

以 WordNet 及 FCA 為基礎進行領域知識本體之合併 以產生模糊化本體知識

陳榮靜

朝陽科技大學資訊科技研究所

暨資管系管理系教授

crching@cyut.edu.tw

葉峻儒

朝陽科技大學

資訊管理系研究生

s9514634@cyut.edu.tw

李宜晏

朝陽科技大學

資訊管理系研究生

s9614630@cyut.edu.tw

摘要

目前建構本體論的工具與方法很多，而建構出來的本體論也會有所差異，在相同的領域下，如果直接使用已被建構好的本體論，可以節省相當多建構本體論的成本，而在相同領域下，本體論的整合便成為一項重要課題。本篇論文主要提出由上而下以模糊正規概念分析為基礎及擷取 WordNet 資訊的本體論整合方法，稱為 FFCA-Merge (Fuzzy Formal Concept Analysis-Merge)，此架構的流程是先將需整合的本體論進行預先處理，並從 WordNet 內容中獲取資訊進行合併及校正，再者利用模糊概念分析技術來彌補 WordNet 無法完成的合併及校正工作，最後產生模糊本體論，目前領域本體論的合併方法主要還是針對不具有模糊概念的本體論，而透過 FFCA-Merge 合併之後，可以轉換成模糊本體論。在處理不明確資料時，模糊本體論具有更強大的彈性與優勢。

關鍵字： 本體論合併、本體論對應、模糊理論、正規概念分析、模糊本體論。

Abstract

Many different contents and structures exist in constructed ontologies even they are at the same domain. If the existed domain ontologies can be used, that will save a lot of time and cost. However, the domain knowledge changes fast.

In addition, the existed domain ontologies maybe need to be updated to solve the domain problems. How to reuse an existed ontology is an important topic for the application of ontologies. So, the integration of existed domain ontologies becomes an important subject. In this paper, we will propose a new method to combine WordNet and Fuzzy Formal Concept Analysis (FFCA) techniques to merge ontologies with the same domain called FFCA-Merge. Through the method, the ontology can be converted to a fuzzy ontology. The new fuzzy ontology has higher flexibility than a general ontology. The experiments results indicate our method can merge the domain ontologies effectively.

Keywords: Ontology merge, Ontology mapping, Fuzzy theory, Formal Concept Analysis, Fuzzy ontology.

1. 前言

本體論這項技術已被視為實現語意網路的重要關鍵，本體論的主要工作在於將獨立的知識進行概念性的分類、建立案例、設定各概念之間的關聯，進一步將含有語意的資訊加以保存，一個建構完備的本體論可視為一個具有語意的知識表達結構，利用本體論將知識做有意義上的分類，進一步的可以利用本體論中的

資訊進行推論，得到新知識。而實現語意網路其中一項重要的工作，便是許多獨立的本體論，透過有效率的方式整合起來，使各個面向的語意資源能夠整合成一個單一旦完備的語意網路

現今本體論的種類、建構方式、建構者的認知、建構語法等，有者相當大的差異性，可能單一領域之下的本體論數量就相當龐大，其中各個本體論間充斥者許多矛盾及衝突，而每個本體論可能都有其存在的價值與應用的優勢，所以如何整合本體論已成為相當重要的研究議題，也是本研究的最主要動機。

本體論整合的部分包含了本體論映射(Mapping)、本體論合併(Merge)、本體論校正(Alignment)等工作。合併的定義是將相異的領域本體論合併成單一旦一致的領域本體論，而校正的定義是將相異且互斥的概念找出其中正確的階層關係，另外映射代表的意義比較廣泛，具體來說，是找出相異本體之間的關聯，建立起彼此的連接(Link)，根據這些連接進一步的可以進行合併或是校正的工作。

整合本體論依不同觀點，各有不同畫分方式，Choi 等作者[5]在本體論整合的議題上，根據議題的不同，而將本體論劃分成三大類，包含了全域本體論(Global ontology)又稱上位本體論(Upper ontology)、區域本體論(Local ontology)、領域本體論(Domain ontology)。全域本體論泛指具有各個領域綱要(Schema)之本體論，它包含各種領域，只有具體綱要並無詳細內容；區域本體論為某一特定區域下本體論之綱要，各個區域本體論屬於不同的領域；領域本體論的範圍針對某一特定領域之下，而在定義上領域本體論具有詳細的概念分類，可視為一個建構較完整之本體論。

而在本體論整合的議題上，主要包含三大範疇[5]：

(1) 全域本體論與區域本體論二者之間的本

體論整合

在此類情形下之本體論對應，由於整合過的全域本體論下已有分享字彙，所以在區域本體論進行整合的情況下，其規則可以較簡單的被發掘及定義。但在現今高度動態變動的環境下，這類的本體論整合議題有延展性和維護性的問題，延展性受限制的主要原因為全域本體論中只要有一個區域本體論有更新、移除、新增的動作，就會影響到之後的本體論對應結果；維護性問題主要是因為在相同領域或相似領域中之本體論，可能存在不一致或互相矛盾的資訊。

(2) 區域本體論之間的本體論整合

此類整合主要是當不同領域的區域本體論之間，透過整合可讓區域本體論之間進行溝通，找出相異本體論之間的關聯。此類對應適合使用在現今高度動態、分散式的環境下，由於區域本體論整合是發生於區域上的，所以區域本體論之間的整合，也較全域本體論與區域本體論之間的整合更具延展性及維護性，但相對的，由於區域本體論間缺少全域本體論的分享字彙，所以在本體論整合的過程上較為困難。

(3) 領域本體論之間的合併(Merging)與校正(Alignment)

此類整合的議題乃是針對相同、重覆領域之本體論，對領域本體論來說，其內容是最豐富且完備的，在整合的過程中需考慮更多的面向，會受到同義詞、特殊名詞、二相異本體論的結構、二相異本體論概念中等，許多矛盾與衝突點的混淆，而如何找出領域本體論之間的相關性及相似性，避免掉上述的矛盾與衝突，在現今本體論整合中，也是一個重要議題。

本篇研究中，主要是針對相同領域本體論間的合併進行探討。表 1 為相同領域本體論間合併、校正技術之比較表。

表 1 各合併、校正技術之比較表

文獻	主要目的	方法	優缺點
1.SMART[16]	合併校正	- 語言學上之相似度(同義字、共用之子字串、句首、句尾)。 - class-name 之比對。	- 半自動化。 - 將語義分析後之結果呈現給使用者，利於進行合併。
2.PROMPT[18]	合併校正	語言學上相似度之比對。	- 半自動化。 - 將語義分析後之結果呈現給使用者，利於進行合併。
3.OntoMorph[2]	合併	- 提供強力的規則語言。 - 結合二種技術：語法改寫和語義改寫。	以強大的規則語言來敘述其本體合併，這對使用者來說可能需要較高的知識門檻。
4.CHIMAERA[15]	合併校正	- 基於 Ontolingual ontology editor 此工具下之合併技術。 - 如果能找出語言學上的配對就能達到自動化合併。	- 全自動化。 - 在不明確的資料中，要完全的找出全部的語言學配對，以達到其全自動化，會有其困難性。
5.HICAL[11]	合併	- 為一 machine-learning 的方法。 - 擷取資料中之實例。	由於不是比較二個字之間的相似度，所以可能將不同的字元分類到同一個概念下。
6.Anchor-PROMPT[17]	合併校正	找出有關聯的關鍵字集合，進行深入研究，以驗證相關性及找出新的相關關鍵字。	- 半自動化。 - 將語義分析後之結果呈現給使用者，利於進行合併。
7.FCA-Merge[26]	合併	- 例子擷取並產生本文。 - 利用 TITANIC 演算法計算出刪減過後之概念點陣。 - 透過專家的知識將刪減過後之概念點陣轉換成本體。	- 半自動化。 - 由概念點陣轉換為本體時，需要人工介入。

從表 1 中可以發現，現今的合併及校正技術，整體來看有二大缺點：(1) 大部分系統的出發點都是以輔助使用者為主，並不是以自動化整合為目標、(2) 合併後本體論不具有模糊本體論之概念。本研究是以自動化整合為出發點，以機器取代人力，在整合的過程中將不需要使用者的參與，最後可以計算出具有模糊資訊的本體論來增強彈性與優勢。

