來源：本研究自行發展。

（一）辨識結構相似概念群（identify groups）：在本步驟中，本研究將導入「分群演算法」，先針對所有詞項進行平面分群，形成多個詞項子集合，透過計算各子群集中心之相似度，進一步實行階層分群。分群演算法實施的結果如圖7，由X軸可看出共區分爲27群，Y軸則爲27個不同概念群間相互的距離。圖8則是兩個語意網路的釋例，該圖以公司治理語意網路中的子群集中心：「企業內部團體」（圖7編號8），以及另一子群集中心：「與企業往來團體」（編號25）爲例，由「分群演算法」的結果可推知，「企業內部團體」和「與企業往來團體」具有高度相似度，推測在本體結構中應該相距不遠，因此，將兩者一起進行比較分析，以便加強建構本體知識的效率。而子群集中心的確認，主要是以「分群演算法」的計算結果爲主，但仍需加入少許人工判斷，以達到最佳分群效果：

1. 更換中心概念1組，加強該群集解釋能力，如：「公開發行公司建立內部控制制度實施要點」（編號4）以「內部控制參考依據」概念取代。
2. 相似概念合併4組，如：將子群集中心「會計控制」（編號13）合併至「控制活動種類」（編號11），以及「剩餘風險」（編號24）合併至「企業風險管理」（編號22）等。

3. 刪去無代表性群集5組。

換言之，沿用分析結果之概念群集及其中心者有17組，再加上2組群集相似度高而合併中心概念者，共19組，整體效率達19 / 27 = 0.703。初步完成之語意網路結構如圖8所示。

（二）調整「分群演算法」結果——定義「中心—相關」概念（define “leading-relative” concepts）：辨識結構相似的概念群與其群集中心之目的，主要是爲了加強建構本體的效率，並不能完全驗證概念間的階層關係，有些概念群間甚至沒有相似的結構但卻具關聯性。因此，需要針對「分群演算法」之結果做出人工調整，本文利用關鍵性問題的提出，針對個別概念群中的中心—相關概念作定義，以釐清中心概念與相關概念間的關係。本研究所提出的關鍵性問題如下：

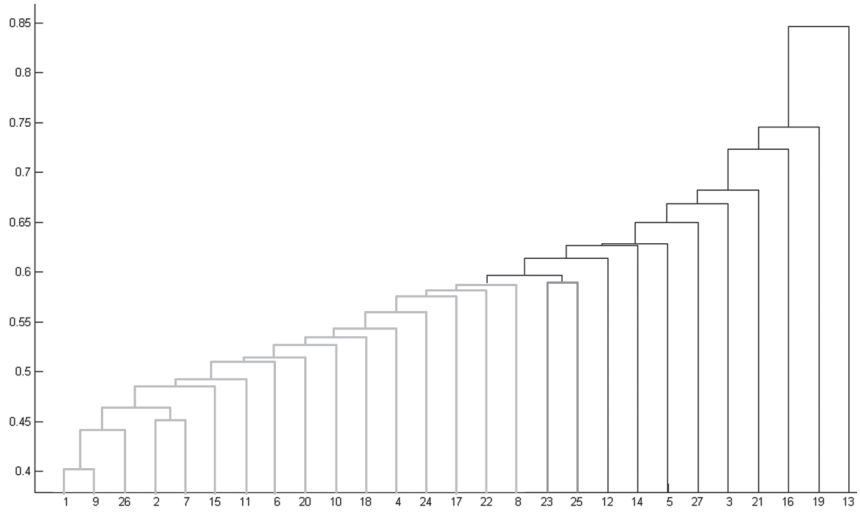

圖7　分群演算法實施結果

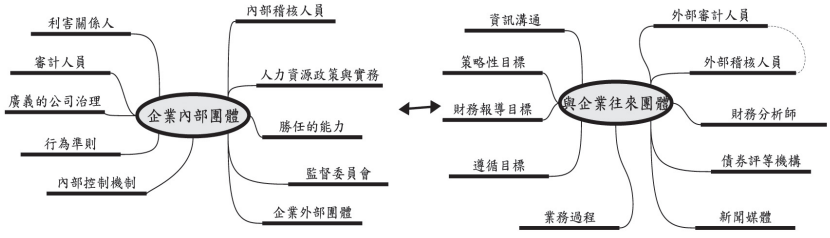

圖8　語意網路結構釋例圖

1. 辨識欲達成管理目標爲何？
2. 辨識欲達成管理目標所參與之團體？
3. 辨識欲達到管理目標應進行之活動內容？

就「企業內部團體」和「與企業往來團體」而言，所欲達成的管理目標爲維持公司治理及營運、財務報導、法令遵循之正常運作。所參與之團體包括內部稽核人員、外部稽核人員、財務分析師、債券評等機構、新聞媒體等，而針對目標所進行之公司活動包括內部團體之行爲準則、內部控制機制、人力資源政策與實務、業務過程、資訊表達與溝通（如圖9）。

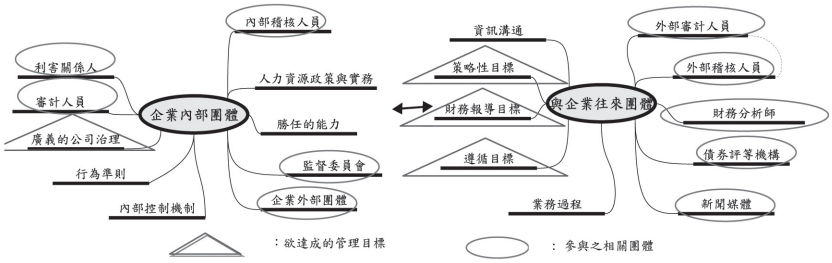

圖9　與企業內部團體及企業往來團體具關係的概念群

（三）發展概念階層關係（develop concept hierarchy）：利用關鍵性問答找出與中心概念有關聯性的相關概念後，本文透過語意結構分析階段所建構之系統輔助，並依據知網的關係義原概念（董振東、董強、郝長伶，2007），找出概念間的概念階層、從屬、同義等關係，並將具相同定義的概念擇一做表達。爲求關係發展之正確性及完整性，在運用系統發展出初步之結構後，專家將依據詞彙定義做進一步的調整與發展。

