可提供語意搜尋之部分知識建構

王宗一　謝東成　陳慶龍　林群貿
國立成功大學 工程科學系
wti535@mail.ncku.edu.tw

摘要

近十年網路的發明帶動著資訊大量的流通。由於許多知識已經數位化，故造成資訊氾濫相關問題。因此，如何正確、有效率的重複利用知識是目前重要的研究議題。目前最常被用來尋找知識的工具為 Google 或 Yahoo 搜尋引擎，所依賴是輸入關鍵字利用資訊檢索的技術獲取網頁相關資訊，再由使用者自行閱讀並組織知識。如果搜尋結果非常龐大，使用者必須花費很多時間去瀏覽內容，以找尋所需的資訊。有鑑於此，本研究將提出一套自動化建構部份知識本體的方法以提供問答系統的語意搜尋資料來源。其方法是將問句本身的語意擴展，再從體育新聞內容中自動建構所屬的部份知識本體，作為回覆答案的資料集。經過實測所得數據顯示本研究方法有顯著的效果。

關鍵詞：資訊檢索、知識本體、問答系統

一、前言

目前知識本體建構必須依賴專家的參與，如果可以自動化建構知識本體，將可節省知識管理的人力，所建置的知識本體也會因為減少人的介入而降低不必要的錯誤。自動化建構的另一項優點就是知識本體的結構較為彈性。人工建置類別架構可能是依據當時的資料內容與屬性才能定義，然而隨著時間的變化資訊內容可能也跟著轉變，如果沿用先前制定的知識本體架構，知識的表達可能就不符合實際狀況而產生不符需求或錯誤。如果知識本體可以隨著時間與資料內容的變化自動調整，資料內容可以更正確的對應到合適的知識類別，在問題的推論上也會更準確。對於問答系統而言可以幫助使用者獲取知識，為了達到此目的則必須結合人工智慧、自然語言處理與資料檢索的技術，並正確的辨識問題類型以回覆適合的答案，但目前問答系統的效能還有很大的進步空間。

因此本研究利用 Formal Concept Analysis, FCA 方法建構部份知識本體，將其整合在問答系統內，利用其概念的關聯性找出正確的資訊給使用者，使問答系統的效能提昇。本研究以網路電子媒體報導體育新聞為題材，如 NBA 籃球新聞。利用中文斷詞系統切割新聞詞句，處理後的資料經過分析過濾，將有意義的詞彙儲存，提供建構知識本體的素材。使用者對問答系統輸入問題，系統分析問句並將問題分類並擴展產生物件集與屬性集，利用 FCA 分析所形成的概念，初步建立階層關係，再依據新聞內容標定關係的描述，由此形成部份體育知識本體。問句的類型也經過剖析、分類，再依據其問題類型與所建構的部份體育知識本體回答使用者問題的答案。

本研究後續架構如下：第二節說明相關文獻探討。第三節說明本研究所提出之系統架構，分別包含三個子系統即文件前處理子系統 (Documents Preprocessing Subsystem)、詢問式知識本體建構子系統 (Query-based Ontology Construction Subsystem)以及語意問答子系統(Semantic Aware Question Answering Subsystem)。第四節為本研究實驗結果部分。最後，第五節則是本研究的結論部分。

二、相關文獻探討

(一) 知識本體(Ontology)

知識本體源自哲學理論，其意義是有系統的解釋存在的現象。主要探討存在現象的一切現實事物的基本特徵。知識本體定義了一個主題領域的構成詞彙，包含詞彙間基本條件與關係，以及延伸詞彙所定義的基本條件與關係的法則，是一種正規化的(Formal)、明確的(Explicit)、概念化的(Conceptualization)、可分享的(Share)描述的邏輯理論[6] [7] 。

知識本體敘述是以名詞所代表真實存在的實體或概念，而概念是知識的最小基本單位或元素。這種具有抽象與具象的元素，很適合表達資訊領域的知識，資訊技術相關領域逐漸藉由知識本體的基本元素：實體、概念及概念間的關係，作為描述真實世界的知識模型[8] 。針對此一趨勢，W3C 組織制定許多知識本體的相關表示語言，例如 RDF、DAML+OIL、OWL…等[4] 。透過這些語言可將知識本體當成語意網的骨架，而知識本體提供人與系統之間可分享的、可理解的與重複使用的溝通平台[10] 。