論文其他部分包括第二節相關技術的介紹，第三節模糊正規概念為基礎的本體論合併

方法之詳細流程架構，第四節實例說明模糊正規概念為基礎的本體論合併方法，最後第五節為結論。

2. 相關技術

以下子章節將針對FFCA-merge所使用到的相關技術及資源，做詳細的介紹。

2.1 WordNet

WordNet[9]是一個大型的英文語意詞

典，也是一個詞彙的本體論，是由Princeton大學認識科學實驗室(Cognitive Science Laboratory)自1985年以來的開發工作，是由心理學教授George A. Miller所帶領的團隊進行指導及維護。在WordNet中有英文詞彙中的各種詞性，例如名詞、動詞、形容詞、副詞，也包含了詞類的各式關係詞，Hyponym(上位詞)、Hypernym(下位詞)、Antonym(反義詞)、Synsets(同義詞)等等。其中又包含各領域的專門類別及詞彙，是一個建構相當完備的知識庫，而SUMO (Suggested Upper Merged Ontology)[32]這個上位本體論也不斷地將其中的詞彙與WordNet進行映射的工作。WordNet中的知識具有高度的標準化，而研究中就是利用WordNet中的資訊來進行整合。圖1為WordNet中包含的資訊範例。然而WordNet上的資訊更新，仍須由人工處理，對於知識變動快速的現今社會，仍存在更新速率不足的問題。

#S: (n) → 詞性 <u>Trial, trial run, test, tryout</u> → 同義詞 <u>(trying something to find out about it)</u> → 定義 <u>"a sample for ten days free trial"</u> → 例句
--

圖 1 WordNet 中的詞彙資訊

2.2 正規概念分析 (Formal Concept Analysis)

本論文提出的 FFCA-Merge 主是要可以將兩個來源本體論進行合併，進而轉換為模糊本體論，使用到的相關技術一部分是基於正規概念分析[8]發展出來的，而正規概念分析的應用相當廣泛，像是文件分析[19][20][25]、本體論建構[21][22]、本體論合併[7][26]、線上學習[24]、專家系統[23]、郵件管理[6][12]、網站導航[14]等。正規概念分析的主要功能在分析資料，使資料形象化、結構化，分析過後之資料將更具有意義及結構[27]。

正規概念分析是根據點陣理論(Lattice theory)[1]所發展出來的，而正規概念分析的建

構步驟大致如下：

- (1) 建構 context
- (2) 建構 concept lattice

本文的定義，包含三個元素 (G, M, I) ， G ：本文的物件群； M ：本文的屬性集合； I ：物件和屬性間的關係。本文中屬性可視為關鍵字(Terms)或特性(Properties)；物件可視為文件(Documents)或輸入(Entities)。而建構本文的目的為找出物件和屬性間的關聯。

在正規概念中還有延伸出二個概念，分別為屬性的內涵(Intent)及物件的延伸(Extent)。

其中屬性的內涵是指一組物件應共同具有的屬性集合。若 A 為一組物件的集合並滿足 $A \subseteq G$ ， A' 稱為 A 的內涵， A' 是指一組屬性的集合，為 A 中所有物件均有的屬性，以下式表達：

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

物件的延伸(Extent)是指滿足特定屬性的物件集合。若 B 為一組屬性的集合並滿足 $B \subseteq M$ ， B' 稱為 B 的延伸， B' 是指一組物件的集合，為滿足所有在 B 中共同的屬性，以下式表達：

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

建構本文後即可建構出概念點陣，概念點陣主要是建構出概念間的繼承關係(Inheritance relation($\leq$))。若從本文中取出二個概念 $(E1, I1)$ 、 $(E2, I2)$ 並建立其之間的繼承關係， $(E1, I1) \leq (E2, I2)$ ，可稱 $(E1, I1)$ 繼承 $(E2, I2)$ ，在概念點陣中稱 $(E1, I1)$ 為子概念(SubConcept)，而 $(E2, I2)$ 為全概念(SuperConcept)。

而建構概念點陣的目的，則為找出概念間的階層關係，具有越多相關物件的屬性，視為越重要，隸屬於越高的層級。底下利用表2及圖1來說明正規概念分析的步驟。

表2 FCA-本文

屬性 文件	球類	足球	武術	柔術
D1	×	×		
D2	×			
D3			×	×
D4			×	×
D5			×	

從表2中可以發現只要文件與屬性間具有相關性，在本文中以×號表示之，經過本文之建立後就能進一步的建構概念點陣。

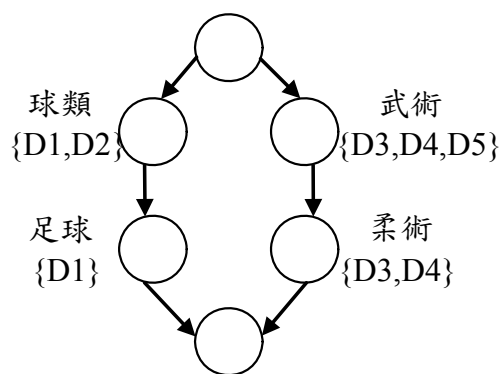


圖 2 FCA-概念點陣

由圖 2 概念點陣中可以發現，本文中具有越多物件相關之屬性，會隸屬於越高層級。例：球類為足球之全概念，而足球為球類之子概念。

2.3 模糊正規概念分析 (Fuzzy Formal Concept Analysis)

Quan 等作者[21][22]提出了將模糊理論合正規概念分析的方法自動建構本體論及其應用。而模糊正規概念分析在本文中之定義，包含三個元素 $(G, M, I = \alpha(G, M))$ ， G ：本文的物件群； M ：本文的屬性集合； I ：物件和屬性之間的關係，其中 I 關係包含了物件和屬性間隸屬值的概念。

模糊正規概念分析與正規概念分析最大的不同點在於，每一個物件和屬性間之相關性

都是具有隸屬值的概念存在，在一個物件中出現頻率越高的屬性就給予越高的隸屬值，而概念之間的關聯也可以利用模糊理論來計算出其相似度。

以下利用表 3、表 4 及圖 3 來說明模糊正規概念分析的步驟。

表3 FFCA-本文中所有隸屬值資訊

屬性 文件	球類	足球	武術	柔術
D1	0.9	0.7	0.4	0.12
D2	0.8	0.33	0.15	0.3
D3	0.11	0.25	0.85	0.7
D4	0.21	0.13	0.85	0.6
D5	0.18	0.1	0.95	0.25

設 $\alpha - cut = 0.6$ ，將數值高於 0.6 的隸屬性視為較有意義：

表 4 FFCA-將本文中低於 0.6 之隸屬值去除

屬性 文件	球類	足球	武術	柔術
D1	0.9	0.7		
D2	0.8			
D3			0.85	0.7
D4			0.85	0.6
D5			0.95	

模糊正規概念分析中具有其隸屬值之概念，可經由 $\alpha - cut$ 的設定，將隸屬值較低之正規概念去除，進一步的得到概念點陣。

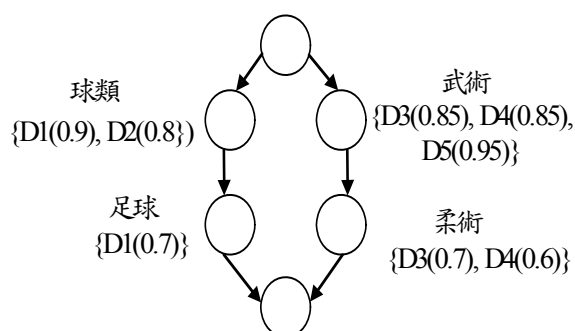


圖 3 FFCA-根據表 4 建構之概念點陣

概念點陣間的每一個正規概念，其包含和物件間的隸屬值資訊，透過這些資訊可以更彈性的分類其正規概念，也可以利用這些隸屬值資訊，計算出正規概念間之相似度。

2.4 FCA-MERGE(Formal Concept Analysis -Merge)

Stumme 和 Maedche[26]在 2001 年提出了 FCA-MERGE 這個技術，主要是將二個相同領域本體進行合併，此方法會進行三個步驟：

- (1) 對二個相同領域之本體進行語意分析並各自產生本文。
- (2) 合併二個領域本體之本文後，產出被修整過之概念點陣。
- (3) 根據修整過之概念點陣產出一個新的本體。

