以文件為對象的概念萃取程序建立知識本體的雛型架構

戚玉樑
中原大學資訊管理學研究所

蔡明宏
中原大學資訊管理學研究所

摘要

對建置本體(Ontology)於知識應用的開發者而言，定義合宜的本體概念是棘手且耗時的工作，目前大多數的實務領域均缺乏可供知識系統直接利用的概念定義，因此在真實世界的認知與系統所處理的知識之間，存有表達上的落差。由於文件為表達人類認知的重要媒介，因此已成為收集知識概念的主要來源之一，但文件原僅為提供人類形而上的理解，對於發展為系統可用的知識概念，則有下列問題須解決：1).如何篩選出有意義的詞彙；2).如何將詞彙發展為命名的概念；3).如何建立概念之間的關係架構。本研究探討如何由文件中發展一套概念萃取的分析程序，它包含了文字探勘、語言學、心理及資料分析等學理方法，依據問題的特性及可行的解決方案整合成為可實際執行的「概念萃取程序」，最終將產出本體的概念雛型架構，以提供本體建置者的藍圖，並進一步發展為正式的領域本體。由實證的評估顯示，本程序可達到 70%以上的主觀接受率，因此可做為建置者在開發本體概念架構初期的重要參考依據。

關鍵字：本體、知識擷取、概念萃取、概念架構。

Knowledge Acquisition Approaches for Building Ontological Conceptual Prototypes in Document

Yu-Liang Chi
Dept. of Management Information Systems, Chung Yuan Christian University

Ming-Hung Tsai
Dept. of Management Information Systems, Chung Yuan Christian University

Abstract

For ontology developers, constructing comprehensible concepts of a knowledge domain is often problematic and time-consuming. The development process can be improved by reusing concept representation. However, the availability of reusable ontology is limited in practice due to the communication gaps existing between human cognition and machine-readable semantic representation required in knowledge engineering. Nowadays, textual documents that act as one of the most important medium to express human cognition are major resources for the acquisition of knowledge in computing systems. This study aims to develop a systematical approach to building reusable ontology prototypes from the documents. Several issues are identified such as how to recognize terminologies in textual corpus, to name concept tags in the terminologies, and to derive conceptual hierarchies. The proposed approach integrates concept extraction techniques that employ text-mining, linguistic, and psychological analysis methods. The empirical assessment reports the usability ratios of both conceptual tags and hierarchies exceed 70 percents. Consequently, the synergy of the elicitation techniques allows the developers to efficiently create and distribute ontology repository in specific domains.

Key words: Ontology, Knowledge Acquisition, Concept Extraction, Conceptual Structure

壹、緒論

本體(Ontology)或稱為知識本體或領域本體，它在人類文明的歷程中佔有很重要的地位。本體是源自於哲學上探究萬事萬物，並加以歸納分析的學說，因此是常用於塑模事物形成的表達方法(Chandrasekaran et al. 1999)。資訊科技引進本體相關技術已有多年的發展經驗，例如人工智慧(Lehmann 1992; Minsky 1975)。本體在資訊應用的研究中有許多不同面向的定義，其中 Gruber(1993)的定義"Ontology is a specification of conceptualization" 最常被引用，因此本體在資訊應用即為對特定事物的概念化及抽象化，統一詮釋外界對該事物的認知。近年來，隨著 XML 技術的成熟與衍生規範的制定，本體的表達方式也逐漸採用註標語言，以期獲得再利用與分享的效益，例如 RDF(Resource Description Framework)與 OWL(Ontology Web Language)等的蓬勃發展。簡言之，以 XML 為核心的本體技術是資訊科技在改良知識表達、知識庫建置及推論應用的新興方式。

典型的本體應用，通常須先對特定領域建置本體的概念定義，再依定義的框架給予事證(Assertions)，最後利用相關技術對本體進行後續的進階應用，例如推論。但建置可供系統理解的概念定義並不容易，因為人類對事物的概念是一種「形而上」的認知，亦即哲學上所稱的“All the logic of human in mind”，因此從形而上的認知到機器處理的知識，必須經由萃取、分析、轉換及呈現等過程(Kayed & Colomb 2002; Sugumaran & Storey 2002)。目前在本體應用的研究中，本體定義通常委由領域專家及協同知識工程人員來共同建置，透過繁複的認知收集，逐步形成本體的概念定義，並需要持續的修正，因此一個成熟的本體定義需要冗長的發展與維護。建置本體定義的另一項問題則是如何去取得大眾對概念的共識，實務上若該領域已有系統化的分類，則可做為本體建置的參考依據，例如自然界的「界、門、綱、目……」等的科學分類，但事實上大部份的領域均欠缺可供參考的系統化分類，因此許多研究也指出：建置本體定義是一項工藝遠勝於科學技術的工作，其過程需要投入大量的時間及人力(Noy & McGuinness 2001; van der Vet & Mars 1998)。簡言之，建置本體的標準、方式及程序，仍處於早期的研發階段(Gillam et al. 2005)。

隨著本體技術在知識應用上的普及，原本應由「專家」來建置本體的工作，其界線也隨之模糊，例如對日常生活的某一事物或工作環境中的某一項議題等，所謂的「專家」將逐漸由發展本體的知識工程人員來替代，因此建置本體也將是普及化的工作。對於缺乏系統化分類的領域而言，建置具有共通性的本體是一項挑戰，而因為文件是最普遍的資料形式，因此也成為主要的知識來源，但由於文件原來只用於表達給人類理解，對系統或機器的處理則沒有幫助，綜合過去由文件做為建置本體素材的研究(Hui & Yu 2005; Lassila & McGuinness 2001; Shamsfard & Barforoush 2004)，下列三項問題須予以克服：(1). 如何在文件中篩選出有意義的詞彙，並去除與概念無關的詞彙。(2). 如何將詞彙命名為具有共識的概念，亦即以語意(Semantic)的角度將具有相同內涵的詞

彙統一成為同一個概念名稱。(3). 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。本研究之目的即為解決上述三項問題，分別分析及提出對應的解決方法，為使抽象的理論轉化為具象可行的機制，本研究也藉整合目前可行的方法成為「概念萃取程序」，具體協助本體開發者應用。

本研究的內容安排如次：第二節針對本體的內涵及目前本體建置工程的相關研究做文獻回顧。第三節探討「概念萃取程序」的設計，並建議一個實施程序。第四至六節分別探討「概念萃取程序」的三項程序及可行的方法。第七節是以學術文章的摘要內容做為概念萃取程序的實證材料，實際產出一個本體的雛型架構。最後，第八節討論應用的結果、評估與經驗分享。

貳、本體的內涵及其建置

本體可視為將事物以概念化架構來塑模認知的知識管理方式，或是一個定義完整的系統化分類(Taxonomy)，因此本體相關技術的資訊應用是以語意處理、知識庫及智慧型系統為主。網際網路標準組織(W3C)曾列舉六大類常應用的類型，而相關應用的研究也非常普遍，例如改善目前網頁內容的知識表達(Martin & Eklund 2000)與入口網站的分類管理(Staab et al. 2000)；文件內容呈現(Motta et al. 2000)；代理人協商的設計(Khedr & Karmouch 2005)；及其他具有語意知識及智慧型系統的應用(Fernandez-Breis & Martnez-Bejar 2000; van Elst & Abecker 2002)。上述的各項應用都須以建置本體定義為基本前提，再進行加值的應用，因此本體定義的內涵是決定應用成效的關鍵。本體的內涵通常包括概念、階層關係及對概念或關係的描述，Corcho et al. (2003)曾建議依本體的實質內涵可分為簡化及完整兩種，前者僅包含概念及其階層關係、後者則加入細節描述或限制條件的公理(Axioms)或邏輯式。

本研究是以文件為建置本體定義的來源，在過去類似的研究中亦指出各種用在此類非結構化資料的分類方式：例如 Huhns & Shingh(1997) 建議依不同的需求，採用 word, keyword, catalog, glossary 等分類型式，提供代理人在知識處理上所需的分類；Lassila & McGuinness(2001)更提出本體由鬆散到嚴謹的等級，並以線性圖表示如圖 1，由左至右代表建置時的周延程度，也代表本體定義將愈來愈精緻完整，圖中央的斜線將線性圖分割為兩區塊，左半部的內容須由目錄、詞彙、詞典、專有名詞到具有階層性，右半部的內容則更趨嚴謹，因此包含如屬性、限制條件及邏輯等。圖 1 的左右區塊的內容非常近似前述 Corcho et al.所建議的簡化及完整兩種本體定義形式。