描述知識本體的重要元素包括有：概念(Class 或 Concept)、屬性(Attribute 或 Property)、實例(Instance)、關係(Relation)。描述知識本體的範例，如圖 1 所示。範例所描述的概念 Car 下有包含有兩個子概念(Subclass) 為 Two-wheel drive 以及 Two-wheel，如果 Car 這個概念具有屬性：引擎、方向盤、輪子…等，其子概念將繼承父概念的所有屬性，也就是說子概念 Two-wheel drive 同樣也有概念 Car 的屬性：引擎、方向盤、輪子…等。概念 Two-wheel drive 的實例為 Altis，實例將繼承父概念所有的屬性，也可能具備與其他實例不同的獨特屬性，以表現與其它實例的差異性，例如：實例 Altis 與實例 Benz ML63AMG 皆有引擎、方向盤、輪子…等屬性，但在內裝配備的屬性值上會大不相同。概念與概念之間或概念與實例之間使用關係描述來表現彼此的關聯性。例如實例 Sailboat 與概念 Ship 存有關係 instance_of，說明 Sailboat 是 Ship 的一個實例。

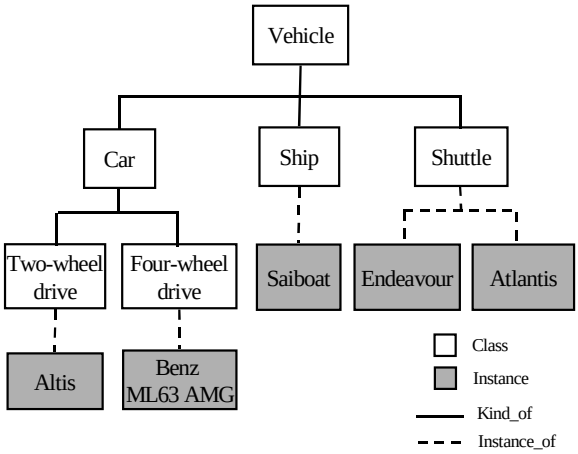

圖1 知識本體概念描述範例

(二) 正規概念分析 (Formal Concept Analysis, FCA)

正規概念分析 (Formal Concept Analysis, FCA)是一種理論方法，用於分析辨識一組資料集合的概念結構[3] 。1982 年 Rudolf Wille 介紹推廣後，相關的研究已迅速增加。FCA 的分析方法成功的應用在各領域之中，例如：醫學、心理學、語言資料庫…等。其原因在於它具有可以產生視覺化內部結構圖表的能力。特別是用於社會科學領域，處理不能被完全定量分析的資料集合之研究。在資訊科學領域，FCA 使用的數學方格(Lattice)可應用在分類系統上。依據他們的類別屬性建構階層。以下說明正規概念分析之定義。

一個正規化的本文具備三個部份表示為 (G,M,I)。G 與 M 分別為集合，$I \subseteq G \times M$ 則表示 I 為 G 與 M 所形成二元關係的集合。G 集合內的元素稱為物件(Objects)，M 集合內的元素稱為屬性(Attributes)。一般表示 $(g,m) \in I$ 可以寫成 gIm，表示為 Object g 存有 Attribute m 其二元關係屬於集合 I 之一[3] 。對所有 $A \in G$ 且 $B \in M$ 時且 A 與 B 為有限集合時，定義如下：

$$A' = \{m \in M \mid (\forall g \in A) g \text{ Im}\}$$

$$B' = \{g \in G \mid (\forall m \in B) g \text{ Im}\}$$

A' 集合是所有物件 A 集合內所有的屬性集合。B' 集合是所有包含屬性 B 集合的物件集合。而本文 (G, M, I) 內的某一概念(Concept)定義為$(A,\ B)$，其中 $A \subseteq G$ 、$B \subseteq M$ 、$A' \subseteq B$ 、$B' \subseteq A$。(A, B)的 A 稱為範圍(Extent)，B 稱為含義(Intent)。