FCA-MERGE 詳細流程如圖 4：

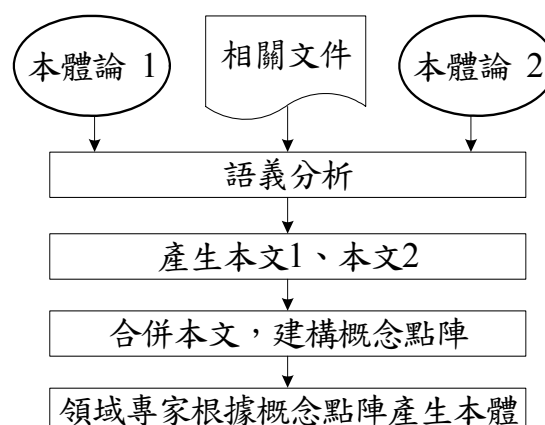


圖 4 FCA-MERGE 流程圖

從上述流程可以發現，FCA-MERGE 在最後一個流程，必須花費大量的人工，需要經由領域專家的背景知識根據概念點陣進行最後的本體論產生；相對而言，我們提出的 FFCA-MERGE 具有隸屬值的概念點陣，將可以大幅度的降低其人工作業，再藉由門檻值之設定，可將相關性低的概念去除，使合併後之本體論更有彈性，而整個流程已接近自動合併本體論之目的。

3 研究架構

圖 5 為本研究提出的 FFCA-Merge 架構圖，在整合的過程中，為了達到重覆利用的目的，及減少定義上的衝突與矛盾點，首先會選擇一個建構較完備、較無爭議、標準化較高的本體論做為基準(命名為基底本體論)；而在不改變基底本體論的前提下，將另一個本體論(命名為修正本體論)透過合併及校正的方式加入其中。

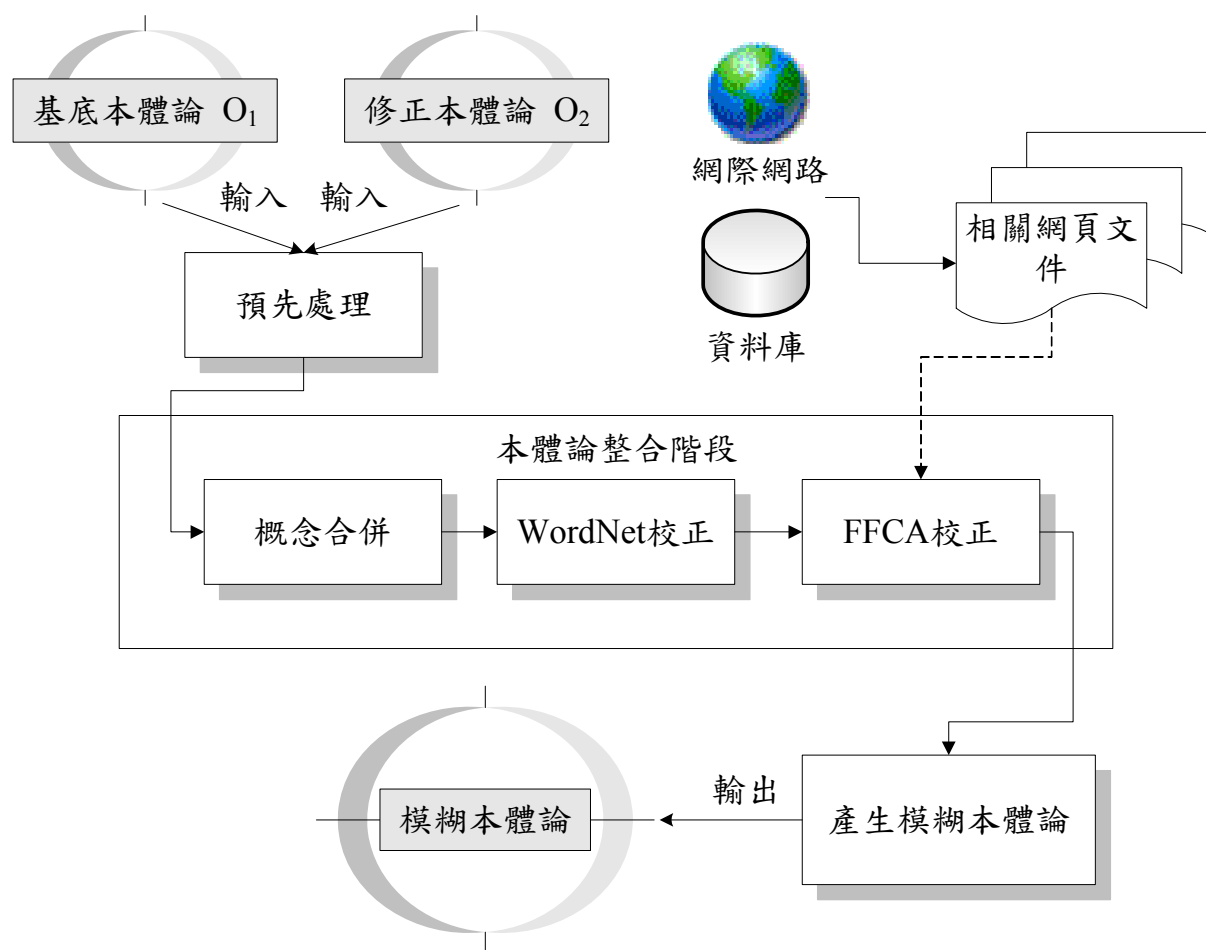


圖 5 FFCA-Merge 架構圖

在整個研究的流程中，包括預先處理，本體論整合階段，分別為合併類別名稱、WordNet 校正、FFCA 校正，其中 FFCA 校正階段需要經由相關網頁文件的收集來達到整合的目的，最後轉換成模糊本體知識。以下章節會對各個流程做詳細的說明。

3.1 預先處理

預先處理這個階段，主要皆為完成本體論整合階段所做的前置作業，其中主要工作有三項：

(1) 加入同義詞及定義標籤：從 WordNet 中擷取並儲存本體論中各個概念的同義詞及定義標籤，而 WordNet 中對特定領域其同義詞的

內容可能有限，在此可以再透過特定領域的同義詞字庫來達到支援的效果，而此階段主要目的為達到概念合併階段的自動化處理。

(2) 詞幹還原：在英文字詞中，詞尾的部分會有一定規律的變化，但其代表的含意不變，像是名詞與動詞會以不同型態出現，例如單、複數型，現在式、過去式等等。有些單字更可延伸有形容詞、副詞的變化，而詞幹還原的目的就在於將字詞的詞尾變化都還原成最原始的狀態，例如 *test*, *tests*, *testing*, *tested* 雖然有著不同的型態，但都代表著同一種概念，而透過詞幹還原可以將上述四種狀態還原為最原始的狀態 *test*。但在實作中，將概念中詞幹還原後可能會造成混淆，原因為某些特殊詞類

如果經過還原後可能會與其他詞類有重覆的情形發生，例如 *software* 透過 Paice/Husk 演算法進行還原的原始狀態為 *softw*，而在網頁檔中比對出現頻率時可能會將 *softwood* 這個完全無關的概念視為同義，而為了避免此類的混淆，在實作中選擇將所有詞幹變化儲存起來的策略。

(3) 篩選網頁文件：在 FFCA 校正的階段中，其整合正確性高低取決於收集網頁文件的質與量，如果在收集的網頁文件中包含了過多的雜訊，那麼不斷地增加網頁文件數量也不一定能讓正確性有所提升，故需要對網頁文件進行篩選的動作，而在大部分的研究中都是先將文件檔案進行分類，以求去除不相關的文件，但此種方式所耗費的成本較高，在本研究中採用一個快速的篩選法，讓網頁文件的質量提升。其主要方式為在 FFCA 校正的階段中，網頁文