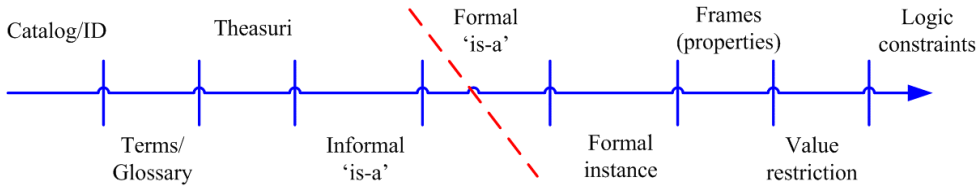


圖 1：Lassila & McGuinness (2001) 提出的 Ontology Spectrum

如前述所言，建置本體至今仍沒有標準的方式及程序。在過去建置本體的研究中，通常將開發的程序稱為本體工程方法(Ontology engineering)，例如 Noy & McGuinness(2001)曾對開發本體提出由決定範圍、概念定義到建立實例等七項步驟，而 Uschold & Grueninger(1996)亦曾建議名為 TOVE 的本體工程方法，它包含了由動機到測試等六項程序。因此，本體工程方法是一般的原則建議，並沒有具體的程序方法。建置本體若是以文件為參考素材時，則需要考量更多的環境限制及需求，部份限制更是本體工程方法沒有涉及的前置事項，例如文件上的干擾字、同義字處理等。

根據以上的分析，概念及階層關係是構成本體內涵的基礎，更精細的描述則須依賴領域特性進一步的分析，因此本研究的目標是以達成 Corcho et al.建議的簡化版或 Lassila & McGuinness 線性圖的左半部為目標範圍。另外，為符合取材來源為文件的特性，我們發展適合的萃取程序以達成有效率的本體建置作業。

參、概念萃取程序

本體建置作業是一項長期性的工作，它至少包含建置、更新及維護等三項階段，本研究目前僅針對建置階段提出因應的方法，特別是當建置本體的參考來源主要是文件資料時，如何將建置所需的資訊素材予以逐步萃取。根據過去探討由文件資料建置本體的相關研究中(Gillam et al. 2005; Jiang et al. 2003; Motta et al. 2000; Richards & Simoff 2001)，建置作業可由不同構面的觀點切入，因而解決問題的重點也相異。為使建置作業能兼顧各項構面，本研究將建置作業定位於萃取概念所需的名稱(Tags)及其階層架構，三項待解決的問題如下：

1. 如何快速有效的在文件資料中萃取有限的領域術語(Terminology)。
2. 如何解決語言學(Linguistic)上的同義詞彙問題，特別是探索詞彙在表達概念時的語意內涵是否相同，以利概念在命名時的統一。
3. 如何確認概念之間的階層關係，並建立結構化的概念系統。

針對上面問題，本研究提出一個「概念萃取程序」機制，並以三項對應的階段步驟來克服，這個機制整合了文字探勘(Text mining)、語言學、心理及資料分析等學理。如圖 2 由上至下所示，分析來源是文件資料，分別經過萃取領域術語、統一概念命名及建立概念階層等三項程序，最終將產出一個本體的雛型架構給開發者參考，圖中各階段的執行與銜接均具有方向及次序性。

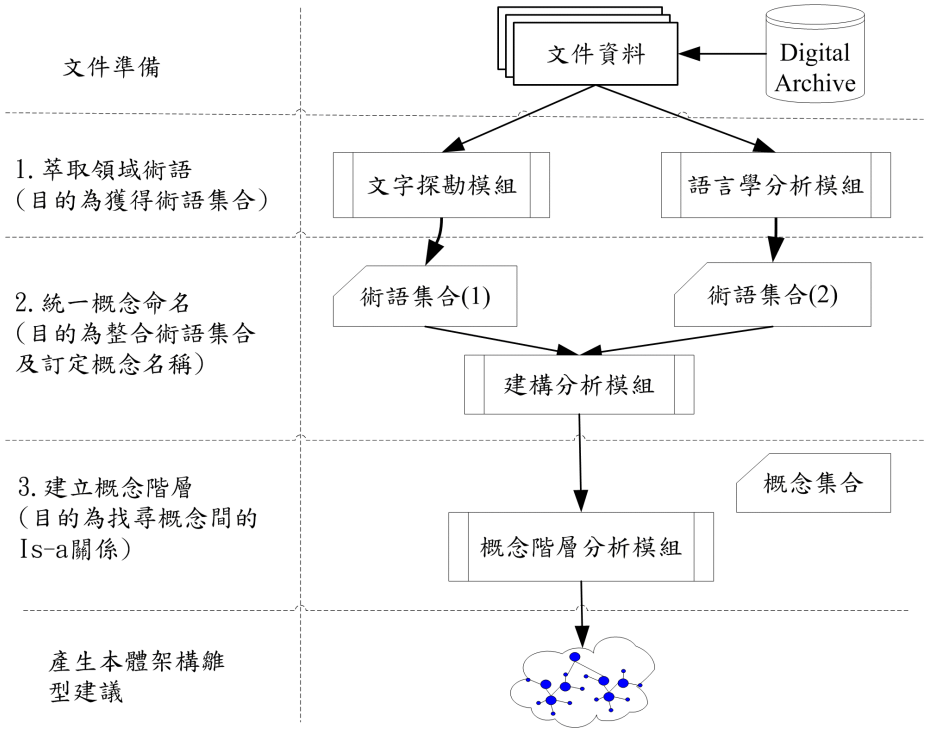


圖 2：概念萃取程序包含三項對應的處理步驟

圖 2 中，第一階段是萃取領域術語，須先處理文件資料中具干擾的雜訊詞彙(Noise words)，其次再處理與主題無關的詞彙，本階段包含兩項分析模組：文字探勘模組及語言學分析模組，其目的是藉兩項不同的模組來互補分析時的缺點；另一方面，本研究也將兩項模組模擬為兩位領域專家，以因應下一階段的需求。第二階段是藉心理學上的建構分析(Construct analysis)，應用於處理前一階段所產出的兩套領域術語詞彙，並企圖解決同義詞的問題，有別於傳統處理同義詞的作法，本研究的同義詞將作為命名概念的依據，因此須探索支持詞彙的「語意」是否相同或其相關的程度，以協助由詞彙發展到概念。最後，第三階段是找尋概念之間的階層關係(亦或“*is-a*”關係)，並產出本體定義的雛型架構。以上三項步驟的細節分別說明於第四至六節。

肆、萃取領域術語

本階段的萃取領域術語將以文字探勘及語言學分析兩項模組來分析文件資料，由於兩項不同的分析觀點將使得領域術語的結果相異，本研究也藉此平衡及互補單一方法過於武斷的問題，兩項模組最終將各自產生其領域術語的集合，繼續提供下階段的分析素材。

一、以文字探勘模組萃取領域術語

文字探勘是利用計算方式處理大量的文件資料，常見於知識發掘的相關研究(Han & Kamber 2001)。本研究考量領域主題與欲探勘的概念應具有相關性，因此利用關聯規則(Association rules)方法，逐步找出文件中的領域術語，部份利用類似方法的研究如 Han & Kamber (2001) 及 Srikant & Agrawal (1997)，均指出關聯規則可提供快速的分析結果，是萃取「主題-詞彙」的常用方式。關聯規則是以典型的因果邏輯式"*if-then*"來表達，在"*if* "的部份稱為前提或先決條件(Premise)，而"*then*"的部份稱為結論(Consequence)，例如本模組產生的詞彙集合以 $I = \{i_1, i_2, .. i_n\}$ 表示，一個關聯規則若表達詞彙 X 與其關聯的詞彙 Y，須以 $X \Rightarrow Y$ 來表示，此時須同時滿足 $X, Y \subseteq I$, 及 $X \cap Y = \phi$。