除此之外，如果(A, B)是唯一的概念，(A, A') 的範圍為 A，若且唯若 $A'' = A$ 且 $B'' = B$。G 的概念子集合 (A, A') 是擁有集合 A 所有範圍的一個獨特的概念，概念 (B, B') 亦同。如果本文內所有的概念集合存在 (A_1, B_1) 與 (A_2, B_2) 這兩概念，若 $A_1 \subseteq A_2$ 則 $(A_1, B_1) \leq (A_2, B_2)$。近一步的說，因為 $A_1'' = A_1$ 且 $A_2'' = A_2$，所以當 $A_1 \subseteq A_2$ 時意味著 $A_1' \supseteq A_2'$ 也就是 $B_1 \supseteq B_2$。

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \subseteq B_2$$

三、系統架構

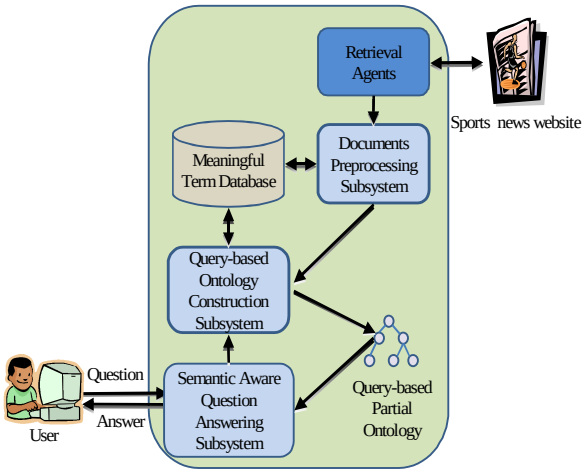

圖2 系統架構圖

圖 2 為本研究所提出的系統架構，分成下列六個部分：

首先，新聞擷取器(Retrieval Agents)將所擷取的新聞資料內容，儲存於系統內，作為系統執行的資料來源。文件前處理子系統(Documents Preprocessing Subsystem)負責處理原始新聞內容，包括有兩個主要功能模組分別是詞彙分析模組(Term Analysis Module)以及內文分析模組(Context Analysis Module)。詞彙分析模組切割原始新聞語句並標記詞性，濾出欲分析的詞彙並將結果存放至有意義詞彙資料庫(Meaningful Term Database)；內文分析模組則是分析新聞內容段落，並且輸出詞彙同現矩陣(Co-occurrence Matrix)。有意義詞彙資料庫紀錄文件前處理子系統產生的有意義詞彙資料集，提供系統執行時可以依照詞彙同現矩陣的紀錄，快速且有效率取用有意義詞彙。

詢問式知識本體建構子系統(Query-based Ontology Construction Subsystem)是依據使用者輸入的問句，利用正規概念分析方法建構詢問式之部份知識本體(Query-based Partial Ontology)，此知識本體初步建構有階層性的概念結構，再參照原始文章的語句描述，標記關係的名稱以提供系統作為回答問題的資料集來源。

語意問答子系統(Semantic Aware Question Answering Subsystem)可分為三個部份即分析問句、回覆問題答案與系統評估。問題分析模組(Question Analysis Module)的目的是將使用者所詢問的問題進行切割、過濾出有意義的字詞並標記問題類型。答案分析模組(Answer Analysis Module)的目的為分析問題的類型，取得詢問式之部份知識本體的資料集做為回答問題的資料來源。最後，後端評估模組(Evaluation Module)將相關的問答資訊皆儲存於資料庫提供專家(Domain Experts)檢視與驗證之用。

本研究主軸分別架構於三個子系統之中，分別是：文件前處理子系統、詢問式知識本體建構子系統以及語意問答子系統，以下詳細說明每一個子系統的處理流程。

(一) 文件前處理子系統

此子系統包括有詞彙分析模組與內文分析模組，首先詞彙分析模組是將新聞擷

取器(Retrieval Agents)所取得的原始新聞資料，利用中央研究院所開發的中文斷詞系統 CKIP [12] 將新聞詞句切割為有意義的詞彙集並標注該詞彙之詞性。再利用詞彙過濾器(Term Filter)篩出具有意義詞彙，如下表 1 為本研究欲保留之具有意義的詞彙集合。其主要目的在去除無意義的虛詞，如：介詞、副詞、連接詞、助詞...，並將結果儲存於有意義詞彙資料庫(Meaningful Term Database)。