件導入的數量以某個倍數成長，在正確率到達所設定的門檻值之後，將最後導入文件數量的區塊保留，其餘的網頁文件視為雜訊過高必須移除。用實驗部份的實例做一個簡單的說明，實驗中隨機收集了 3243 筆網頁文件，並無經過篩選的步驟，其中可能包含許多雜訊，在 FFCA 校正時將網頁數量以 2 倍數成長，觀察正確率的變化，初始網頁文件數量為 100，而在網頁數量 3200 的時候正確率高過一定的門檻值，只保留最後一個區塊編號 1600 到 3200 網頁文件。而在實驗的部分我們將 3243 筆資料視為 3200 筆進行導入，也會針對此篩選法進行比較，證明此方法對文件質量有顯著提升。

前二項前置處理工作架構以圖6表示，圖中概念內容裡三項灰色區塊為經過前置處理後，於概念中新增的資訊。

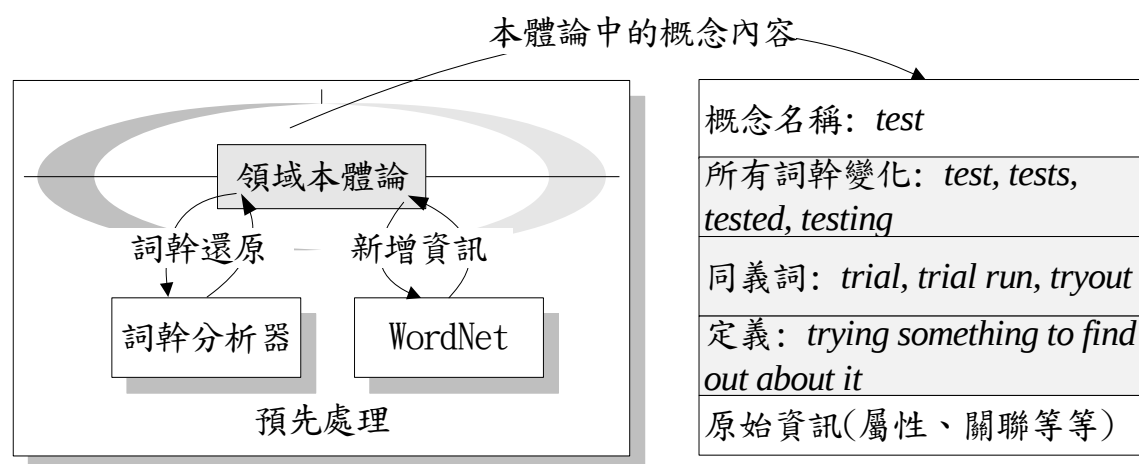


圖 6 前置處理架構

4. 本體論整合階段

此階段為本篇研究的核心，整合的原則有二項，其一採用由上而下的整合方式，此目的為降低整合過程中的複雜度，主要因為當上層概念可以被合併或校正後，將此概念的下層概念直接繼承原先架構，因為將本體論中概念階層架構都視為是正確的，故不需做任何改

變；其二不更改基底本體論中的階層架構，只針對修正本體論中的內容進行調整，此目的在降低整合過程中的複雜度，及避免認知不一的矛盾與衝突點。整合階段中主要分為三個部份來達到本體論整合的目的，以下子章節針對各階段做詳細說明。

4.1 概念合併

在此階段主要是利用預先處理後，於概念中新增的資訊，來達到概念合併的目的，而 WordNet 扮演者詞義分析的角色，透過 WordNet 中的定義及同義詞資訊，只要概念名稱相同，或具有相同同義詞並且定義相同，則視為意義相同的概念，需進行合併，而此類概念我們定義為 Equal concept，定義如下：

(1) Equal concept

$$\begin{aligned} \text{Equal} := \{ & (c_1, c_2) \mid c_1 \in O_1 \wedge c_2 \in O_2 \wedge O_1 \neq O_2 \wedge \\ & (\text{Class name}(c_1) = \text{Class name}(c_2)) \vee \\ & \text{Synsets}(c_1) \cap \text{Synsets}(c_2) \neq \emptyset \wedge \\ & \text{Definition}(c_1) = \text{Definition}(c_2) \} \end{aligned}$$

透過上述處理，本體論中的所有同等概念視為合併完畢，而接下來的工作為找出概念之間正確的階層架構，進行校正。

4.2 WordNet 校正

此階段利用 WordNet 資料庫中的資訊來達到校正的目的，同樣地根據 WordNet 中對關係詞的定義，利用其中三項較廣泛的詞類關係，定義如下：

(2) Hyponym concept

$$\text{Hyponym} = \{ (c_1, c_2) \mid c_1 \in O_1 \wedge c_2 \in O_2 \wedge O_1 \neq O_2 \wedge ((c_2 \leq c_1) \in W) \}$$

(3) Hypernym concept

$$\text{Hyponym} = \{ (c_1, c_2) \mid c_1 \in O_1 \wedge c_2 \in O_2 \wedge O_1 \neq O_2 \wedge ((c_2 \leq c_1) \in W) \}$$

(4) Sibling concept

$$\begin{aligned} \text{Sibling} := \{ & (c_1, c_2) \mid c_1 \in O_1 \wedge c_2 \in O_2 \wedge O_1 \neq O_2 \wedge \\ & (\text{Sister term}(c_1, c_2)) \vee \\ & (\text{Superconcept}(c_1, d) \cap \\ & \text{Superconcept}(c_2, d) \neq \emptyset) \} \end{aligned}$$

上述四種詞類關係，利用圖 7 說明之，在兩相異本體論中有二個概念分別為 *Test* 與 *Tryout* 透過預先處理中的新增資訊，符合 equal concept 的定義，先將這二個概念進行合併；*Problem solving* 在 WordNet 資料庫的定義中是屬於 *Inquiry* 的上層概念也就是上位詞，符合 Hypernym concept 的定義；*Experiment* 在 WordNet 資料庫的定義中是隸屬於 *Inquiry* 的下層概念也就是下位詞，符合 Hyponym concept 的定義；*Pilot experiment* 與 *Test* 在 WordNet 資料庫的定義中，兩者是屬於同一層級的概念具有兄弟的關係，符合 Sibling concept 的定義。經過上述四項詞類關係的定義，兩相異本體論之間的階層關係，便有依據可以進行合併或校正的工作，而整合後的本體論結果如圖 7 所示。