關聯規則通常也伴隨兩項評估指標來表達其精確的程度，第一項指標稱為支持度(Support)，該值是指我們鎖定的處理項目在全部處理項目的比率值，例如本研究中若全部處理項目為 100 項，詞彙 X 與 Y 同時出現的次數為 3，則其支持度為 3%，支持度以 *Support* $(X \Rightarrow Y)$ 表示如下：

$$Support(X \Rightarrow Y) = Support(X \cup Y) = \frac{包含X和Y的處理項目}{所有處理項目}$$

支持度亦可用條件機率表示如下：

$$Support(X \Rightarrow Y) = P(X \cap Y)$$

第二項指標稱為信賴度(Confidence)，該值是指在出現詞彙 X 的前提下，同時出現含有詞彙 X 與 Y 的比率值。假設 X 的總出現次數為 10 次，若 X 與 Y 同時出現 3 次，則信賴度為 30%。信賴度以 *Confidence* $(X \Rightarrow Y)$ 表示式如下：

$$Confidence(X \Rightarrow Y) = Support(X \cup Y) / Support(X)$$

信賴度亦可用條件機率表示如下：

$$Confidence(X \Rightarrow Y) = P(Y \mid X) = P(X \cap Y) / P(X)$$

在進行文字探勘分析時，兩項前置程序如下：

1. 建立須刪除的詞彙名單：這類在分析時須略去的詞彙，一般稱為干擾或雜訊的詞彙，例如"of", "it", "are"等。本研究項在訂定此詞彙名單是參考 Frakes & Baeza-Tates (1992)的建議，建立如介係詞、代名詞及 be 動詞等。
2. 採用詞類正規化演算法(Stemming algorithm)將相同意義但不同詞類的詞彙予以統一。詞類正規化是常用於自然語言及資訊萃取的前置分析技術，例如"ontology", "ontologies", "ontological"經由演算法的協助，可視為相同的詞彙。

在關聯規則的作業中，使用者須設定最小支持度及最小信賴度的門檻值，最小支持度意指是否有足夠的資料支持該規則的成立，最小信賴度是指該規則成立之前提與其結果之間的可信度，因此門檻值的高低將會決定關聯規則的數量，圖 3 列出利用文字探勘工具的部份關聯規則結果。圖中表列的各欄分別為：前提(Premise)、結果(Consequence)、支持度及信賴度，這些在關聯規則中的詞彙可視為與主題相關的詞彙，分析者須進一步整理成為領域術語。簡言之，本研究的文字探勘模組利用關聯規則來剔除與主題無關的詞彙，並篩選出有限的關鍵詞彙作為領域術語。

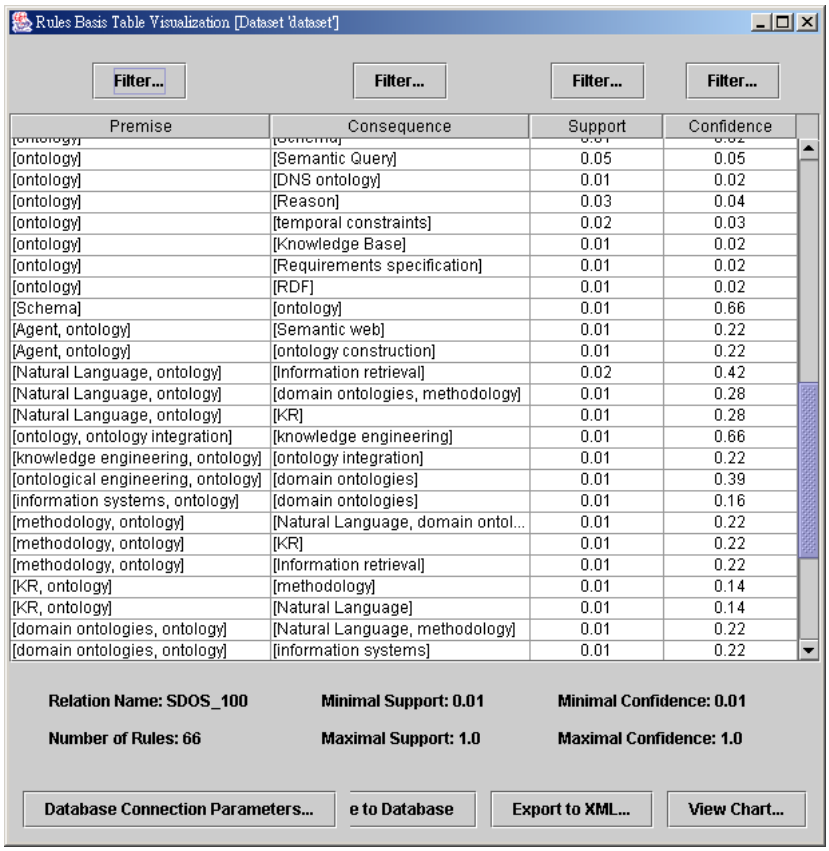


圖 3：利用 TextAnalyst 工具執行文字探勘的關聯規則

二、以語言學分析模組萃取領域術語

關聯規則雖然在文字探勘的觀點上，可視為獲得重要關鍵詞的有效方法，然而在語言學或語意的觀點上卻不一定成立，由於前述的關聯規則分析已先行剔除低頻率的詞彙，因此有些重要的關鍵詞因而被忽略，為彌補文字探勘方法在萃取領域術語上的盲點，如何利用語言學的分析也更加重要(Chung & Moldovan 1995)。

在語言學的觀點中，萃取領域術語是經由詞彙的語意關係來逐步發掘，目前具體的分析方式是利用自然語言處理(NLP, Natural Language Process)的方法(Cordon et al. 2002; Shamsfard & Barforoush 2004)。自然語言處理將文句利用語法及語意二種觀點來分析，語法注重文句的組成形式，語意則在探索句子的意義內涵。本模組主要是應用詞彙的語意關係權重，判斷語意是否相對的重要，其作法是根據詞彙的語法結構，予以統計及建立圖形化的結構，經重新正規化的詞彙權重和關係，稱之為語意權重，若以圖示結構來呈現則稱之為語意地圖(Semantic map)，亦可視為分析文件組成後所總結的語意概念(Semantic concepts)。

在進行自然語言處理工作時，也須先行去除干擾分析的雜訊詞彙，剔除這些詞彙的方式則類似前述文字探勘模組的作法。在進行萃取領域術語過程中，本研究採用 Alcala et al. (2003)的 linguistic rule learning 方法進行權重分析：首先將詞彙取出並讀入字元框格(n-character window)，再利用遞迴式類神經網路(RHNN, Recurrent Hierarchical Neural Network)來處理語意計算，最後將詞彙頻率予以記錄。整理語意權重的計算結果，可依權重排列獲得所需的詞彙表，亦即萃取後的領域術語。另外，也可根據詞彙的語意權重，將詞彙與其他詞彙串接起來，形成一個語意地圖，圖 4 是應用 TextAnalyst 工具執行後所產生的部份語意地圖，雖然圖中的各詞彙成階梯排列，但這項排列僅是語意的相關性，它們並不具有任何概念性的階層關係。

上述兩項領域術語的分析方法均能產出一組領域術語(關鍵詞)的集合，由於分析觀點不同，兩項術語集合的內容也有所差異，我們將以下階段的分析方法，將兩項術語集合併為一項。