表 1 有意義詞組範例表

詞性標注	詞性描述
Na	普通名詞 (Common noun)
Nb	專有名詞 (Proper noun)
Nc	地方名詞 (Place noun)
Nd	時間名詞 (Time noun)
VA	動作不及物動詞 (Active intransitive verb)
VC	動作單賓動詞 (Active transitive verb)
VD	雙賓動詞 (Ditransitive verb)
VHC	狀態使動動詞 (Stative unaccusative verb)
VH	狀態不及物動詞 (Stative intransitive verb)
VJ	狀態單賓動詞 (Stative transitive verb)

經過詞彙分析模組初步處理新聞資料後，內文分析模組將進行新聞資料內文分析。根據前模組處理後可以產生兩個集合，分別為：新聞文件集合 $D_i = \{D_1, D_2, D_3, ..., D_n\}$ 與詞彙過濾器所得有意義詞彙集合 $T_j = \{T_1, T_2, T_3, ..., T_m\}$。

$$X_{ij} = \begin{cases} 1, & D_i \text{ co - occurs with } T_j, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$M_{ij} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix} \quad (2)$$

利用公式(1)產生新聞文件與有意義詞彙的同現陣列。其表示方式為公式(2)之 $n \times m$ 矩陣。其中 n 表示文件的篇數，m 為有意義詞彙的個數。此步驟所建立之同現矩陣資訊將作為詢問式知識本體建構子系統的二元關係矩陣產生器所取用。

(二) 語意問答子系統

語意問答系統處理使用者所輸入的問句並依其問題的類型回覆答案。使用者輸入的詢問句是由問題分析模組(Question Analysis Module)處理，答案分析模組(Answer Analysis Module)則是將答案結果回覆給使用者。

表 2 問句樣式範例表

樣式類型	範例
WHEATHER	熱火隊教練是萊里嗎? (Nb + DE + Na + SHI + Nb + T ?)
WHERE	騎士隊主場城市在哪裡? (Nb + DE + Na + Na + P + Ncd ?)
WHICH_ONE	諾威茲基是哪隊的前鋒? (Nb + SHI + Nep + Nf + DE + Na ?)
WHO	湖人隊的後衛是誰? (Nh + SHI + Nb + DE + Na?)
WHOSE	請問歐尼爾暱稱是什麼? (VE + Nb + DE + Na + SHI + Nep ?)

首先問題分析模組部份，使用者的問句先經由 CKIP 切割並標記詞性，再經由詞彙過濾器篩出有意義的詞彙，藉由此步驟系統能夠擷取出具有意義的辭彙傳送至詢問式知識本體建構子系統中進行建構。其後，則會標注問句焦點(Question Focus)

標籤與問句焦點描述(Question Focus Description)標籤[9] ，如標注<QF> Term A </QF>或<QFD> Term B </QFD>的標籤，其目的在了解使用者所詢問的問題焦點，例如，使用者詢問 “誰是熱火隊的教練?”，經由標注的結果為誰是<QFD>熱火隊</QFD>的<QF>教練</QF>?

系統對問句類型的分析結果將是選取答案的重要依據。對問句類型的辨識與解析係依據事先收集的多種問句型式與內容分析得來，經過所收集問句的訓練，歸納整理出問句基本的樣式如表 2 所示。當使用者輸入問句經過分析，如符合系統預設的樣式類型，該問題將被標記為所符合的問句類型。系統依符合的樣式類型標記<QT VALUE=”XXXX” />的標籤，標記後的問句範例如表 3 所示。

表 3 問句類型標記

問句範例	註記標籤的問句
請問小牛隊的教練是誰?	請問<QFD>小牛隊</QFD>的 <QF> 教練 </QF> 是誰?<QT VALUE="WHO" />
請問火箭隊的主場在哪?	請問<QFD>火箭隊</QFD>的 <QF> 主場 </QF> 在哪?<QT VALUE="WHERE" />
請問姚明是那一隊?	請問<QF>姚明</QF>是那一隊 ?<QT VALUE="WHICH_ONE" />

答案分析模組依據註記標籤決定問題答案的範圍，並且從建置好的部份知識本體之中搜尋適當的答案提供給使用者。圖 3 則是根據上面所提到的問題 “誰是熱火隊的教練?”進行建構的部份體育知識本體。此架構已具有階層結構與關係的描述，可做為回答問題的依據。