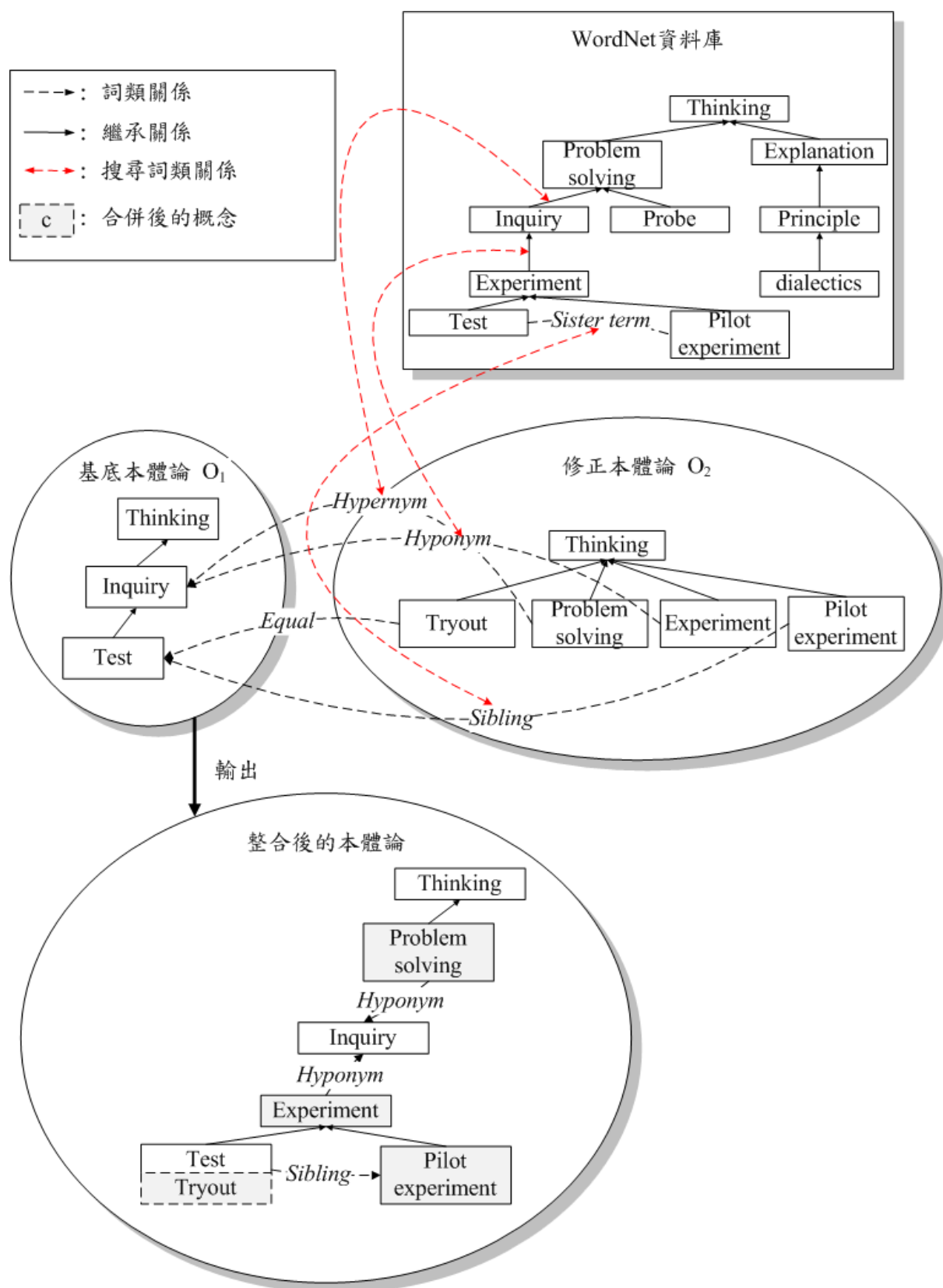


圖 7 獲取詞類關係說明圖

在理想的情況底下，如同上述的範例，可以獲取到全部的資訊，但實際的情形下，本體論中的概念並不一定全部存在於 WordNet 資料庫中，故 Hypernym concept 與 Hyponym concept 定義中的搜尋過程，不限定於單一層級，也就是說會不斷地往上層或下層搜尋，有祖父關係、子孫關係的上下位詞概念，也視同於符合 Hypernym concept 與 Hyponym concept 的定義，並會進行校正的工作；另外 Sibling

concept 的定義較為特殊，由於概念並不一定全部存在於 WordNet 資料庫中，故將 Sibling concept 做較有彈性的定義，以圖 8 說明，圖中 Part 1 的部分為理想的 Sibling concept 關係，而 Part 2 的例子可以發現， C_1 與 C_2 概念共有相同上層概念，且 C_1 與 C_2 二者之間並無 Hypernym concept 與 Hyponym concept 的關係，對於此類的概念在定義中也歸納於 Sibling concept。

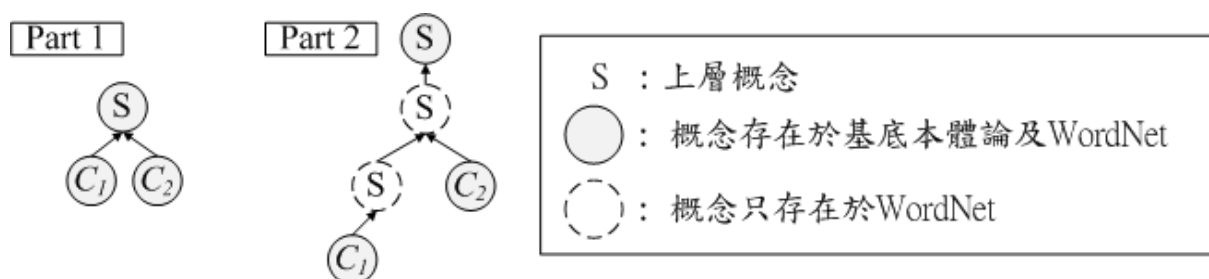


圖 8 Sibling concept 關係說明圖

4.3 FFCA 校正

前兩階段的整合工作，主要參照 WordNet 資料庫中的資訊進行整合，但領域本體論時常出現特定領域的特殊詞彙與組合詞，在 WordNet 資料庫中沒有此類詞彙的資訊可以使用，為了彌補這項缺失，我們利用模糊概念分析的技術，來找出此類概念的整合結構。

圖 9 為 FFCA 校正的流程圖，首先會針

對基底本體論及修正本體論尚未整合的概念，分別建立各自的 Context，進行聯集後計算出各 Context 的隸屬值，透過 α -cut 的設定，將相關性低的文件去除。接著計算未整合概念與基底本體論的相似度，找出最大相似度的概念，透過轉換成 Concept lattice 找出最適當的階層位置進行校正。以下子章節會針對流程細部做詳細說明。

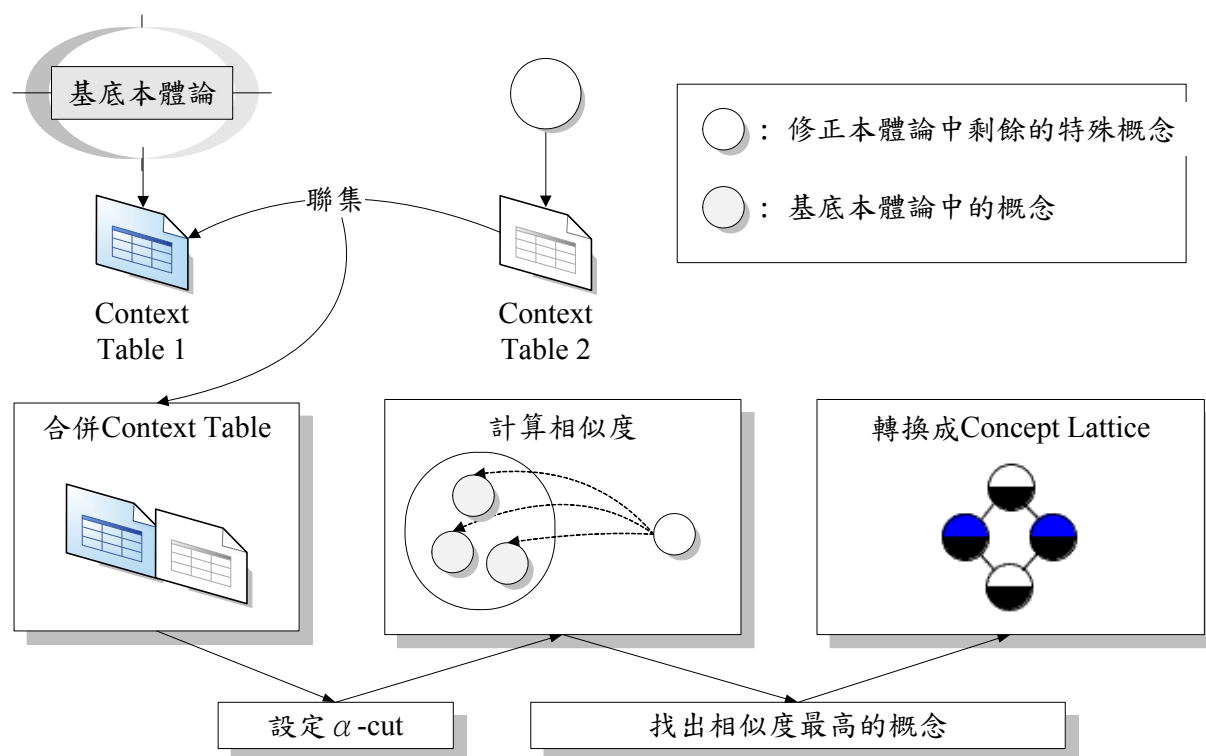


圖 9 FFCA 校正流程圖

(1) 合併 Context Table

FFCA-merge 整合架構中的 FFCA 校正，主要是參照收集而來的相關網頁文件中，找尋相異領域本體論概念之間的關聯，此階段主要是將各 Context Table 進行聯集後，最主要目的為計算出 Context 內的隸屬值來代表每一概念的重要性，設定方式採用 TF(Term Frequency) 與 TF-IDF(Term Frequency-Inverse Document Frequency)分別計算出每一個概念的隸屬值程度，在實驗的部分會對此二種隸屬值的計算方式，比較二者之間的優劣。