1. 詞彙共現性分析：概念階層關係（以共有詞「會」爲例）：我們使用本研究所建立之詞彙共現性分析工具，先針對共有詞「會」篩選出7個複合詞，經過檢視後，發現包含共有詞「會」之複合詞可分爲三類，並進一步選擇「委員會」爲共有詞進行篩選，最後可建構出概念階層關係如圖10。
2. 句型分析（圖11）：

   從屬關係：A屬於B

   -- 風險評估**屬於**內部控制的組成要素之一。

   -- 內部環境**屬於**企業風險管理的組成要素之一。

   同義關係：A又稱爲B

-- 內部審計人員又稱為內部稽核人員。
-- 查核人員又稱為外部審計人員。
-- 外部董事又稱為獨立董事。
反義關係：A是B的相反
-- 內部控制有效是內部控制無效的相反。
-- 股權集中視股權分散的相反。

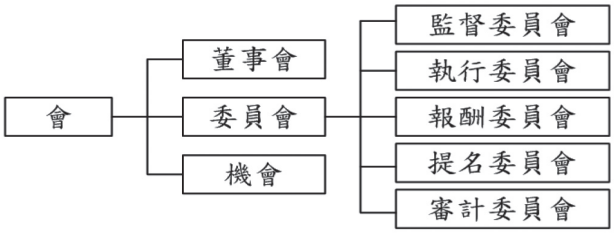

圖10　包含共有詞「會」之概念階層關係

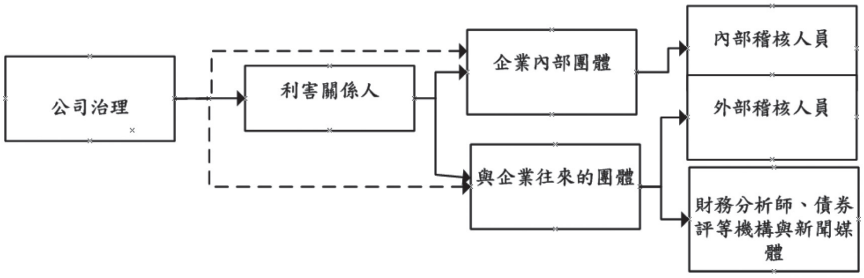

圖11　企業內部團體與與企業往來團體之上下階層關係

（四）組織與調整（organize and adjust）：經過結構相似概念群的辨識提升本體建置效率、提出重要問題找出具關聯性的概念、並利用語意結構分析概念階層關係後，本步驟進行概念階層的整合與調整，將多個分散的本體知識組織起來，使階層架構擴大，並在整合過程中調整概念間的關係。若組織過程中遇上問題或本體階層不夠明確，就重新在語意網路中找出其他相關的概念群進行萃煉與修正，經過建置與調整後，呈現簡單結構的公司治理本體知識（圖12）。以下舉出兩項調整之案例：

1. 從屬關係（以企業內部團體與與企業往來團體爲例，如圖11）：
擁有一個品質優良企業內部團體之目的是維護公司治理水準，因

此，公司治理與企業內部團體具從屬關係。而與企業往來團體維持良好關係之目的亦是維護公司治理水準，因此，與企業往來團體和公司治理具從屬關係。同時企業內部團體及與企業往來團體皆包含公司治理之利害關係人概念，因此，企業內部團體及與企業往來團體與利害關係人具從屬關係。