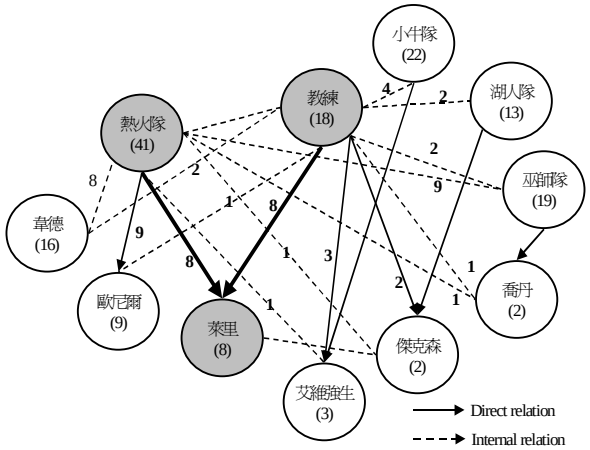

圖3 部分體育知識本體之範例

由於詢問的問題類型不同，有時候產生的部份體育知識本體之答案候選可能為多個或無法直接從階層架構找出答案時，必須挑選最合適的答案給予使用者。因此，本研究在此部分將依據詞彙間共同出現頻率，使用公式(3)計算資料集內與問題有關的詞彙分數。其中 *w* 為一權重值，當概念之間具有 Direct relation 時作為增加權重的指標，其值為 0.7，相反則為 0.3。最後，系統則挑選總分數最高者為答案提供給使用者。

以上面的問題為例，本研究必須計算出總分數為最高者作為提供使用者答案。因此，熱火隊出現的句子共有 41 句；教練出現的句子共有 18 句；熱火隊與萊里共同出現的句子共有 8 句；教練與萊里共同出現的句子有 8 句。因兩兩概念之間具有 Direct relation 故 *w* 為 0.7，再利用公式(3)可計算出 *Score*(熱火隊, 萊里)=0.137 與 *Score*(教練, 萊里)=0.311 為最高值，因此系統將挑選萊里作為答案提供給使用者。

$$Score = \frac{Term_x 與 Term_y 共同出現的句子個數}{Term_x 出現的句子個數} \times w \quad (3)$$

(三) 詢問式知識本體建構子系統

此系統主要目的是依據使用者所詢問的問題建構一詢問式之部分知識本體(Query-based Partial Ontology)。因此依據

語意問答系統的問題分析模組所取得的有意義片段描述和文件前處理子系統的同現矩陣與有意義詞彙資料庫的資料集，送到此子系統中進行建構部份知識本體的步驟。

此子系統使用 Formal Concept Analysis 方法，藉由概念所包含的物件集與屬性集來決定概念之間隱含的關係或屬性的階層[2] [5] [11] 。其建構步驟如下：

Step1: 產生二元關係矩陣與概念集

以符號(O, A, X)表示有物件集合O與屬性集合 A 存在二元關係集合X，換言之即$X \subseteq O \times A$。物件集合 O 的產生是由問題分析模組所找出的問題焦點與問題焦點描述的重要詞彙，從該重要詞彙的所屬的文件集合中所有的描述語句的集合即為物件集合 O。而 A 為該文件描述語句集合中所有屬於有意義的詞彙的集合。

要產生一個概念集合 C，假設 S 為物件 O 的部份集合，Q 為屬性 A 的部份集合，則$S \subseteq O$ 且$Q \subseteq A$。該物件 S 集合的所有屬性值為：

$$\sigma(S) = \{a \in A \mid \forall o \in S : (a, o) \in X\]$$

而其包含Q屬性集合的所有物件集合為：

$$\tau(Q) = \{o \in O \mid \forall a \in Q : (a, o) \in X\]$$

而概念則是一組(S, Q)所組成。換言之即一組有意義的詞彙(屬性)與一組擴展衍生的文章描述語句(物件)其關係為$Q = \sigma(S)$且$S = \tau(Q)$[11] [1] 。此步驟是取得文件前處理子系統產生之同現矩陣有關物件的資料與有意義詞彙資料庫內的資料集，再依據其矩陣的描述比對資料庫內的有意義矩陣的關聯性，產生具有物件集合、屬性集合與物件與屬性二元關係的資料結構。並且依照該資料結構生成一組概念集合。

Step2: 建立所有的概念的階層關係

在產生概念集合後，概念之間的關係就可以被建立。假設某概念(A_0, B_0)為概念(A_1, B_1)的子概念(Sub-concept)之一，可表示為$(A_0, B_0) \subseteq (A_1, B_1)$。換句話說$c_0 = (A_0, B_0)$為$c_1 = (A_1, B_1)$的子概念[5] 。