詞彙頻率(TF)是代表詞彙出現在所有網頁文件中的相對頻率，也就是代表在網頁文件中出現頻率越高的概念，具有越高的重要性；而文件頻率(DF)是代表全部網頁文件中，出現此概念詞彙的網頁文件數量，而概念詞彙出現於較少數的文件，其區別性越強，越能代表其重要性，故會採用逆文件頻率(IDF)來計算該值。TF 及 TF-IDF 的計算公式如

$$TF_j(a_i) = \frac{fre(a_i)}{\text{MAX}\{fre(a_i)\}}$$

$$TF - IDF_j(a_i) = TF_j(a_i) \times \log \frac{N(D_i)}{N(a_i)}$$

i ：為概念編號，由 1 到 i 。

j ：為文件編號，由 1 到 j 。

$TF_j(a_i)$ ：代表 TF 隸屬值計算方式，為概念 a_i 在網頁文件 j 中的重要性，其值為介於 0~1。

$TF - IDF_j(a_i)$ ：代表 TF-IDF 隸屬值計算方式，為概念 a_i 在網頁文件 j 中的重要性，其值為介於 0~1。

$fre(a_i)$ ：為概念 a_i 出現在網頁文件中的頻率。

$IDF_j(a_i)$ ：計算方式為 $\log \frac{N(D_i)}{N(a_i)}$ ， $N(D_i)$ 為

全部網頁文件數量， $N(a_i)$ 為出現過概念 a_i 的

網頁文件數量。

概念與網頁文件之間有了隸屬值資訊後，再透過 α -cut 的設定將相關性弱的文件去除，減少低相關文件的干擾，以提升整合的正確性，而在實驗的部份會針對 α -cut 值的設定進行調校，找出特定領域之下最合適的 α -cut 數值。

(2) 計算相似度

透過合併 Context Table 的階段處理後，利用概念與網頁文件之間的隸屬值資訊，可以計算出兩相異概念之間的相似程度，因相似度的計算公式考慮了結構性的資訊，故會將每一個上層概念及其下層概念進行平均值的動作，其公式如下：

$$\text{similarity}(BO, RC) = \frac{\sum_{k=1}^n \frac{|BO_k \cap RC|}{|BO_k \cup RC|}}{n}$$

BO：為基底本體論，包含其所有概念。

RC：為修正本體論中，尚未整合的所有概念。

n：為基底本體論中概念層級的深度。

BO_k：為基底本體論中第 k 層級的所有概念。

在相似度的計算採用了二種不同的策略：

- i. 獨立層級：根據由上而下的原則，只計算出基底本體論中最上層概念與未整合概念的相似度，而後再利用轉換成 concept lattice 來找出相似度最高的概念集合，以找出未整合概念應該校正於上

層概念及其下層概念中的階層關係，此策略較不易受到網頁文件中雜訊的干擾，正確性較高。

- ii. 整體層級：計算每一概念間的相似度，最後只需將未整合概念校正於相似度最高的概念底下即可，此一策略無須經由 concept lattice 的轉換，其校正成本較低，但容易受到網頁文件中雜訊的干擾，影響其正確率。

接下來利用圖 10 來說明相似度的計算方式與上述二種策略的不同之處，圖中 Brazilian Jujitsu 為修正本體論中未整合概念，根據由上而下的整合原則，計算第一層級的概念 Football game 與 Martial art，相似度計算過程如下：

$$\begin{aligned} \text{similarity}(\text{Football game}, \text{Brazilian Jujitsu}) &= \frac{0.15}{0.9+0.8+0.7+0.35} + \frac{0}{0.7+0.7+0.35} + \frac{0}{0.5+0.7+0.35} \\ &= \frac{0}{3} \\ &= 0.01 \end{aligned}$$

$$\begin{aligned} \text{similarity}(\text{Martial art}, \text{Brazilian Jujitsu}) &= \frac{0.7+0.35}{0.85+0.85+0.95} + \frac{0.7}{0.7+0.6+0.35} = 0.9 \end{aligned}$$

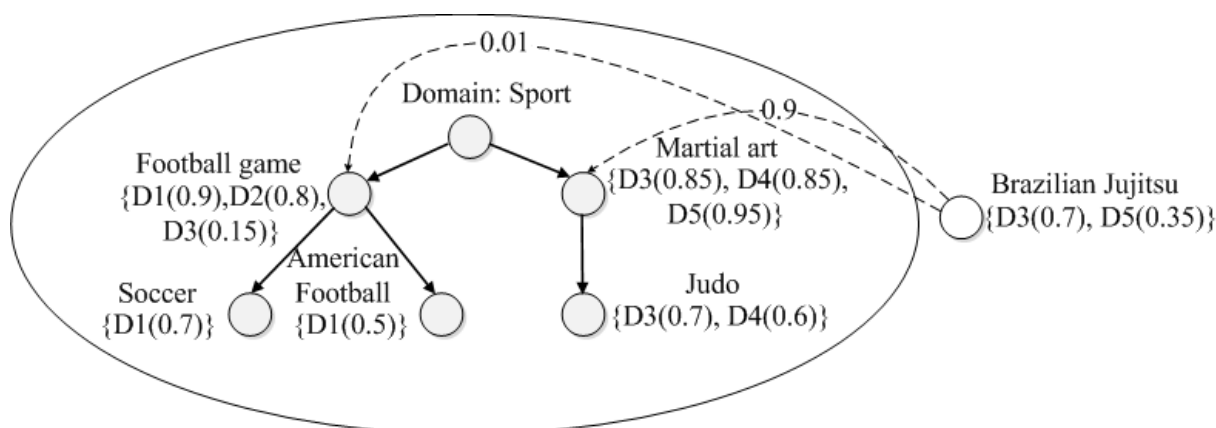


圖 10 相似度計算圖例

根據獨立層級的策略，找到最相似的上層概念 Martial art，再將其轉換為 concept lattice 找出其正確的整合位置。如果是採用整體層級的策略，便會將 Brazilian Jujitsu 分別與 Football game、Martial art、Soccer、American Football、Judo 計算其相似度，最後找出 Brazilian Jujitsu 乃是與 Martial art 相似性最高，便會直接將概念校正至 Martial art 概念底下，在實驗的部分也會針對這二種策略進行比較。

(3) 轉換成 concept lattice

採用獨立層級策略，在找到最相似的上層概念後，必須轉換成 concept lattice，實作的部份使用 Concept Explorer[30]這個 FCA 分析軟體來進行 concept lattice 的轉換，此軟體提供了一個框架來建立 context table 及轉換 concept lattice 使資料形象化，

根據上述例子，會將 Brazilian Jujitsu 與 Martial art 透過轉換成 concept lattice 找出二概念的階層關係進行校正，如圖 11 所示，Brazilian Jujitsu 便會校正至 Martial art 下層概念，與 Judo 隸屬於同一層級。Concept lattice 中概念越大者代表含有越多的相關文件，在這個例子中，文件總數為 5 佔全部文件的 100%，其他概念表達方式以此類推。

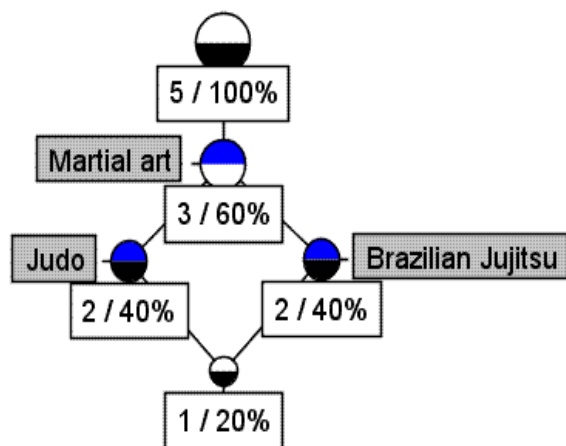


圖 11 轉換 concept lattice 圖例

(4) 產生模糊本體論

經過模糊概念分析後，每個概念都會與其相關網頁文件有著隸屬值的資訊，可以進一步的計算出概念 $C_2 \subseteq C_1$ 的附屬程度[21][22]，公式如下：

$$\text{subsethood}(\text{Subconcept}, \text{Superconcept}) = \frac{|\text{Subconcept} \cap \text{Superconcept}|}{|\text{Superconcept}|}$$