2. 同義關係：外部審計人員與外部稽核人員名稱雖不一樣，事實上具相同意義，因此，在建構本體時，將以外部稽核人員表達呈現。

五、雛型本體產出

依據圖12的結果，本研究進一步地把語意網路轉換爲以OWL語言所表達之公司治理本體知識，完成知識本體之正式雛型。

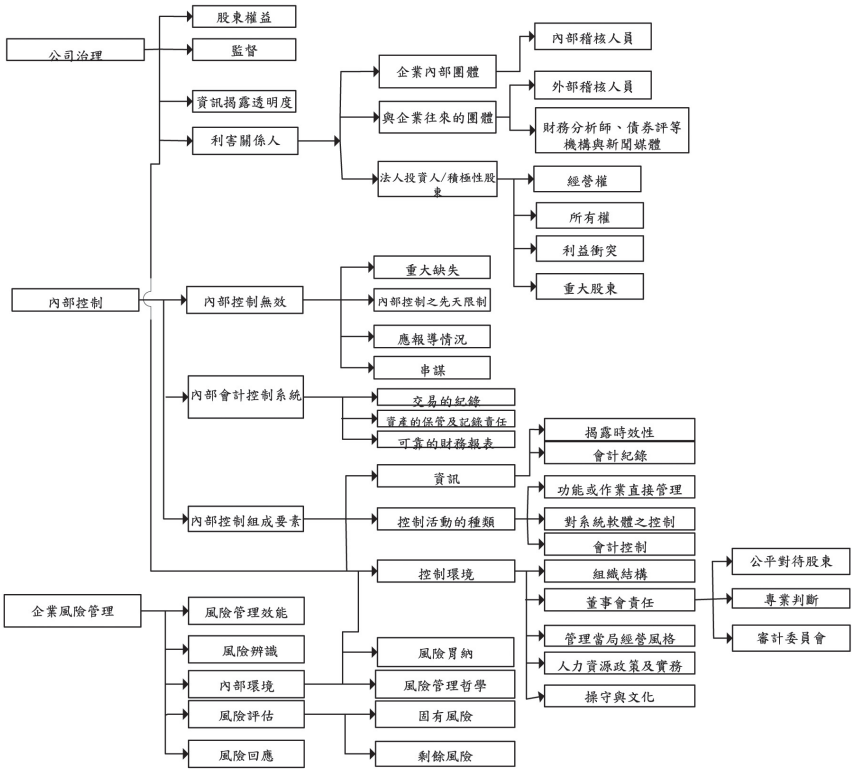

圖12　公司治理議題本體知識（部分）

六、雛型本體評估

本研究採用問卷方式進行「雛型本體」之評估，問卷內容針對本體的詞彙、概念、階層結構、及建置方法等設計（請參考附錄），我們委由五名領域專家進行檢驗本研究之結果，領域專家包含：任教於國內大專教師兩名（均有超過十年以上之會計教學經驗）、會計師三名（均有超過五年以上之會計師事務所實務經驗），調查結果彙整如表3，由於召回率（recall rate）不易衡量，因此以五名專家個別的正確率（precision rate）或同意比例爲指標。評估結果顯示：本研究提出的建置方法，五名專家對於該雛型本體的同意題數比例都非常高（多數均超過九成），據此，未來領域專家應可以此初始的本體爲發展基礎，再修正爲應用本體。

表3　公司治理本體雛型評估結果

	同意題數	不同意題數
專家甲	60	1
專家乙	59	2
專家丙	54	7
專家丁	60	1
專家戊	57	4

伍、結論與建議

一、研究結論

本體建置是針對特定領域，以概念、階層化結構、及關係等呈現，是一種探討領域是如何組成的方法論，而發展一個實務應用的本體，通常也須考量待解決問題或任務的形式，但大致仍循著知識擷取、知識塑模、知識表達、再修正等步驟進行，因爲參與開發的各類角色眾多，開發過程也相對冗長。本研究爲簡化人爲的作業程序，因而在知識擷取階段及塑模的部分階段，改以文本資料作爲建模的素材來源，期能由機器產生一個初始的「雛型本體」，再交由領域專家持續修正爲應用本體。近年來，雖有學者也是以文件資料爲知識擷取對象，但普遍缺乏一致性的方法步驟，因此不易爲其他領域所參考應用。另一方面，有

別於其他以文本資料爲主的本體建構，本研究提出的方法，特別強調建置領域詞典（或詞彙表）的重要性，由於詞典（或詞彙表）在外顯知識（explicit）的定義具有固定形式，因此當進行詞彙篩選與句型分析時，較能有效提升知識之召回率（Brewster, et al., 2003; Hearst, 1992）。另外，本研究的範例－公司治理領域，目前並沒有適合之領域詞典（或詞彙表），因此本研究也一併說明如何建立領域詞彙表。

本研究主要的貢獻是將A-G方法，改良爲可執行的本體建置程序，這種稱爲AGOE的方法，主要將A-G中的原則性描述，具體以循序步驟及配套技術來實踐。本研究在第肆節中，即利用此AGOE方法來發展「公司治理」的知識本體，由於這個本體主要是以階層化呈現其概念架構，它是初始的雛型本體而不是最終的應用本體，因此本研究的評估主要針對此概念架構的詞彙、概念、階層結構、及建置方法等，檢驗是否達到快速建置「雛型」本體之目的。本研究最後採用問卷調查，並對五名領域專家進行調查，平均同意比例已達九成以上，初步顯示領域專家應可根據此雛型本體爲基礎，再繼續修正爲最終之應用本體。

二、後續研究建議

本研究提出AGOE的本體開發程序，它能協助專家快速建置本體的初始架構，有效縮短本體的開發時程。三項後續的研究建議如下：

（一）本體之屬性擷取：除了利用相似度針對詞彙進行歸類及運用詞彙共現性、句型分析擷取詞彙間關係，以建構本體之基本結構外，在使用TFIDF分解詞彙定義以計算詞彙間相似度的過程中，可以發現敘述該詞彙之定義已被分解爲具備不同權重之關鍵詞，這些關鍵詞有機會進一步轉換爲該定義所敘述概念（詞彙）之屬性。因此，後續研究者可針對描述概念之關鍵詞轉換爲屬性方面，做進一步的研究。

（二）整合正規概念分析法：本研究提出之語意結構分析，主要協助建立概念之結構，後續研究者可使用正規概念分析法，用於概念之發掘與確認，以達到互爲驗證之作用。

（三）提升公司治理本體知識的外部效度：本研究以會計觀點來建置本體知識，然而，公司治理範圍甚廣，包含了財金、管理等領域，若只以會計領域所定義的公司治理知識作爲研究基礎，將會限制公司治理本體的效度。因此，後續研究者可蒐集不同領域的知識，擴大公司治理本體的可應用性及可參考性。

參考文獻

中央研究院（2009）。*中文斷詞系統*。上網日期：2009年7月1日，檢自：http://ckipsvr.iis.sinica.edu.tw/

行政院經濟建設委員會（2003）。*強化公司治理政策綱領暨行動方案*。臺北市：作者。

周濟群、連子杰（2011）。運用文字探勘與XBRL技術提升企業資訊擷取與整合效益之研究。*當代會計*，*12*(1)，85-114。

易明秋（2003）。*公司治理*。臺北市：弘智。

財團法人中華民國證券暨期貨市場發展基金會（2011）。*臺灣公司治理簡介*。臺北市：作者。

馬秀如（2006）。*企業風險管理：整合架構*。臺北市：財團法人會計研究發展基金會。

馬秀如、賴森本、阮中祺、李美雀（2005）。企業風險管理。*會計研究月刊*，*238*，28-78。

戚玉樑、蔡明宏（2007）。以文件爲對象的概念萃取程序建立知識本體的雛型架構。*資訊管理學報*，*14*(3)，47-66。

董振東、董強、郝長伶（2007）。知網的理論發現。*中文信息學報*，*21*(4)，3-9。

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise？ The information content of internet stock message boards. *The Journal of Finance*, *59*, 1259-1294.