在概念階層上給定兩個元素(I_1, J_1)與(I_2, J_2)，他們的下確界(Infimum)定義為$(I_1, J_1) \cap (I_2, J_2) = (I_1 \cap I_2, \sigma(I_1 \cap I_2))$

而上確界 (Supremum) 定義為$(I_1, J_1) \cup (I_2, J_2) = (\tau(J_1 \cap J_2), J_1 \cap J_2)$

所有的概念階層關係的建構皆由上確界與下確界來決定[1] 。此步驟決定概念間的階層關係，利用上述定義決定每個概念是否具有父概念或子概念。概念間產生關聯性後，將提供系統尋找或判斷答案之用。另一方面也會搜尋文章內容的描述，標記該階層關係適合的屬性值。

Step3: 建立概念間之內部關係

概念之間除了繼承的關係外，可能也存有部份交集的關係。部份交集可以說明兩概念間存有某種程度的相似性。藉此步驟彌補 FCA 方法無法呈現概念間部份交集，藉由此步驟找出概念間內部的關係[11] 。假設有兩個概念$c_0 = (A_0, B_0)$與$c_1 = (A_1, B_1)$，使$\exists Set_x$且$\exists Set_y$，當$Set_x \subset B_0$、$Set_y \subset B_1$且$Set_x = Set_y$，則概念c_0與c_1之間存有內部的關係。系統會依據上述原則決定概念間是否存在共同的屬性集合，存有該屬性集合表示存有內部關係，並且逐一搜尋新聞內容找尋適合標注內部關係的屬性值。

經過上述建構步驟與流程，此子系統將建構出詢問式之部份知識本體(Query-based Partial Ontology)，提供語意問答系統的答案分析模組做為回覆問題的答案來源。

四、實驗結果

本研究是以 FCA 方法建構詢問式之部份知識本體 (Query-based Partial Ontology)以提供語意搜尋使用。故實作一套問答系統並以體育新聞文件作為實驗素材，其使用者操作介面如圖 4 所示。並且進行兩種類型的實驗測試，其目的在驗證

本系統的準確率與實用性。分別是實驗一：比較本研究所實作的問答系統與關鍵字搜尋引擎所得答案之準確率。實驗二：使用本研究方法，將素材改為其他類型的體育新聞並測試其答題準確率。

圖 4 使用者操作介面

(一) 實驗一

首先，新聞擷取器從聯合新聞網[13]中擷取美國職業籃球相關新聞總共 180 篇作為實驗素材，其資料如表 4 所示。表 5 為辭彙分析模組處理 180 篇新聞文件後之詞彙結果，其中相異詞彙數量表示為該詞性不同詞彙的各數。

表 4 文件與詞彙數量表

	文件數量	句子數量	詞彙過濾前之數量	詞彙過濾後之數量
Total	180	1,728	66,296	25,493

表 5 中文自然語言處理分析之詞類統計

詞性	詞性標注	詞彙數量	相異詞彙數量
名詞	Na	8,127	2,286
	Nb	5,084	331
	Nc	766	144
	Nd	1,271	93
動詞	VA	1,492	350
	VC	3,679	775
	VHC	208	45
	VH	3,376	837
	VJ	1,204	188
	VD	286	37
總數		25,493	5,086

實驗環境設定

本測試之問題是從網路使用者對本研究感興趣的問題與網路論壇所收集的常見問題集來做為實驗材料，共 5 種題型、210 個問句，題型與問句統計如表 6 所示。

表 6 問題集的題型、說明與例句

問題類型	例句	數量
WHERE	湖人隊的主場是哪一個城市?	36
WHETHER	歐尼爾是中鋒嗎?	11
WHOSE	賈奈特的外號是什麼?	7
WHICH_ONE	詹姆斯是哪個球隊的呢?	13
WHO	姚明的教練是誰?	143
合計		210

實驗評估方法

實驗方法是將所收集的問句分別輸入至系統處理，個別將其處理的結果紀錄下來，交由該領域專家評估判斷其答案的正確性，再統計回答問題的準確率，所使用的評估準確率的公式(3)如下所示。評估本實驗之答題準確性由熟悉該體育知識背景的專家擔任。

$$\text{準確率(Precision)} = \frac{\text{正確回答問題的總題數}}{\text{輸入實驗問題的總題數}} \quad (3)$$