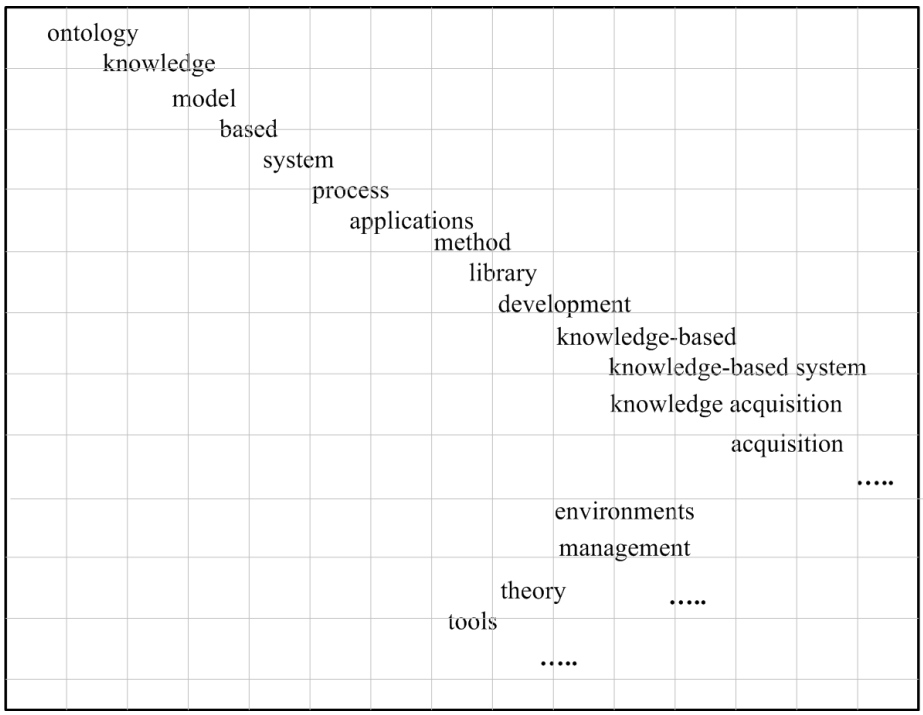


圖 4：利用 TextAnalyst 工具執行語意權重分析所產生的部份語意地圖

伍、概念命名的統一

Shaw & Gaines (1989)曾對詞彙篩選的語意問題提出稱為 COCO 的分類，並以四個區隔象限，包括：(1). 同詞同義(COnsensus)，詞彙與對應的意義概念皆相同；(2). 異詞同義(COrrespondence)，不同詞彙對應相同的意義概念；(3). 同詞異義(COnflict)，

相同詞彙但對應不同的意義概念；(4). 異詞異義(COntrast)，詞彙與對應的意義概念皆不相同。其中除同詞同義的理想情況外，其餘三項皆是詞彙處理所面臨的問題。前述萃取領域術語階段，本研究利用文字探勘及語言學兩種方法進行分析，兩種方法分別利用其計量方式來判斷，在處理同詞異義及異詞異義上獲得相當程度的正確性，但在異詞同義(亦即同義字)部份，因涉及高度的人為認知，故其效果較受限，因此所萃取的領域術語仍存在有同義字問題，不利於我們對概念的命名統一。

本研究利用凱利方格(RGT, Kelly's Repertory Grid Technique)方法來分析同義字問題，凱利方格原為應用在概念探索及其形成的分析方法，凱利方格是源自於 George Kelly 所提出的個人心理建構理論(PCP, Personal Construct Psychology)，心理建構所強調的重點在溯源事件形成的真正原因而不是它最終被稱為什麼，由於概念的形成是一種學習過程的結果，因此須回溯形成概念的構成因素，稱之為「建構」(Construct)，藉此判斷不同的概念是否為相同來源的依據(Gaines & Shaw 1993)。近年來，凱利方格已引進至智慧型系統的知識塑模分析上，特別是用於解析及萃取專家的認知，對於判別近似的認知有顯著的成效，許多利用凱利方格的研究也佐證了探索事物形成的建構，對收集概念有很好的效益(Batty & Kamel 1995; Ford et al. 1991)。簡言之，凱利方格為實際執行個人心理建構理論的計量方法，我們將藉凱利方格來接續詞彙分析，並據此發現具有相同建構的同義字。凱利方格分析可概分為認知評估及構念引出等兩階段進行，說明如下：

(1) 認知評估階段：我們將前述萃取領域術語的兩種方法視為兩位專家，它們以各自的觀點萃取領域術語，這兩組領域術語將做為建立凱利方格的元素及建構的資料來源。圖 5 是利用 WebGridIII 的凱利方格分析工具(參考 http://tiger.cpsc.ucalgary.ca/)，圖中位於矩陣方格兩側的項目稱為建構，分別表達兩種極端的建構，例如"is-a Methodology"及"not a Methodology"，圖中的右下方分別列出稱為元素的領域術語，矩陣方格即為使用者對該項認知的評估。評估分數由 1 代表「非常不同意」及 5 代表「非常同意」，使用者通常在建立「元素-建構」時，即須評估加權分數，以做為凱利方格的分析依據。

(2) 構念引出階段：構念引出是指隨著「元素—建構」的增刪或評估分數的改變而更動原有構念定義，任何的精進作為皆可稱為構念的「再引出」，這項構念再引出可重複到沒有更具體的變更為止。例如元素"Ontology Language"和"Domain Ontology"在凱利方格的認知評估階段，若兩組的元素因為擁有極為相似的建構(亦即構成的因素)，因此將呈現聚集狀態。為區別兩者的差異，分析者可加入新的建構，例如"is Markup Language"及"not Markup Language"，圖 6 為經凱利方格再次驗證後，我們以空間節點圖展示兩項構念已完全分開。

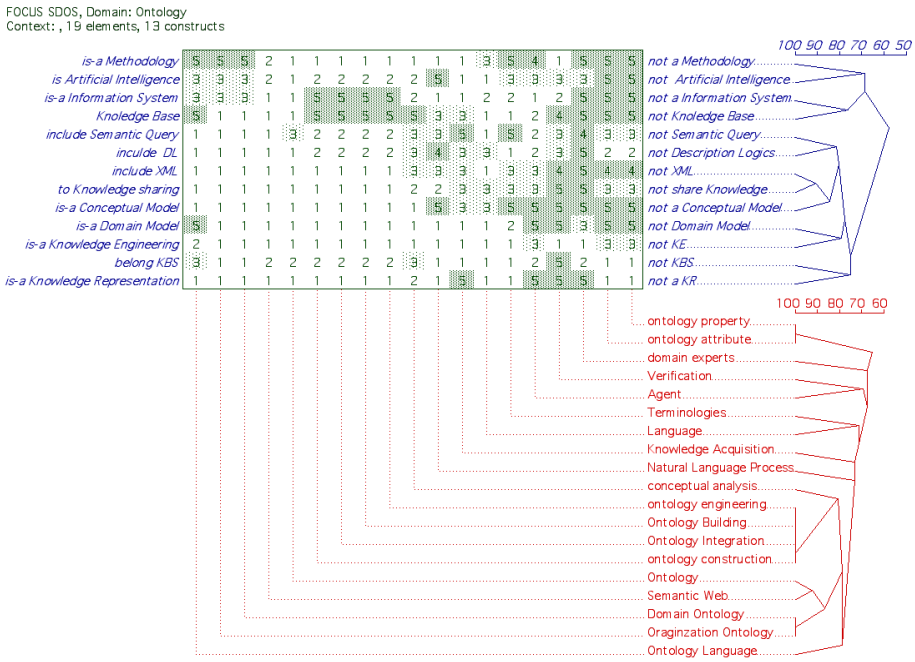


圖 5：使用 WebGridIII 執行凱利方格圖，分析元素與建構之間的構念形成

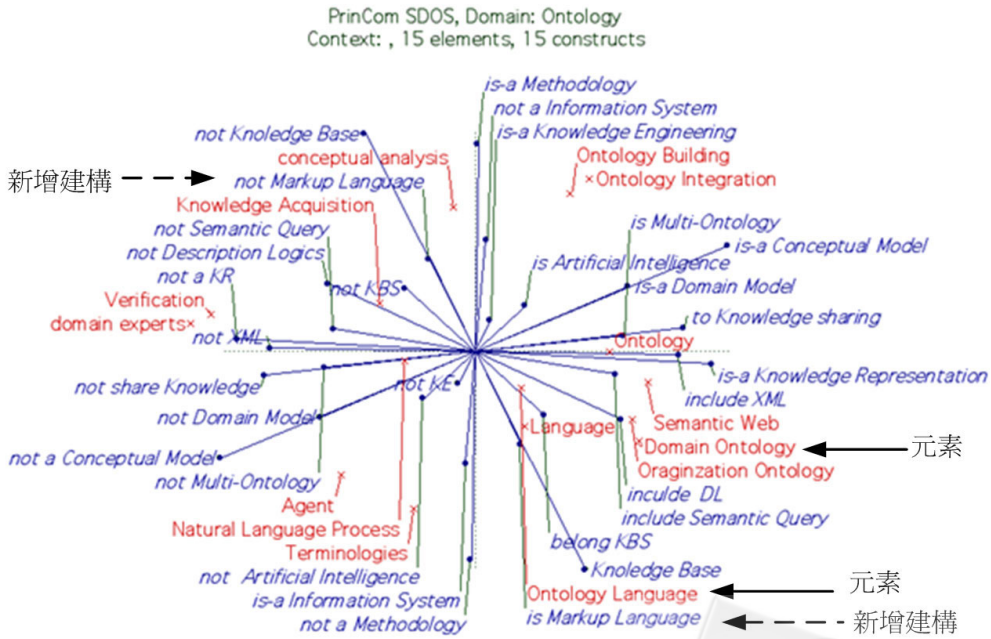


圖 6：加入兩項新增建構(如圖中標示之虛線箭頭)，Ontology Language 與 Domain Ontology 兩項元素已獲區隔(如圖中標示之實線箭頭)