Aussenac-Gilles, N., Biébow, B., & Szulman, S. (2000, October). *Revisiting ontology design: A method based on corpus analysis*. Paper presented at the 12th International Conference on Knowledge Engineering and Knowledge Management, Juan-les-Pins, France.

Biébow, B., & Szulman, S. (1999, May). *Terminae: A linguistic-based tool for the building of a domain ontology*. Paper presented at the 11th European Workshop on Knowledge Acquisition, Modeling and Management, Dagstuhl Castle, Germany.

Brewster, C., Ciravegna, P., & Wilks, Y. (2003, August). *Background and foreground knowledge in dynamic ontology construction*. Paper presented at the SIGIR Semantic Web Workshop, Toronto, Canada.

Chi, Y.-L. (2007). Elicitation synergy of extracting conceptual tags and hierarchies in textual document. *Expert Systems with Applications*, *32*, 349-357.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22-29.

Cimiano, P. (2006). *Ontology learning and population from text: Algorithms, evaluation and applications*. New York: Springer.

Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point？ *Data & Knowledge Engineering*, *46*, 41-64.

Downey, D., Etzioni, O., Soderland, S., & Weld, D. S. (2004, July). *Learning text patterns for web information extraction and assessment.* Paper presented at the American Association for Artificial Intelligence Workshop on Adaptive Text Extraction and Mining, San Jose, CA.

Engelberg, J. (2008, January). *Costly information processing: Evidence from earnings announcements*. Paper presented at American Finance Association Annual Meeting, San Francisco, CA.

Fernández-López, M., & Gómez-Pérez, A. (2002). Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, *17*, 129-156.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological engineering: With examples from the areas of knowledge management, e-commerce and the semantic web*. New York: Springer.

Grüninger, M., & Fox, M. (1995, August). *Methodology for the design and evaluation of ontologies*. Paper presented at IJCAI 1995, Workshop on Basic Ontological Issues in Knowledge Sharing, Quebec, Canada.

Hamon, T., & Nazarenko, A. (2001). Detection of synonymy links between terms: Experiment and results. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 185-208). Amsterdam: John Benjamins.

Hearst, M. A. (1992, August). *Automatic acquisition of hyponyms from large text corpora*. Paper presented at the Fourteenth International Conference on Computational Linguistics, Nantes, France.

Hindle, D. (1990, June). *Noun classification from predicate-argument structures*. Paper presented at the 28th annual meeting on Association for Computational Linguistics, Pittsburgh, PA.

Hiroko, F., Simmons, D. B., Newton, C. E., & Robert, E. S. (1997). Knowledge conceptualization tool. *IEEE Transactions on Knowledge and Data Engineering*, *9*, 209-220.

Lee, C. T., Huang, I., & Fang, K. T. (2010, July). *A study of building tax knowledge-based system: An ontological orientation -- Using cases under the national tax administration of central Taiwan province, ministry of finance*. Paper presented at Technology Management for Global Economic Growth, Phuket, Thailand.

Lin, D. (1998, August). *Automatic retrieval and clustering of similar words.* Paper presented at the 17th International Conference on Computational Linguistics, Quebec, Canada.

Liu, P., Hu, Y., Wang, X., & Liu, K. (2011, May). *A methodology for domain ontology construction in information science.* Paper presented at 2011 International Conference on E-Business and E-Government, Shanghai, China.

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, *16*(2), 72-79.

Morgan, A., Hirschman, L., Yeh, A., & Colosimo, M. (2003, July). *Gene name extraction using FlyBase resources*. Paper presented at the ACL Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan.

Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protege-2000. *IEEE Intelligent Systems*, *16*(2), 60-71.

Rajsiri, V., Lorré, J.-P., Bénaben, F., & Pingaud, H. (2010). Knowledge-based system for collaborative process specification. *Computers in Industry*, *61*, 161-175.

Reed, S., & Lenat, D. B. (2002, July). *Mapping ontologies into Cyc*. Paper presented at AAAI 2002 Conference Workshop on Ontologies for the Semantic Web, Edmonton, Canada.

Salton, G., Fox, E. A., & Wu, H. (1983, September). *An automatic environment for boolean information retrieval*. Paper presented at the IFIP 9th World Computer Congress, Paris, France.

Sanderson, M., & Croft, B. (1999, August). *Deriving concept hierarchies from text*. Paper presented at the 22nd Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval, Berkeley, CA.

Swartout, B., Ramesh, P., Knight, K., & Russ, T. (1997, March). *Toward distributed use of large-scale ontologies*. Paper presented at AAAI Spring Symposium on Ontological Engineering, Stanford, CA.

Tang, S., & Cai, Z. (2010, July). *Tourism domain ontology construction from the unstructured text documents*. Paper presented at 9th IEEE International Conference on Cognitive Informatics, Beijing, China.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139-1168.