為了驗證本系統的準確率以及實用性，因此比較本研究所實作的問答系統和關鍵字搜尋引擎所的得答案之準確率。其搜尋引擎使用目前最深受歡迎的關鍵字搜

尋引擎 Google 擁有 80 億個網址索引[14]，將所收集的問題集分別輸入本研究之問答系統與 Google 搜尋引擎，本實驗設定 Google 搜尋模式為一般全域搜尋模式，全域搜尋模式可以搜尋世界所有網站的文件資料，指定搜尋字體為繁體中文，輸入的例句如：小牛隊的教練是誰?，範例其結果如圖 5 所示。由於 Google 搜尋引擎每次提供搜尋結果的片段資料為 10 筆，因此搜尋結果的 10 筆資料片段如果包含問題答案任一項，即視為正確回答問題，如果答案不出現在該次 10 筆片段資料內，則視為錯誤的答案內容。檢視 Google 資料片段正確性之專家評估介面如圖 6 所示。藉此評估相同問句於兩種不同系統之準確率。

圖 5 Google 檢視回傳之 10 筆 Snippet 內容

圖 6 使用 Google 查詢問句後的資料片段

實驗結果

表 7 則為本研究所提出方法之實驗結果，其準確率為 65.24%。表 8 則為使用 Google 查詢後的 Snippet 資料，驗證其準確率為 70%。

表 7 使用詢問式之部分知識本體之結果

題型	正確性	題數	準確率 %
WHERE	Correct	31	86.11%
	Wrong	5	13.89%
WHOSE	Correct	4	57.14%
	Wrong	3	42.86%
WHETHER	Correct	10	90.91%
	Wrong	1	9.09%
WHICH_ONE	Correct	11	84.62%
	Wrong	2	15.38%
WHO	Correct	81	56.64%
	Wrong	62	43.36%
Total	Correct	137	65.24%
	Wrong	73	34.76%

表 8 使用 Google 查詢問句之結果

題型	正確性	題數	準確率 %
WHERE	Correct	27	75%
	Wrong	9	25%
WHOSE	Correct	1	14.28%
	Wrong	6	85.71%
WHETHER	Correct	9	81.82%
	Wrong	2	18.18%
WHICH_ONE	Correct	8	61.54%
	Wrong	5	38.46%
WHO	Correct	102	71.33%
	Wrong	41	28.67%
Total	Correct	142	70%
	Wrong	68	30%

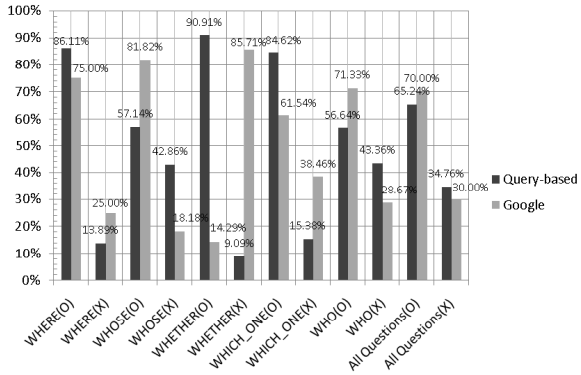

圖 7 使用詢問式之部份知識本體與 Google 搜尋引擎之準確率比較

由實驗結果可以觀察到本研究之問答系統的實驗準確率 65.24%與 Google Search 使用 Key Word 尋找問題答案的準確率 70%約略相近，如圖 7 所示。不過在本實驗設定下，當使用者利用 Google 找尋答案是必須檢閱 10 筆網頁部份資料，再確認是否有答案在其中，此實驗 210 個問句的搜尋結果總計有 2,100 筆資料、共 228,804 個字，對於使用者來說瀏覽大量資料是很沈重的負擔。在準確率相近的情況下，使用本研究之問答系統直接提供一個或一段答案的描述是比較有效率。

(二) 實驗二

圖 8 實驗二之專家檢視畫面

實驗二則是針對不同新聞題材進行比較，其文章內容取自中華職棒大聯盟(Chinese Professional Baseball League, CPBL)新聞共 50 篇進行實驗。其實驗步驟如同實驗一，開放使用者輸入問題測試，系統將紀錄所有可處理之問題與系統答覆之答案，提供專家檢視驗證如圖 8 所示。