Subconcept：下層概念(子概念)

Superconcept：上層概念(父概念)

承上例，計算 Martial art 與 Judo 之間的 subsethood 附屬程度：

$$\text{subsethood}(\text{Judo}, \text{Martial art}) = \frac{0.6 + 0.7}{0.85 + 0.85 + 0.95} = 0.49$$

本體論轉換為模糊本體論後，會包含二種的隸屬值資訊，一種是概念本身與相關網頁文件的隸屬值，另一種為在概念與概念之間、概念與實例之間，除了彼此的關係外還包含著附屬程度。

4 實驗評估

在實驗的部份，針對棒球此一領域的本體論來進行整合，二個領域本體論的來源一個是從 DAML 本體論資料庫中取得的棒球領域本體論[29]，其中包含了 98 個概念；另一個是從我們之前的研究中[3][4]利用類神經網路相關技術自動建構的棒球領域本體論，其中包含了 74 個概念。

而其中由 DAML 本體論資料庫中取得的棒球領域本體論，由於是由專家手動建構的本體論，視為其正確性會較自動建構出的本體論為高，故將其設定為基底本體論，而自動建構的棒球領域本體論視為修正本體論。以下針對實驗的地方進行說明，主要分成網頁收集、獲

取 WordNet 資訊、模糊概念分析三個部分。

4.1 網頁收集

首先從 Google 搜尋引擎[31]以「棒球」為關鍵字，共收集了 3243 筆網頁文件，這些網頁文件是隨機選取的，所以其中必定含有相當多的雜訊可能會影響到整合後的正確性，透過預先處理階段提出的篩選網頁文件的方式，將前面 1600 筆網頁文件視為雜訊過高，只保留最後 1643 筆作為 FFCA 校正階段的網頁文件來源。

在實驗的部分，主要針對幾個項目進行探討，整合後的正確性、從獨立層級與整體層級來檢視篩選網頁文件方式的效益。而用來量測正確性的指標如下：

Precision

$$= \frac{\text{the number of mapping and correct mapping}}{\text{the number of mapping}}$$

Recall

$$= \frac{\text{the number of mapping and correct mapping}}{\text{the number of correct mapping need to be found}}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Precision：用來量測利用我們的方法找到的整合路徑中，正確的整合結果有多少。

Recall：用來量測全部應該被找到的正確的整合路徑中，正確的整合結果有多少。

F-measure：綜合以上二個值的平均量測指標。

4.2 初步實驗

首先在本體論整合階段中，經過概念合併、WordNet 校正的前二個階段處理，已經可將二個實驗本體論的整合工作幾近完成，只剩下一個上層概念 *baseball ground* 及其下層概念無法整合，而對於由上而下的整合方式來說，只要將 *baseball ground* 概念找到正確的位置，其下層概念便可自動化的進行整合。而前二階段的合併工作完全是依照 WordNet 資料庫中對字詞定義出的階層結構，均視為正確的整合位置。以下便針對修正本體論上的 *baseball ground* 概念進行 FFCA 校正，其校正結果如下圖 12。

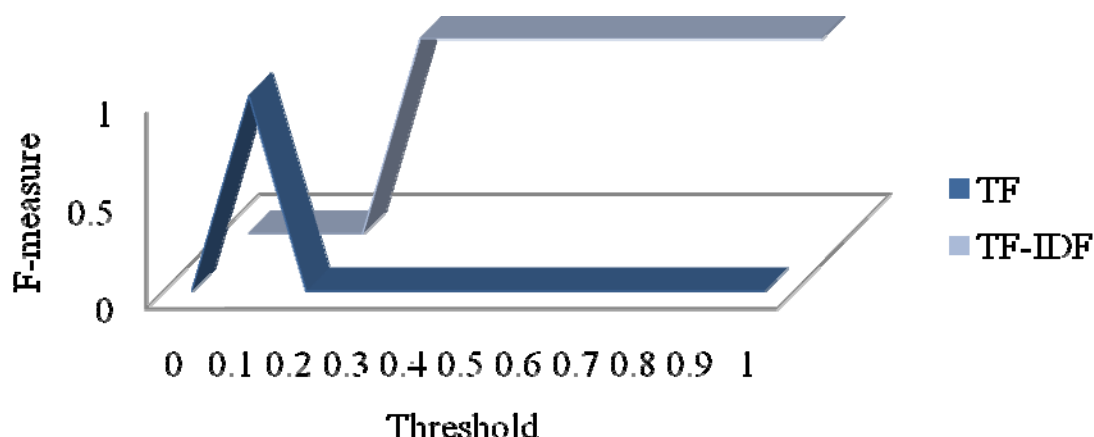


圖 12 FFCA 校正結果

由圖中可發現，隸屬值經由 TF-IDF 的設定，其正確率較 TF 高出許多，TF 值僅在 α -cut 為 0.1 時可以找到正確校正路徑，而 TF-IDF 在 α -cut 為 0.2 到 1 的設定皆能找到正確的校正路徑。

透過實驗結果可以證明 FFCA-merge 確實可以有效的達到領域本體論的合併目的，其中 WordNet 資料庫建構完備，已可以完成絕大部份的整合工作，而 FFCA 校正配合篩選網頁文件及 TF-IDF 值的隸屬值計算，同樣地可以

獲得不錯的效果。

4.3 FFCA 校正階段物件變數帶來的影響

首先針對文件數量來進行實驗，檢視增加網頁文件數量是否可以提升 FFCA 校正階段的正確性。在實驗的過程中，將總數 3243 的總網頁文件數目，以 2 的倍數逐漸加入，來進來量測，實驗結果如圖 13 所示。

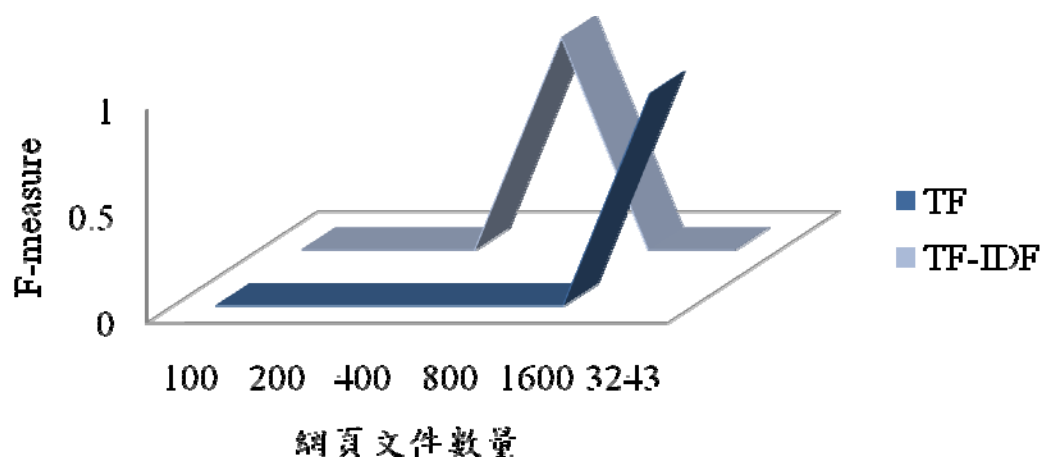


圖 13 網頁文件數量對 FFCA 校正的影響

由實驗中發現增加網頁文件數量不一定可以增加 FFCA 校正的正確性，由 TF-IDF 可以看出，在網頁文件數量越高時，其正確性反而又會下降，追究其主因乃是因為網頁文件是隨機的從網際網路上收集，並無經過分類或篩選的步驟，其中網頁文頁可能含有過多的雜訊，以致於影響到 FFCA 校正的正確性。