Tzeng, J. S., Liou, W. C., & Sun, C. M. (2007). Constructing a lexical semantic network based on a domain dictionary. *Kansei Engineering International*, *7*(1), 47-54.

Uschold, M., & King, M. (1995, August). *Towards a methodology for building ontologies*. Paper presented at IJCAI 1995, Workshop on Basic Ontological Issues in Knowledge Sharing, Quebec, Canada.

Waterson, A., & Preece, A. D. (1999). Verifying ontological commitment in knowledge-based systems. *Knowledge-Based Systems*, *12*, 45-54.

Yang, C. C., Luk, J. W. K., Yung, S. K., & Yen, J. (2000). Combination and boundary detection approaches on Chinese indexing. *Journal of the American Society for Information Science*, *51*, 340-351.

Zhou, L. (2007). Ontology learning: State of the art and open issues. *Information Technology and Management*, *8*, 241-252.

附錄　公司治理本體評估問卷

親愛的受試者：

本問卷主要目的在瞭解您對公司治理概念相關性之認同程度，以作爲本研究改善分類本質的參考。問卷爲無記名，個別資料絕不公開披露，請依照你的真實感受填答。你的意見是本研究持續改善的重要參考依據，懇請撥冗詳實填答，感謝你的合作！

臺北商業技術學院會計資訊學系教授 周濟群

一、個人基本資料

1. 大學科系：□會計系 □非會計系（請填科系）_____

二、以下是有關公司治理議題與代理理論的相關性問題 以個人經驗與認知程度，勾選適當的答案。	同意	不同意
1. 你認同「代理問題」是「市場環境機制與考量」議題的一部分嗎？	□	□
2. 你認同「代理問題」可能起因於「自利行爲」嗎？	□	□
3. 你認同「自利行爲」可能起因於「資訊不對稱」嗎？	□	□
4. 你認同「自利行爲」可能起因於「契約不完備」嗎？	□	□

三、以下是有關公司治理議題一般性問題 以個人經驗與認知程度，勾選適當的答案。		
1. 你認同「股東權益」是「公司治理」議題架構的一部分嗎？	□	□
2. 你認同「資訊揭露透明度」是「公司治理」議題架構的一部分嗎？	□	□
3. 你認同企業內部「操守與文化」是「公司治理」議題架構的一部分嗎？	□	□
4. 你認同「監督」是「公司治理」議題架構的一部分嗎？	□	□
5. 你認同企業內部「組織結構」是「公司治理」議題架構的一部分嗎？	□	□
6. 你認同「董事會責任」是「公司治理」議題架構的一部分嗎？	□	□
7. 你認同「利害關係人」是「公司治理」議題架構的一部分嗎？	□	□
8. 你認同「內部控制組成要素」是「公司治理」議題架構的一部分嗎？	□	□

9. 你認同「董事會責任」包含了「公平對待股東」項目嗎？ ☐ ☐
10. 你認同「董事會責任」包含了設置「審計委員會」嗎？ ☐ ☐
11. 你認同「董事會責任」包含了擁有「專業判斷」能力嗎？ ☐ ☐
12. 你認同監督遵循「相關主管機關」法令屬於「審計委員會」責任範圍內嗎？ ☐ ☐
13. 你認同接洽「財務會計師」屬於「審計委員會」責任範圍內嗎？ ☐ ☐
14. 你認同「利害關係人—企業內部團體」包含了「內部稽核人員」嗎？ ☐ ☐
15. 你認同「利害關係人—與企業往來的團體」包含了「財務分析師」嗎？ ☐ ☐
16. 你認同「利害關係人」包含了「法人投資人／積極性股東」嗎？ ☐ ☐
17. 你認同「法人投資人／積極性股東」會關注企業「經營權」議題嗎？ ☐ ☐
18. 你認同「內控要素—對資訊處理的控制」是「公司治理」議題內容的一部分嗎？ ☐ ☐
19. 你認同「內控要素—控制活動」是「公司治理」議題內容的一部分嗎？ ☐ ☐
20. 你認同「內控要素—授權」是「公司治理」議題內容的一部分嗎？ ☐ ☐

四、以下是有關公司治理議題與企業風險管理的相關性問題以個人經驗與認知程度，勾選適當的答案。

1. 你認同「企業風險管理」包含了「風險辨識」概念嗎？ ☐ ☐
2. 你認同「企業風險管理」包含了「風險評估」概念嗎？ ☐ ☐
3. 你認同「企業風險管理」包含了「ERM監督」概念嗎？ ☐ ☐
4. 你認同「企業風險管理」包含了「ERM風險評估」概念嗎？ ☐ ☐
5. 你認同「企業風險管理」包含了「風險回應」概念嗎？ ☐ ☐
6. 你認同「企業風險管理」包含了「風險事項辨認」程序嗎？ ☐ ☐

7. 你認同「風險胃納」是「ERM風險評估」內容的一部分嗎？ ☐ ☐
8. 你認同「風險容忍度」是「ERM風險評估」內容的一部分嗎？ ☐ ☐
9. 你認同「固有風險」是「風險評估」內容的一部分嗎？ ☐ ☐
10. 你認同「剩餘風險」是「風險評估」內容的一部分嗎？ ☐ ☐