本研究利用構念的引出及再引出，可將同義詞藉其構成的建構特徵予以區隔或聚集，例如“Ontology property”及“Ontology attribute”在其構念上具有很高的相似程度，因此可將兩詞彙判定為相同的領域術語，當命名該概念時，分析者可參考同義詞在前一階段的重要性而決定最終的命名，其他的同義詞則在概念定義時，賦予邏輯上的等價關係。本階段將可完成詞彙轉換為概念的命名程序。

陸、概念階層關係的確認

凱利方格促使領域術語能以一致性的命名詞彙詮釋概念，前述各階段已完成對概念的萃取，但仍欠缺概念架構所需的階層關係，本節將說明如何建立及確認階層關係的做法。

本研究利用正規概念分析法(FCA, Formal Concept Analysis)來解析概念及其階層關係，正規概念分析是衍生自格狀理論(Lattice theory)的資料分析方法，根據 Wille(1992)的研究，它最早開始於 1980 年代，主要應用於資料的篩選分析，所謂「正規」是指以數學模式來表達概念的意義，因此在建立時雖然以文字實施，但實際內容為一種數學形式，稱為 Formal Context。根據 Ganter &d Wille(1997)的介紹，Formal Context 通常以三個欄位的*(G, M, I)*公式表示，其中 *G* 代表 Context 的物件群，*M* 代表 Context 的屬性集合，*I* 則是 *G* 與 *M* 之間的二元關係，亦即一個物件 *g* 及一個屬性 *m*，具有 *gIm* 或 $(g,m)\in I$ 的關係。因此，假設 A 及 B 為二個概念時，若使得 $A\subseteq G$ 及 $B\subseteq M$ 成立，我們須定義

$A' := \{m \in M \mid (g,m)\in I \text{ for all } g \in A\}$

$B' := \{g \in G \mid (g,m) \in I \text{ for all } m \in B\}$

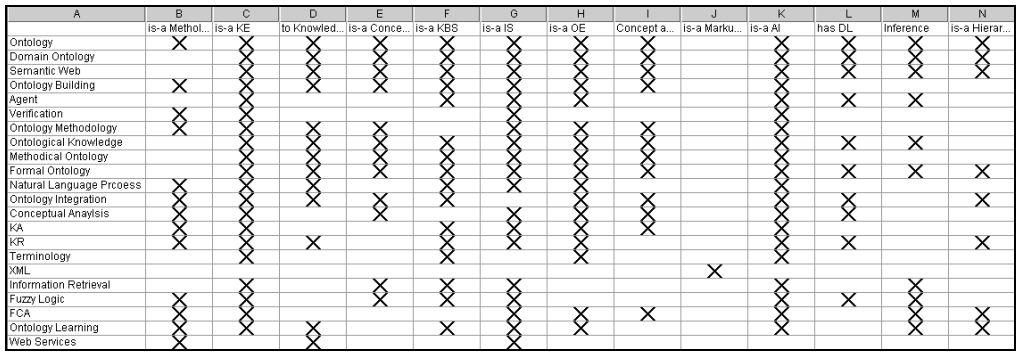
當(A,B)是 Formal Context *(G, M, I)*的一個正規概念時，必須符合(iff)

$A\subseteq G, B\subseteq M, A'=B \text{ and } B'=A$

此時，A 被稱為概念(A,B)的外在延伸(Extension)，B 則是概念(A,B)的內部組成(Intension)，外在延伸是指具有相同概念的其他物件，而內部組成是這些物件所具有的共同特徵(或屬性)。而在符合此正規概念的前提下，若代入具體的項目內容來重新表示概念時，則很容易依據 FCA 的原理發展出各種關係，例如包含關係：$(A_1, B_1)\leq(A_2, B_2): \Leftrightarrow A_1\subseteq A_2 (\Leftrightarrow B_2\subseteq B_1)$。最後，將這些有序的正規概念集合起來，則稱為 Concept lattice，並以 *B(G, M, I)*表示。

正規概念分析在概念階層的建立應用上，首先是以收集物件、屬性及關係等，再以方格圖建立。本研究在採用此方法所需的上述資料，是以前一階段的結果為來源，例如圖 7 是一個正規概念分析的方格圖，圖的左側是物件名稱，上方則是一些它們的屬性。因此 Formal Context 的*(G, M, I)*分別是物件(*G*)包括{*Ontology, Domain ontology, Semantic Web…*}、屬性(*M*)包括{*is-a Methodology, is-a KE, is-a KBS…*}、關係*(I)*是以符號"X"表示物件與屬性的二元關係，例如物件“*Ontology*”擁有屬性“*is-a Methodology*”。例如在圖 7 中，若以"Domain ontology"所具有的屬性(如 *is-a KE, is-a KBS…*)視為一個

正規概念時，此概念的外在延伸包括了{*Ontology, Semantic Web, Ontological knowledge…*}等物件，此概念的內部組成則包含{ *is-a KE, is-a concept, is-a KBS…*}等屬性。簡言之，正規概念藉由外在延伸及內部組成的內容界定，將有助於釐清概念間的階層關係(Guarino 1995)。



A	B	C	D	E	F	G	H	I	J	K	L	M	N
	is-a Method...	is-a KE	to Knowled...	is-a Conce...	is-a KBS	is-a IS	is-a OE	Concept a...	is-a Marku...	is-a AI	has DL	Inference	is-a Hierar...
Ontology	X	X	X	X	X	X	X	X		X	X	X	X
Domain Ontology		X	X	X	X	X	X	X		X	X	X	X
Semantic Web		X	X	X	X	X	X	X		X	X	X	X
Ontology Building	X	X	X	X	X	X	X	X		X			
Agent		X			X	X	X			X	X	X	
Verification	X	X				X				X			
Ontology Methodology	X	X	X	X		X	X	X		X			
Ontological Knowledge		X	X	X	X	X	X	X		X	X	X	
Methodical Ontology		X	X	X	X	X	X	X		X			
Formal Ontology		X	X	X	X	X	X	X		X	X	X	X
Natural Language Prcoess	X	X	X		X	X	X			X			
Ontology Integration	X	X	X	X	X		X	X		X	X		X
Conceptual Anaylsis	X	X		X		X	X	X		X	X		
KA	X	X			X	X	X	X		X			
KR	X	X	X		X	X	X			X	X		X
Terminology		X			X		X			X			
XML									X				
Information Retrieval		X		X	X	X				X		X	
Fuzzy Logic	X	X		X	X	X				X	X	X	
FCA	X	X				X	X	X		X		X	X
Ontology Learning	X	X	X		X	X	X			X		X	X
Web Services	X		X			X							

圖 7：使用 Galicia 工具建立物件與屬性的 Context lattice

建立正規概念分析的 Context lattice 是分析工作的初步階段，目前一般的正規概念分析工具通常也結合其他的演算法，提供使用者進一步尋找資料之間的隱性關係，例如本研究利用的 Galicia (http://www.iro.umontreal.ca/~galicia)，它提供以屬性為基礎的分析機制，當使用者完成方格圖後，正規概念分析工具即以屬性為標的，逐項計算資料之間的隱性關係，藉由互動問答方式，強制使用者完成關係的確認。例如某項物件並沒有某些屬性，但經系統計算發現與該物件有關的其他物件卻具有某些屬性，因此會探詢使用者的意圖，若使用者並不同意系統的質疑，則須提供具體的案例，這些互動式問答將持續到沒有隱性關係及衍生的關係為止，而每一次的問答將會改變 Formal Context 的組成，進而影響概念階層的架構關係。當各項正規概念確立時，概念之間的階層架構也發展完成，我們可藉由格線圖或表列方式呈現結果。