然而透過我們在預先處理提出的篩選網頁文件的方式，其正確性可獲得大幅度的改善(圖 12)，接下來實驗的部分根據獨立層級及整

體層級進行探討，在 FFCA 校正的部分主要採用的是獨立層級的策略來進行，主因乃是因為要避免掉網頁文件帶來太多的雜訊，因此只計算上層概念的相似性再轉換為 concept lattice，相對整體層級而言，尚需要花費轉換的成本，而整體層級如未經由篩選網頁，是完全無法找出正確整合路徑，但經由篩選網頁文件的前處理，可將整體層級策略的正確率相對提升不少，實驗結果如圖 14 所示。

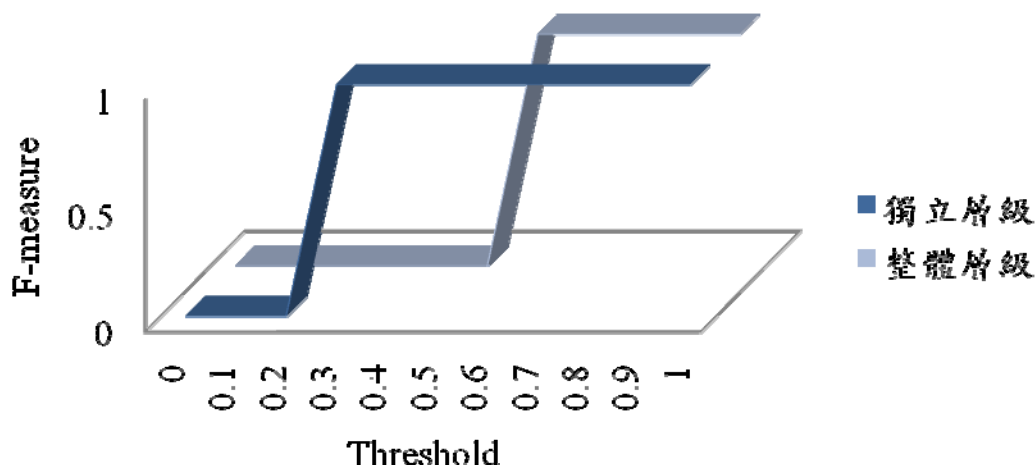


圖 14 獨立層級與整體層級比較

透過實驗可以發現，雖然獨立層級整體的正確率較高一些，但整體層級確實也可以達到一定的正確率，透過此實驗也可再次驗證篩選文件的方式確實有效。

5 結論

本研究主要透過模糊正規概念分析的技術及 WordNet 資訊的支援，將領域本體論進行有效率的整合，透過預先處理將詞幹變化、WordNet 資訊、篩選網頁文件，再進行概念合併、WordNet 校正、FFCA 校正，將本體論整合後，可以利用 FFCA 校正過程中和相關網頁文件的隸屬值資訊進而轉換成模糊本體論，而透過實驗也可以證明此合併架構確實有不錯的正確率。經由 FFCA-Merge 合併後，可以建構出具有模糊概念之本體論，在未來有足夠的能力面對現今資料庫中日益龐大的不明確資訊。

參考文獻

- [1] Birkoff, G., "Lattice Theory," *American Mathematical Society, Providence*, 1967.
- [2] Chalupsky, H. C., "Ontomorph: A Translation System for Symbolic Knowledge," *Principles of Knowledge Representation and Reasoning*, 2000.
- [3] Chen, R. C. and Chuang C. H., "Automating Construction of a Domain Ontology Using a Projective Adaptive Resonance Theory (PART) Neural Network and Bayesian Network. Expert Systems" *The Journal of Knowledge Engineering(SCI)*, 2008.
- [4] Chen, R. C., Chuang, C. H. and Tseng, C. C., "Constructing an Ontology Automatically by Projective ART Neural Network," *International Computer Symposium (ICS 2006)*, Dec. 4-6, 2006.
- [5] Choi, N., Song, I. Y. and Han, H., "A Survey on Ontology Mapping," *SIGMOD Record*, Vol. 35, No. 3, 2006.

- [6] Cole, R. and Stumme, G., "CEM—a Conceptual Email Manager," *Proceedings of the International Conference on Computational Science*, pp. 438-452, 2000.
- [7] Ganter, B. and Stumme, G., "Creation and Merging of Ontology Top-Levels," *Proceedings of the International Conference on Computational Science (ICCS '03)*, pp. 131-145, 2003.
- [8] Ganter, B. and Wille, R., "Formal Concept Analysis: Mathematical Foundations," *Springer-Verlag New York, Inc. Secaucus, NJ, USA*, 1997.
- [9] George, A. M., "WordNet: a lexical database for English," *Communications of the ACM*, Vol. 38, No. 11, pp. 39-41 1995.
- [10] Gruber, T., "Translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, pp. 199-220, 1993.
- [11] Ichise, R., Takeda, H. and Honiden, S. "Rule Induction for Concept Hierarchy Alignment," *Proceedings of the Workshop on Ontology Learning at the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.
- [12] Jones, K., *View Mail Users Manual*, <http://www.wonderworks.com/vm>, 1999.
- [13] Kalfoglou, Y. H. and Hu, B., "CROSI Mapping System (CMS) Results of the 2005 Ontology Alignment Contest," *K-CAP Integrating Ontologies Workshop*, 2005.
- [14] Kent, R.E. and Neuss, C., "Creating a Web Analysis and Visualization Environment," *Computer Networks and ISDN Systems*, Vol. 28, No. 1-2, pp. 109-117, 1995.
- [15] McGuinness, D., Fikes, R., Rice, J. and Wilder, S., "The Chimaera Ontology Environment," *In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, 2000.
- [16] Noy, N. F. and Musen, M. A., "Smart: Automated Support for Ontology Merging and Alignment," *Proceedings of the Twelfth Banff Workshop on Knowledge Acquisition, Modeling, and Management*, 1999.
- [17] Noy, N., and Musen, M., "Anchor-PROMPT: Using Non-Local Context for Semantic Matching," *Proceedings of the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.
- [18] Noy, N., and Musen, M., "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment," *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2000.
- [19] Petersen, W., "A Set-Theoretical Approach for the Induction of Inheritance Hierarchies," *Theoretical Computer Science*, Vol. 51, 2002.
- [20] Priss, U., "Linguistic Applications of Formal Concept Analysis," *Proceedings of the 1th International Conference. Formal Concept Analysis*, 2003.
- [21] Quan, T. T., Hui, S. C. and Fong, A.C.M., "Automatic fuzzy ontology generation for semantic help-desk support," *IEEE Transactions on Industrial Informatics*, Vol. 2, No. 3, pp 155 – 164, 2006.
- [22] Quan, T. T., Hui, S. C., Fong, A.C.M. and Cao, T. H., "Automatic Fuzzy Ontology Generation for Semantic Web," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 6, pp. 842-856, 2006.

- [23] Richards, D. and Compton, P., "Combining Formal Concept Analysis and Ripple Down Rules to Support Reuse," ***Proceedings of the 9th International Conference Software Eng. and Knowledge Eng. (SEKE '97)***, 1997.
- [24] Schmitz, C., Staab, S., Studer, R., Stumme, G. and Tane, J., "Accessing Distributed Learning Repositories through a Courseware Watchdog," ***Proceedings of the E-Learn 2002 World Conference E-Learning in Corporate, Government, Healthcare, and Higher Education***, pp. 909-915, 2002.
- [25] Sporleder, C., "A Galois Lattice Based Approach to Lexical Inheritance Hierarchy Learning," ***Proceedings of the International Conference ECAI 2002 Workshop Machine Learning and Natural Language Processing for Ontology Eng. (OLT '02)***, 2002.
- [26] Stumme, G. and Maedche, A., "FCA-Merge: Bottom-Up Merging of Ontologies," ***In proceeding of the International Joint Conference on Artificial Intelligence IJCAI101***, 2001.
- [27] Wille, R., "Restructuring lattice theory: an approach based on hierarchies of concepts," ***I.Rival (Ed.), Symposium on Ordered Sets, Reidel***, 1982.
- [28] Cycorp, <http://www.cyc.com/>
- [29] DAML Ontology Library, <http://www.daml.org/2001/08/baseall/baseall-ont.daml>
- [30] Formal Concept Analysis, <http://www.upriss.org.uk/fca/fca.html>
- [31] Google, <http://www.google.com>
- [32] Suggested Upper Merged Ontology, <http://www.ontologyportal.org>