五、以下是有關公司治理議題與內部控制的相關性問題

1. 你認同「內部控制」包含了「內部控制無效」概念嗎？ ☐ ☐
2. 你認同 「內部控制」包含了「自行檢查評估結果」概念嗎？ ☐ ☐
3. 你認同 「內部控制」包含了「內部會計控制系統」概念嗎？ ☐ ☐
4. 你認同 「內部控制」包含了「內部控制組成要素」概念嗎？ ☐ ☐
5. 你認同「內部控制先天限制」是「內部控制無效」內容的一部分嗎？ ☐ ☐
6. 你認同「重大缺失」是「內部控制無效」內容的一部分嗎？ ☐ ☐
7. 你認同「顯著缺失」是「內部控制無效」內容的一部分嗎？ ☐ ☐
8. 你認同「控制缺失」是「內部控制無效」內容的一部分嗎？ ☐ ☐
9. 你認同「應報導情況」是「內部控制無效」內容的一部分嗎？ ☐ ☐
10. 你認同「持續監督」屬於「自行檢查評估結果」的內容範圍嗎？ ☐ ☐
11. 你認同「報導缺失」屬於「自行檢查評估結果」的內容範圍嗎？ ☐ ☐
12. 你認同「交易紀錄」屬於「內部會計控制系統」的內容範圍嗎？ ☐ ☐
13. 你認同「資產保管與紀錄責任」屬於「內部會計控制系統」的內容範圍嗎？ ☐ ☐
14. 你認同「內部控制組成要素」包含了「對資訊處理之控制」概念嗎？ ☐ ☐
15. 你認同「內部控制組成要素」包含了「資訊」概念嗎？ ☐ ☐

16. 你認同「內部控制組成要素」包含了「控制活動」概念嗎？	☐	☐
17. 你認同「內部控制組成要素」 包含了「內部控制參考依據」概念嗎？	☐	☐
18. 你認同「授權」屬於「對資訊處理之控制」的內容範圍嗎？	☐	☐
19. 你認同「應用控制」屬於「對資訊處理之控制」的內容範圍嗎？	☐	☐
20. 你認同「資訊揭露」屬於「資訊」的內容範圍嗎？	☐	☐
21. 你認同「揭露時效性」屬於「資訊」的內容範圍嗎？	☐	☐
22. 你認同「直接功能或作業管理」是「控制活動」內容的一部分嗎？	☐	☐
23. 你認同「對系統軟體之控制」是「控制活動」內容的一部分嗎？	☐	☐
24. 你認同「會計控制」是「控制活動」內容的一部分嗎？	☐	☐
25. 你認同「交易流程」屬於「會計控制」的內容範圍嗎？	☐	☐
26. 你認同「會計制度」屬於「會計控制」的內容範圍嗎？	☐	☐
27. 你認同「公開發行公司建立內部控制制度處理準則」是「內部控制參考依據」內容的一部分嗎？	☐	☐

六、你認爲本研究尚可納入哪些公司治理議題內容。

七、如對本研究有任何具體建議，敬請不吝提出！

Journal of Library and Information Science Research 6:2 (June 2012)

A Research on How to Construct the Prototype of Knowledge Ontology Based on Glossary -- Using the Domain Knowledge of "Corporate Governance" as an Illustration

Chi-Chun Chou
Associate Professor
Department of Accounting Information
National Taipei College of Business

Yu-Liang Chi
Professor
Department of Information Management
Chung Yuan Christian University

Jian-Shiun Tzeng
Ph.D. Candidate
Department of Information Management
National Taiwan University

Globalization is affecting complex institutional structures, such as corporate governance, worldwide, especially with the harmonization of civic society demands for accountability, transparency and information disclosure to stakeholders concerned with corporate sustainability reporting and corporate social responsibility (CSR). In recent years, computer systems have developed domain ontologies to help achieve interoperable information sharing and resource collaboration. Ontologies provide conceptually distinct and discrete descriptions, which convey abstract processes in discretely characterized steps, which are frequently used in computer searches, information exchange and semantic processing, permitting pervasive interdisciplinary conceptual sharing and interoperability. Ontologies are the critical structural element in knowledge management systems facilitating knowledge storage and use. However, developing a specific ontology requires depending on domain experts experience and knowledge, along with analysis, summarization, abstraction, and revision in a labor-intensive and time-consuming process. Therefore, this

study utilizes text mining techniques such as ontology learning (Cimiano, 2006; Maedche & Staab, 2001) or ontology design based on corpus analysis (Aussenac-Gilles, Biébow, & Szulman, 2000) to auto-construct or semi-auto-construct domain expertise for a prototype corporate governance ontology.

Aussenac-Gilles et al. (2000) proposed ontology engineering, deploying natural language processing tools for semantic relation extraction from massive corpuses, permitting semantic network elucidation of conceptual relations and classes for a prototype ontology. Such methods do not however explain the procedural principles underlying the knowledge or its interconnectedness, or the standard techniques or tools involved, so this study deploys a modified version (A-G ontology engineering, AGOE) which relies on text exploration and language analysis techniques for the auto-extraction of prototypical ontologies for revision by domain experts.

The method proposed in this study is different from other text-based Ontologies, as the system glossary (gloss tables) have explicit definition forms, so when deploying collocation screening and sentence analysis, the system is better able to effectively enhance the domain knowledge recall rate (Brewster, Ciravegna, & Wilks, 2003; Hearst, 1992). Among the domain content collected for the corpus used in this study were texts, journal articles and theses/dissertations dealing with corporate governance, internal controls and enterprise risk management, to build the system glossaries. Further, to identify corporate governance domain keywords, this study used the domain corpus contents to produce an N-gram, then used Term Frequency and Mutual Information (MI) for keyword screening. After filtering out the keywords, the study deployed full-text search to find the location of the key words in the texts, then relied on domain experts to manually delineate the definition and interpretation of the keywords. Then the interpretative contents of the annotated glosses were analyzed using the Academia Sinica's Institute of Chinese system for Chinese word segmentation, after which appropriate emendations were made, generating the final corporate governance domain lexicon used in this study. We then used the corporate governance domain lexicon as the basis for a Vector Space Model (VSM; Salton, Fox, & Wu, 1983), deploying the TFIDF (term frequency inverse document frequency) mode, to calculate the appropriately weighted term characteristics, generating the word item feature vector to elucidate a semantic network.