柒、實例驗證結果

為驗證概念萃取程序的各項階段，本研究測試資料來源是利用 ScienceDirect OnSite (或稱 SDOS)的電子期刊文章，由於現階段的電子期刊資料庫是以關鍵字查詢為主，我們給予主題字"*Ontology*"為本次收集文件的實證標的，隨機選出共計 150 篇期刊文章的摘要部份，並開發文件匯入系統來管理資料如圖 8，另為瞭解文章數量對執行概念萃取程序的影響程度，本研究分別以四組不同的文章數量(80 篇、100 篇、120 篇、150 篇)做為測試情境，並在相同的條件下進行實驗，以下是以文章數量為 100 篇的實證情形：

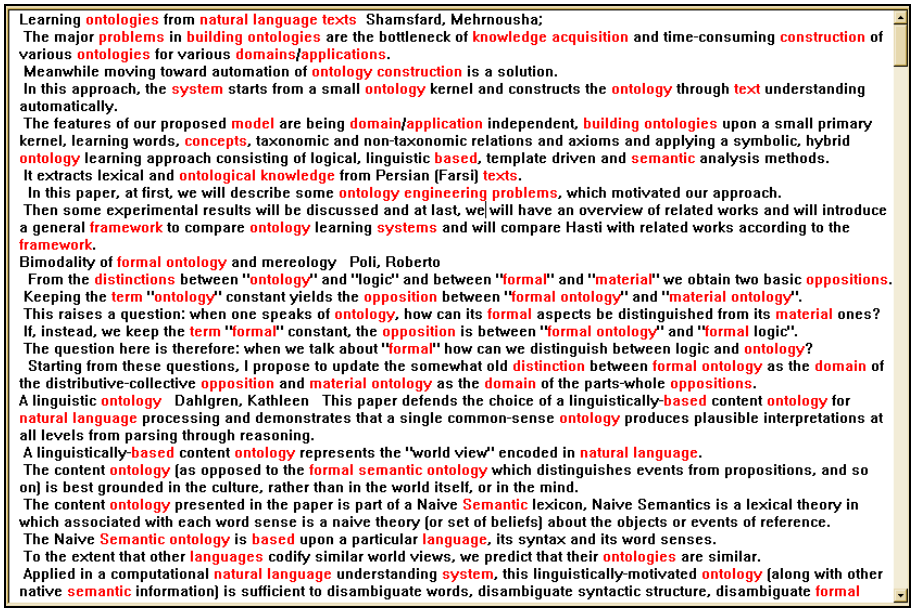

Learning ontologies from natural language texts Shamsfard, Mehrnousha;
The major problems in building ontologies are the bottleneck of knowledge acquisition and time-consuming construction of various ontologies for various domains/applications.
Meanwhile moving toward automation of ontology construction is a solution.
In this approach, the system starts from a small ontology kernel and constructs the ontology through text understanding automatically.
The features of our proposed model are being domain/application independent, building ontologies upon a small primary kernel, learning words, concepts, taxonomic and non-taxonomic relations and axioms and applying a symbolic, hybrid ontology learning approach consisting of logical, linguistic based, template driven and semantic analysis methods.
It extracts lexical and ontological knowledge from Persian (Farsi) texts.
In this paper, at first, we will describe some ontology engineering problems, which motivated our approach.
Then some experimental results will be discussed and at last, we will have an overview of related works and will introduce a general framework to compare ontology learning systems and will compare Hasti with related works according to the framework.
Bimodality of formal ontology and mereology Poli, Roberto
From the distinctions between "ontology" and "logic" and between "formal" and "material" we obtain two basic oppositions.
Keeping the term "ontology" constant yields the opposition between "formal ontology" and "material ontology".
This raises a question: when one speaks of ontology, how can its formal aspects be distinguished from its material ones?
If, instead, we keep the term "formal" constant, the opposition is between "formal ontology" and "formal logic".
The question here is therefore: when we talk about "formal" how can we distinguish between logic and ontology?
Starting from these questions, I propose to update the somewhat old distinction between formal ontology as the domain of the distributive-collective opposition and material ontology as the domain of the parts-whole oppositions.
A linguistic ontology Dahlgren, Kathleen This paper defends the choice of a linguistically-based content ontology for natural language processing and demonstrates that a single common-sense ontology produces plausible interpretations at all levels from parsing through reasoning.
A linguistically-based content ontology represents the "world view" encoded in natural language.
The content ontology (as opposed to the formal semantic ontology which distinguishes events from propositions, and so on) is best grounded in the culture, rather than in the world itself, or in the mind.
The content ontology presented in the paper is part of a Naive Semantic lexicon, Naive Semantics is a lexical theory in which associated with each word sense is a naive theory (or set of beliefs) about the objects or events of reference.
The Naive Semantic ontology is based upon a particular language, its syntax and its word senses.
To the extent that other languages codify similar world views, we predict that their ontologies are similar.
Applied in a computational natural language understanding system, this linguistically-motivated ontology (along with other native semantic information) is sufficient to disambiguate words, disambiguate syntactic structure, disambiguate formal

圖 8： 隨機選取自 SDOS 電子期刊資料庫，主題為"Ontology"的 100 篇期刊文章摘要，由本研究的系統彙整

- 在萃取領域術語階段，由兩項模組對同一資料進行分析，並各自獲得一組領域術語集合。文字探勘模組以關聯規則分析測試資料，分別設定最小支持度及最小信賴度的門檻值分別為 0.05 及 0.3，共計整理獲得 243 個領域術語。語言學分析模組採用 TextAnalyst 工具，協助語意關係的權重分析法，共計整理獲得 377 個領域術語。
- 在概念命名階段，本研究利用凱利方格工具來解決領域術語發生同義詞的問題，經合併前述階段的結果及去除重複的領域術語後，共計有 275 個元素及 194 個建構產生，受限於工具操作上的複雜性，須分割成較小的凱利方格進行分析，本階段最後引出 63 組聚集的元素，我們可據此將元素做為概念的統一命名。
- 在概念階層的確認階段，本研究利用正規概念分析法(FCA)將前述階段的 63 項概念做為 Context lattice 的物件，屬性則是參考凱利方格的建構，利用以屬性為基礎的隱性關係分析機制，各項正規概念之間的階層已逐漸確認，本階段最後歸併為 51 項正規概念。圖 9 是部份正規概念所構成的樹狀格線圖，圖中包含節點、格線及概念名稱，不同層且有格線聯繫的上下概念，代表“is-a”的從屬關係，例如圖中的正規概念“Knowledge Building”, “Knowledge Base”及“Knowledge Engineering”等三項，是在正規概念“*Ontology*”之下的主要概念，因此圖 9 為概念的階層關係，亦即提供本體開發者參考的本體概念架構雛型。