實驗結果

系統收集有效問句 116 個，經由專家檢視驗證其資料之正確性，其表 9 為實驗之準確率結果。

表 9 實驗二之查詢問句結果

題型	正確性	題數	準確率 %
WHERE	Correct	3	50
	Wrong	3	50
WHOSE	Correct	4	44.44
	Wrong	5	55.56
WHETHER	Correct	14	72.97
	Wrong	2	27.03
WHICH_ONE	Correct	5	45.45
	Wrong	6	54.55
WHO	Correct	54	68.97
	Wrong	20	31.03
Total	Correct	80	68.97
	Wrong	36	31.03

從此實驗中可發現，由於所收錄的文件數量較少，因此是非問句題型的答題準確率較高，反應 FCA 方法使用於小範圍文件時，其建構屬性之間的階層關係較明顯，對於是非問句題型的分析判斷較有利。球員相關資訊題型的答題準確率較低的原因可能是文件量未達規模，導致此部份判斷的資訊不足。整體上，系統答題準確率可達六成以上，可說明本研究對於問答系統上之實用性。

五、結論

本研究運用正規概念分析方法將無結構性的文字敘述轉成可被使用的知識，並

且將文章內容自動的建構成具組織性、階層性的資料集，使資料可被使用。並且實作一問答系統，此系統內包含文件前處理子系統，負責處理原始資料使之成為可被使用的資料集。詢問式知識本體建構子系統則建構知識本體，提供語意問答子系統回答問題的資料集。語意問答子系統則對問題進行剖析、問題焦點與描述的萃取、辨識問題的類型，最後提供使用者答案。本研究方法相對於搜尋引擎所提供大量、冗餘而無結構性的資料而言，較能夠節省使用者彙整資料的時間，初步達到智慧型網路的願景，並經過實際測試，其回答問題之準確率驗證具有可用性。

六、致謝

本研究承蒙國科會計畫 NSC95-2221-E-006-158-MY3 經費部分補助，特此感謝。

七、參考文獻

[1] F. Buchli, “Detecting Software Patterns using Formal Concept Analysis,” in der Philosophisch-naturwissenschaftlichen Fakult¨: University of Bern, 2003.

[2] P. du Boucher-Ryan, and D. Bridge, “Collaborative Recommending using Formal Concept Analysis,” Knowledge-Based Systems, vol. 19, pp. 309-315, 2006.

[3] B. A. Davey, and H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, pp. 65-84, 2002.

[4] D. Fensel, F. van Harmelen, I. Horrocks, D. L. McGuinness, and P. F. Patel-Schneider, “OIL: an ontology infrastructure for the Semantic Web,” Intelligent Systems, vol. 16, pp. 38-45, 2001.

[5] A. Formica, “Ontology-based concept similarity in Formal Concept Analysis,” Information Sciences, vol. 176, pp. 2624-2641, 2006.

[6] T. R. Gruber, “A Translation approach to portable ontology specifications,” Knowledge Acquisition, vol. 5, pp. 199-220, 1993.

[7] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. R. Swartout, “Enabling technology for knowledge sharing,” Ai Magazine, vol. 12, pp. 36-56, 1991.

[8] N. F. Noy, and D. L. McGuinness, “Ontology development 101: A guide to creating your first ontology,” Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[9] C. W. Shih, C. W. Lee, M.-Y. Day, T. H. Tsai, T. J. Jiang, C. W. Wu, C. L. Sung, Y. R. Chen, S. H. Wu, and W. L. Hsu, “ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA,” Proc. of NTCIR-5 Workshop, Tokyo, Japan, pp. 202-208, 2005.

[10] B. Swartout, R. Patil, K. Knight, T. Russ, K. Knight, and Tom Russ, “Toward Distributed Use of Large-Scale Ontologies,” In Symposium on Ontological Engineering of AAAI Stanford, California, 1997.

[11] S. S. Weng, H. J. Tsai, S. C. Liu, and C. H. Hsu, “Ontology construction for information classification,” Expert Systems with Applications, vol. 31, pp. 1-12, 2006.

[12] Chinese Knowledge Information Processing Group, “CKIP AutoTag,” Academic sinica, 1998.

[13] http://udn.com/NEWS/main.html.

[14] http://www.google.com/intl/zh-TW/why_use.html.