This study presents the use of text mining technology to build an ontology of knowledge for corporate governance issues, necessitating therefor, analysis of a complex semantic network and appropriate streamlining, in order to elucidate the ontology's hierarchical relationships. Initially, the semantic network was analyzed to identify structural Identify Groups, then applying a clustering algorithm, deploying plane clustering for all the words, thus forming word subsets for calculation of similarity of each sub-cluster center. Then, according to the results of the clustering algorithm, establish hierarchical clustering for each concept group, identifying the key issues, to Define "Leading-Relative" Concepts, in order to clarify the relationships among the central and related concepts. Then after using key issues to identify the related concepts associated with the central concept, this study used an auxiliary system for semantic structure analysis, lexical co-occurrence analysis, and sentence analysis, to Develop Concept Hierarchy. After enhancing the efficiency of establishing the ontology through deploying structural Identify Groups, clarifying key issues for conceptual relevance, and using semantic structure analysis of the conceptual hierarchical relationships, we finally undertook to Organize and Adjust, hierarchically expanding an arrangement of the multiple dispersed ontological knowledge while adjusting the relationships between concepts during this integration process. If during the organizational processes problems were encountered or ontology classes were not clear enough, we ascertained other conceptual groups amidst the semantic network for extraction and amendment.

This study's primary contribution lies in developing the AGOE method as an enhanced application of the A-G method, and clearly delineating the require processes and technical suite of tools necessary for real world implementation using the corporate governance domain as an example. As this ontology primarily relies on stratification of the conceptual framework, it remains an initial prototype ontology rather than a final application ontology, so this study focused on establishing the glossary, concepts, class structure, and prototyping methods, appropriate to empirically demonstrate expeditious elucidation of this prototype ontology. Finally, this study relied on questionnaires and surveys of five domain experts, achieving over 90% agreement among them, thus preliminarily indicating the domain experts could rely on the prototype ontology to undertake amendments yielding a final, usable ontology

application. Therefore we conclude that the AGOE method is empirically demonstrated to facilitate pre-processing to elicit initial prototype ontologies, especially for those domains requiring text mining techniques to accomplish the ontology design, through deploying a combination of tools to create a glossary, build the semantic network and analyzing the corpus semantic relations for deriving a practical corporate governance domain ontology.

Efforts such as this study, are essential to informing robust global corporate social responsibility and sustainability reporting requirement template and rubric harmonization. Future research should pursue discrete examples of corporate governance domains in geographical terms, such as eliciting the traditional Japanese institutional arrangements, or so-called relational contingent governance, among the panoply of culturally diverse examples of global corporate governance styles. As the world demands green, Smart Cities, and Stock Exchanges require corporate sustainability reporting, harmonization of corporate governance domain standards will be necessitated to ensure common access to transparent information for concerned stakeholders. Smart Grids deployed in Smart Cities will require interoperable access to extensive corporate governance information permitting pervasive business process sharing which promises to greatly enhance energy-efficiency and carbon reduction through flexible common semantics permitting the corporate knowledge infrastructure to revivify as documents and data learn to speak to each other as information that thinks. Appropriate ontologies overcome information disparity and the digital divide to help organize disperse information as the basis for efficacious knowledge exploitation that renders the otherwise discrete universe of data into effectively usable experience-based wisdom, much as ISO standards and global industry standard harmonization help advance the globalization of business knowhow and ensure universal corporate governance integrity.

References

Academia Sinica. (2009). *Zhong wen duan ci xi tong* [Chinese word segmentation system]. Retrieved July 1, 2009, from http://ckipsvr.iis.sinica.edu.tw/ [Text in Chinese].

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise？ The information content of internet stock message boards. *The Journal of Finance*, *59*, 1259-1294.

Aussenac-Gilles, N., Biébow, B., & Szulman, S. (2000, October). *Revisiting ontology design: A method based on corpus analysis*. Paper presented at the 12th International Conference on Knowledge Engineering and Knowledge Management, Juan-les-Pins, France.

Biébow, B., & Szulman, S. (1999, May). *Terminae: A linguistic-based tool for the building of a domain ontology*. Paper presented at the 11th European Workshop on Knowledge Acquisition, Modeling and Management, Dagstuhl Castle, Germany.

Brewster, C., Ciravegna, P., & Wilks, Y. (2003, August). *Background and foreground knowledge in dynamic ontology construction*. Paper presented at the SIGIR Semantic Web Workshop, Toronto, Canada.

Chi, Y.-L. (2007). Elicitation synergy of extracting conceptual tags and hierarchies in textual document. *Expert Systems with Applications*, *32*, 349-357.

Chi, Yu-Liang, & Tsai, Ming-Hung. (2007). Knowledge acquisition approaches for building ontological conceptual prototypes in document. *Journal of Information Management*, *14*(3), 47-66.

Chou, Chi-Chun, & Lian, Zih-Jie. (2011). Enhancing effectiveness of business information retrieval and integration via text mining and XBRL technology. *Journal of Contemporary Accounting*, *12*(1), 85-114. [Text in Chinese].

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22-29.

Cimiano, P. (2006). *Ontology learning and population from text: Algorithms, evaluation and applications*. New York: Springer.

Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point？ *Data & Knowledge Engineering*, *46*, 41-64.

Council for Economic Planning and Development. (2003). *Accountability of companies in Taiwan: Policy agenda and action plan to strengthen corporate governance*. Taipei, Taiwan: Author. [Text in Chinese].