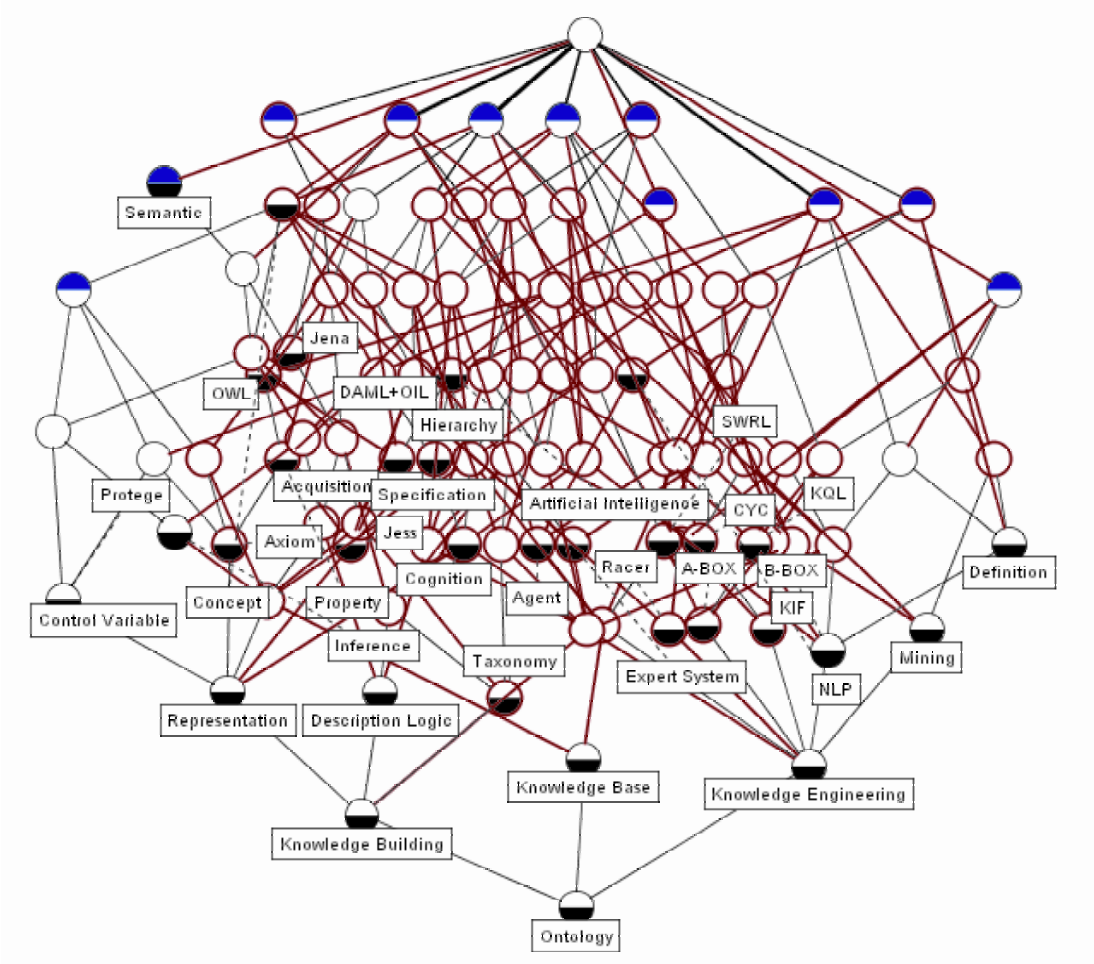


圖 9：本體概念架構以正規概念分析的格線圖呈現，最下階層表示 Root

表 1 為綜合四組不同情境的實驗數據。在「萃取領域術語」項目中，各數據表示所萃取的術語詞彙總數；在「統一概念命名」項目中，我們以 X→Y 表達：X 個元素進入凱利方格(RGT)並產出 Y 組概念，例如文章樣本數為 100 篇時，275 個元素進入 RGT 分析並產出 63 組概念；在「建立概念階層」項目中，X→Y 表達：X 個元素進入正規概念分析法(FCA)並產出 Y 組具有階層的概念，例如文章樣本數為 100 篇時，63 個概念進入 FCA 分析並產出 51 項具階層的概念(表 1 加註*符號者)，此即為最終產出之本體概念架構雛型。

表 1：以四組情境進行概念萃取程序的實驗

概念萃取程序 \ 分析數量		文章樣本數			
		n=80 篇	*n*=100 篇	*n*=120 篇	*n*=150 篇
		術語詞彙、元素或概念數			
萃取領域術語	文字探勘模組	229	243	262	284
	語言學分析模組	346	377	391	413
統一概念命名(凱利方格)		242→56	275→63	303→71	311→75
建立概念階層(正規概念分析)		56→44*	63→51*	72→54*	75→55*

由於本研究所提出的「概念萃取程序」，目前並沒有統一標準或通用的評估方法可供利用，因此本研究參考 Jiang et al. (2003)的評估模式，以問卷調查方式進行「主觀接受率」評估。我們邀請曾修習知識相關課程及至少有一年本體建置經驗的 23 位知識工程師進行評估實驗，評估者以前述四組文章數量別的電子期刊摘要為建置的參考來源，分別各利用 10 天建置其理想的本體概念架構。在完成傳統方式的本體概念建置之後，我們再給予每位建置者本研究的結果進行實體比較，並設計調查表供其評估差異情況。調查表針對所選取的概念及階層架構進行逐項的調查，評估方式是以李克特五點量表(Likert's five-point scale)計算，所有項目由非常不同意至非常同意(共分五級)程度分別給予 1 至 5 的分數，當受測者反應為 4 或 5 時，將被視同建置者在主觀上接受該項目，受測者的「主觀接受率」是累計其接受項目的次數除以全部的項目。本研究的評估標的包含概念選取及階層關係等二部份。經統計所有評估者的主觀接受率，我們以 Mean ± (S.D.)呈現如表 2 所示，其中文章樣本數大於等於 100 篇時，在概念選取及階層關係的主觀接受率皆超過 70%。

表 2：知識工程師對 4 種本體概念架構雛型在概念選取及階層關係的主觀接受率

項目別 \ 數量別	文章樣本數			
	n=80 篇	*n*=100 篇	*n*=120 篇	*n*=150 篇
	主觀接受率 Mean ± (S.D.)			
概念選取	68.42± (2.6)	75.42± (5.4)	77.81± (5.2)	78.67± (5.4)
階層關係	67.14± (5.6)	73.27± (3.8)	73.92± (2.0)	74.66± (2.2)

捌、討論與結論

本研究針對目前知識工程人員在建立本體時的問題，特別是如何由文件資料中建立本體定義，深入分析其問題關鍵、解決方案設計及可行的分析方法。在處理文件資料的情境中，本體概念通常隱含在非結構化的文件中，因此必須藉由反覆的分析才可獲得概念架構，萃取工作須分階段逐項實施。建置一個本體的概念架構除須界定領域

的範圍外，另須限定問題的構面，故須納入知識提供者對概念的認知與詮釋，對知識系統的應用而言，概念不僅是稱為什麼詞彙的術語，而是它在特定前提下究竟表達的意義是什麼，例如在一般的情況下，“Bank”可能用於是表達銀行的概念，但若是在不同的領域或問題構面下，“Bank”可能是表達河岸的概念，因此詞彙並不能全等於概念，而完善的概念定義更包括如語意、關係及階層的表達等，這也是過去研究認為概念建置是工藝遠勝於科學技術的主要原因。

本研究提出的「概念萃取程序」包含了三項主要程序，在分析方法的選擇及銜接上是以允許「人為參與」為前提，以期提供符合本體概念架構發展的要求，為補足因人為因素所引起的錯誤，因此須納入具有計算或邏輯推導能力的分析方法。例如凱利方格的構念引出即可修正涵意模糊的概念，正規概念分析法中的屬性探索(Attribute exploration)利用逐項計算屬性資料之間的隱性關係，修正人為判定所發展的矛盾或差異。

本研究提出的概念萃取程序已在第七節中以 SDOS 電子期刊資料庫的文章摘要進行驗證，以下我們討論三項與實驗進行有關的問題：

(1) 測試文件的合理數量。本研究並不建議實際的文件數量，我們認為合理的數量仍應視分析之標的物不同而須彈性調整，例如原始的資料來源若為內容鬆散的資料(例如網頁)，此時宜增加文件數量；反之，若資料來源的性質單純，則可酌量減少文件數量。另外，本研究曾以不同的文件數(80 篇、100 篇、120 篇、150 篇)區隔為四組情境，並在相同的條件下進行實驗，由表 1 的結果顯示：在萃取領域術語階段(包含文字探勘模組及語言學分析模組)，萃取的詞彙術語依四組情境而明顯遞增，但在進入後續的統一概念命名及建立概念階層等階段時，我們利用了如凱利方格及正規概念分析法等方法，將詞彙術語逐步轉換為概念，因此在最終的概念數量上雖然呈現遞增，但其增額差異已相差不大且呈收斂狀態。

(2) 門檻值的選定問題。門檻值的高低可做為控制萃取詞彙術語的數量，本研究也不作建議應選定的門檻值，仍須視實務的狀況而定，但我們提供兩項參考原則：(a).視領域及問題構面二者的大小，若以象限分佈來搭配，共可獲得如{大大、大小、小大、小小}等 4 組，範圍愈大者可考慮提高門檻值；(b).視分析文件的內容及來源性質，若內容及來源均鬆散則應可考慮提高門檻值。以本研究分析的文件為例，領域及問題構面二者均屬小範圍，可考慮降低門檻值，但文件的內容偏向廣泛(我們僅以 Ontology 篩選)，可考慮提高門檻值，綜合兩項原則，最後給予傾向中性偏高的門檻值。

(3) 主觀接受率的問題。我們曾邀請 23 位具有本體建置能力的知識工程師，對本研究的本體概念架構雛型進行主觀接受率評估調查。由表 2 的統計結果顯示：文件數量的多寡確實與主觀接受率呈正相關，但其主觀接受率的提升幅度亦明顯縮小。Mineau et al. (2000)即曾指出：概念塑模的建置過程，普遍面臨完整性與複雜度兩者之間的犧牲交換(trade-off)，當完整性到達一定的程度後即面臨瓶頸。由於過多的文件數量確實造成分析上的困難，因此我們建議達到 70%的主觀接受率是較經濟的做法。