Dong, Zhen-Dong, Dong, Qiang, & Hao, Chang-Ling. (2007). Theoretical findings of HowNet. *Journal of Chinese Information Processing*, *21*(4), 3-9. [Text in Chinese].

Downey, D., Etzioni, O., Soderland, S., & Weld, D. S. (2004, July). *Learning text patterns for web information extraction and assessment.* Paper presented at the American Association for Artificial Intelligence Workshop on Adaptive Text Extraction and Mining, San Jose, CA.

Engelberg, J. (2008, January). *Costly information processing: Evidence from earnings announcements*. Paper presented at American Finance Association Annual Meeting, San Francisco, CA.

Fernández-López, M., & Gómez-Pérez, A. (2002). Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, *17*, 129-156.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological engineering: With examples from the areas of knowledge management, e-commerce and the semantic web*. New York: Springer.

Grüninger, M., & Fox, M. (1995, August). *Methodology for the design and evaluation of ontologies*. Paper presented at IJCAI 1995, Workshop on Basic Ontological Issues in Knowledge Sharing, Quebec, Canada.

Hamon, T., & Nazarenko, A. (2001). Detection of synonymy links between terms: Experiment and results. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 185-208). Amsterdam: John Benjamins.

Hearst, M. A. (1992, August). *Automatic acquisition of hyponyms from large text corpora*. Paper presented at the Fourteenth International Conference on Computational Linguistics, Nantes, France.

Hindle, D. (1990, June). *Noun classification from predicate-argument structures*. Paper presented at the 28th annual meeting on Association for Computational Linguistics, Pittsburgh, PA.

Hiroko, F., Simmons, D. B., Newton, C. E., & Robert, E. S. (1997). Knowledge conceptualization tool. *IEEE Transactions on Knowledge and Data Engineering*, *9*, 209-220.

Lee, C. T., Huang, I., & Fang, K. T. (2010, July). *A study of building tax knowledge-based system: An ontological orientation -- Using cases under the national tax administration of central Taiwan province, ministry of finance*. Paper presented at Technology Management for Global Economic Growth, Phuket, Thailand.

Lin, D. (1998, August). *Automatic retrieval and clustering of similar words.* Paper presented at the 17th International Conference on Computational Linguistics, Quebec, Canada.

Liu, P., Hu, Y., Wang, X., & Liu, K. (2011, May). *A methodology for domain ontology construction in information science.* Paper presented at 2011 International Conference on E-Business and E-Government, Shanghai, China.

Ma, Xiu-Ru. (2006). *Qi ye feng xian guan li: Zheng he jia gou* [Business risk management: An integrated approach]. Taipei, Taiwan: Accounting Research and Development Foundation. [Text in Chinese].

Ma, Xiu-Ru, Lai, Sen-Ben, Ruan, Zhong-Qi, & Li, Mei-Que. (2005). Qi ye feng xian guan li [Business risk management]. *Accounting Research Monthly*, *238*, 28-78.

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, *16*(2), 72-79.

Morgan, A., Hirschman, L., Yeh, A., & Colosimo, M. (2003, July). *Gene name extraction using FlyBase resources*. Paper presented at the ACL Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan.

Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protege-2000. *IEEE Intelligent Systems*, *16*(2), 60-71.

Rajsiri, V., Lorré, J.-P., Bénaben, F., & Pingaud, H. (2010). Knowledge-based system for collaborative process specification. *Computers in Industry*, *61*, 161-175.

Reed, S., & Lenat, D. B. (2002, July). *Mapping ontologies into Cyc*. Paper presented at AAAI 2002 Conference Workshop on Ontologies for the Semantic Web, Edmonton, Canada.

Salton, G., Fox, E. A., & Wu, H. (1983, September). *An automatic environment for boolean information retrieval*. Paper presented at the IFIP 9th World Computer Congress, Paris, France.

Sanderson, M., & Croft, B. (1999, August). *Deriving concept hierarchies from text*. Paper presented at the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA.

Securities & Futures Institute. (2011). *Taiwan gong si zhi li jian jie* [Introduction to corporation governance in Taiwan]. Taipei, Taiwan: Author. [Text in Chinese].

Swartout, B., Ramesh, P., Knight, K., & Russ, T. (1997, March). *Toward distributed use of large-scale ontologies*. Paper presented at AAAI Spring Symposium on Ontological Engineering, Stanford, CA.

Tang, S., & Cai, Z. (2010, July). *Tourism domain ontology construction from the unstructured text documents*. Paper presented at 9th IEEE International Conference on Cognitive Informatics, Beijing, China.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139-1168.

Tzeng, J. S., Liou, W. C., & Sun, C. M. (2007). Constructing a lexical semantic network based on a domain dictionary. *Kansei Engineering International*, *7*(1), 47-54.

Uschold, M., & King, M. (1995, August). *Towards a methodology for building ontologies*. Paper presented at IJCAI 1995, Workshop on Basic Ontological Issues in Knowledge Sharing, Quebec, Canada.

Waterson, A., & Preece, A. D. (1999). Verifying ontological commitment in knowledge-based systems. *Knowledge-Based Systems*, *12*, 45-54.

Yang, C. C., Luk, J. W. K., Yung, S. K., & Yen, J. (2000). Combination and boundary detection approaches on Chinese indexing. *Journal of the American Society for Information Science*, *51*, 340-351.

Yi, Ming-Qiu. (2003). *Gng si zhi li* [Corporation governance]. Taipei, Taiwan: Hong Zhi. [Text in Chinese].

Zhou, L. (2007). Ontology learning: State of the art and open issues. *Information Technology and Management*, *8*, 241-252.