概念萃取是整合三項步驟的集成，過程中或有文件數量別、門檻值高低、詞彙多寡等的選擇，而不同的選擇勢必影響各階段產出物的數量，然而萃取程序的設計主軸是由文件、詞彙、概念到最終的概念架構雛型，因此在本研究第二、三階段所採用的方法，均具有修正上述不同選擇所造成的差異影響，使得最終的概念架構雛型能趨向穩定(如表 1 所示)，主觀接受率的調查對象是最終的概念架構雛型，因此在過程中的數量差異將不致大幅影響主觀接受率。本研究並未宣稱由概念萃取程序所獲得的「概念架構雛型」可做為最終的領域本體定義，由於傳統塑模本體概念架構的期程過長，因此本研究發展的萃取程序確可提供建置者在建置初期的重要參考，並依據獲得的概念架構雛型再行精進為最終且正式的概念架構。本研究特別對於某些缺乏分類且須以文字資料為主要參考來源的領域，提供有效的本體概念建置協助。

誌謝

本研究感謝審查委員的修正建議及國科會一般型研究計畫提供之部份經費支援(計畫編號：NSC95-2416-H-033-009)。

參考文獻

1. Alcala, R., Cano, J.R., Cordon, O., Herrera, F., Villar, P. and Zwir, I. "Linguistic modeling with hierarchical systems of weighted linguistic rules," *International Journal of Approx. Reasoning* (32:2-3) 2003, pp: 187-215.
2. Batty, D. and Kamel, M.S. "Automating knowledge acquisition: A propositional approach to representing expertise as an alternative to repertory grid technique," *IEEE Transactions on Knowledge and Data Engineering* (7:1) 1995, pp: 53-67.
3. Chandrasekaran, B., Josephson, J.R. and Benjamins, V.R. "What are ontologies, and why do we need them?" *IEEE Intelligent Systems* (14:1) 1999, pp: 20-26.
4. Chung, M. and Moldovan, D.I. “Parallel natural language processing on a semantic network array processor,” *IEEE Transactions on Knowledge and Data Engineering* (7:3) 1995, pp: 391-405.
5. Corcho, O., Fernandez-López, M. and Gomez-Perez, A. “Methodologies, tools and languages for building ontologies. Where is their meeting point?” *Data & Knowledge Engineering* (46:1) 2003, pp: 41-64.
6. Cordon, O., Herrera, F. and Zwir, I. “Linguistic modeling by hierarchical systems of linguistic rules,” *IEEE Transactions on Fuzzy Systems* (10:1) 2002, pp: 2-20.
7. Fernandez-Breis, J.T. and Martnez-Bejar, R. "A cooperative tool for facilitating knowledge management," *Expert Systems with Applications* (18:4) 2000, pp: 315-330.

8. Ford, K.M., Petry, F.E., Adams-Webber, J.R. and Chang, P.J. “An approach to knowledge acquisition based on the structure of personal construct systems,” *IEEE Transactions on Knowledge and Data Engineering* (3:8) 1991, pp: 78-88.
9. Frakes, W.B. and Baexa-Tates, R. *Information Retrieval: Data structures and algorithms*, Prentice-Hall, Englewood Cliffs, 1992.
10. Gaines, B.R. and Shaw, M.L.G. "Knowledge Acquisition Tools based on Personal Construct Psychology," *Knowledge Engineering Review* (8:1) 1993, pp: 49-85.
11. Ganter, B. and Wille, R. *Formal Concept Analysis: Mathematical Foundations*, Springer-Verlag, New York, 1997.
12. Gillam, L., Tariq, M. and Ahmad, K. “Terminology and the construction of ontology,” *Terminology* (11:1) 2005, pp: 55-81.
13. Gruber, T.R. “A translation approach to portable ontologies,” *Knowledge Acquisition* (5:2) 1993, pp: 199-220.
14. Guarino, N. "Formal ontology, conceptual analysis and knowledge representation," *International Journal of Human-Computer Studies* (43:5) 1995, pp: 625-640.
15. Han, J. and Kamber, M. *Data mining: Concepts and Techniques*, Morgan-Kaufman, San Mateo, 2001.
16. Huhns, M.N. and Singh, M.P. "Ontologies for agents," *IEEE Internet Computing* (1:6) 1997, pp: 81-83.
17 Hui, B. and Yu, E. "Extracting conceptual relationships from specialized documents," *Data & Knowledge Engineering* (54:1) 2005, pp: 29-55.
18. Jiang, G., Ogasawara, K., Endoh, A. and Sakurai, T. “Context-based Ontology Building Support in Clinical Domains using Formal Concept Analysis,” *Journal of Medical Informatics* (71:1) 2003, pp: 71-81.
19. Kayed, A. and Colomb, R.M. "Extracting ontological concepts for tendering conceptual structures," *Data & Knowledge Engineering* (40:1) 2002, pp: 71-89.
20. Khedr, M. and Karmouch, A. “ACAI: agent-based context-aware infrastructure for spontaneous applications,” *Journal of Network and Computer Applications* (28:1) 2005, pp: 19-44.
21. Lassila, O. and McGuiness, D. “The Role of Frame-Based Representation on the Semantic Web,” *Stanford Knowledge Systems Laboratory Technical Report KSL-01-02*, 2001.
22. Lehmann, F. (ed) *Semantic Networks in Artificial Intelligence*, Elsevier Science Inc., New York, 1992.
23. Martin, P. and Eklund, P.W. “Knowledge Retrieval and the World Wide Web,” *IEEE Intelligent Systems* (15:3) 2000, pp: 18-25.
24. Mineau, G. W., Missaoui, R. and Godin R. “Conceptual modeling for data and knowledge management,” *Data & Knowledge Engineering* (33:2) 2000, pp: 137-168.

25. Minsky, M. "A framework for representing knowledge," In Winston, P.J. (ed) *The psychology of computer visions*, McGraw-Hill, New York, 1975.
26. Motta, E., Shum, S.B. and Domingue, J. “Ontology-driven document enrichment: principles, tools and applications,” *International Journal of Human-Computer Studies* (52:6) 2000, pp: 1071-1109.
27. Noy, N.F. and McGuinness, D.L. “Ontology Development 101: A Guide to Creating Your First Ontology,” *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05*, 2001.
28. Richards, D. and Simoff, S.J. “Design Ontology in Context- A Situated Cognition Approach to Conceptual Modeling,” *Artificial Intelligence in Engineering* (15:3) 2001, pp: 121-136.
29. Shamsfard, M. and Barforoush, A.A. “Learning ontologies from natural language texts,” *International Journal of Human-Computer Studies* (60:1) 2004, pp: 17-63.
30. Shaw, M.L.G. and Gaines, B.R. “Comparing Conceptual Structures: Consensus, Conflict, Correspondence and Contrast,” *Knowledge Acquisition* (1:4) 1989, pp: 341-363.
31. Srikant, R. and Agrawal, R. “Mining Generalized Association Rules,” *Future Generation Computer Systems* (13:2-3) 1997, pp: 161-180.
32. Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.-P., Studer, R. and Sure, Y. "Semantic Community Web Portals," *Computer Networks* (33:1-6) 2000, pp: 473-491.
33. Sugumaran, V. and Storey, V.C. “Ontologies for conceptual modeling: their creation, use, and management,” *Data & Knowledge Engineering* (42:3) 2002, pp: 251-271.
34. Uschold, M. and Grueninger, M. “Ontologies: principles, methods and applications,” *Knowledge Engineering Review* (11:2) 1996, pp: 93-155.
35. van der Vet, P.E. and Mars, N.J. “Bottom-up construction of ontologies,” *IEEE Transaction on Knowledge and Data Engineering* (10:4) 1998, pp: 513-526.
36. van Elst, L. and Abecker, A. “Ontologies for information management: balancing formality, stability, and sharing scope,” *Expert Systems with Applications* (23:4) 2002, pp: 357-366.
37. Wille, R. “Concept Lattices and Conceptual Knowledge System,” *Computers Math. Application Supporting Ontological Analysis of Taxonomic Relationships* (23:6-9), 1992, pp: 493-515.