導入知識本體與學習能力於FAQ系統

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

戴紹國
朝陽科技大學資訊管理系助理教授
sgdai@cyut.edu.tw

許智皓
朝陽科技大學資訊管理所研究生
s9514645@cyut.edu.tw

黃盛麟
朝陽科技大學資訊管理所研究生
s9614645@cyut.edu.tw

摘要

近年來在語意網技術的發展下，本體論的應用已經越趨多樣化，尤其以資料檢索方面的應用為主，主要訴求為利用本體論，來針對使用者的需求提供更正確的資訊。但目前大多數的應用，主要是以被動式的資料檢索，缺乏與使用者的互動。作為使用者與機器之間溝通的管道，本體論應該要能隨著資訊的快速變動而隨時更新，所以能隨著新資訊的加入而擴充便成為本體論的一個課題。因此我們提出一個具有學習能力，並且可以與使用者互動的自動化線上FAQ(Frequently Asked Questions)系統。除了可以讓使用者能快速得到回覆，更能依據新得到的資訊，對本體論加以擴充以應付更多的需求。

關鍵字：本體論、FAQ、知識本體、TF-IDF

Abstract

In recent years, the application of ontology has been already toward the diversification under the development of the semantic web technology. The main application of ontology is information retrieval. With the utilization of ontology, we expect to offer more correct information for users. Although, the most of application of ontology is information retrieval but it lacks of the interaction with users. As the communicated bridge of the users and the machines, ontology should be upgraded while the information changes. Thus it is an important subject to expand ontology with the changing information. So we propose an automation on-line FAQ (Frequently Asked Questions) system that has the learning ability. We except that the system can reply answers and extend ontology.

Keywords：Ontology、FAQ、TF-IDF

1.緒論

1.1 研究背景與動機

近年來網際網路的快速發展，使得大多數的企業紛紛將自己公司 E 化，並建構自己所屬的公司網頁。其中因應而生的線上客服系統，亦逐漸成為消費者與企業間的溝通橋樑。有鑒於以往使用者在詢問問題時，常發生幾種可能情況，如：問題的重複詢問、客服人員回覆速度過慢等情況。使得使用者無法在第一時間得到想要的答覆。再以企業觀點來看，客服人員亦是人力成本之一。因此本篇提出以本體論建構出一個線上客服系統，利用本體論來找尋使用者所需要的解答，並於資料庫查不到所需要資料時，在負責單位回覆問題給使用者同時透過本系統將概念新增至資料庫與本體論架構

中，達成具有學習能力且可以自動回覆的線上 FAQ 系統。

1.2 問題描述

我們以朝陽科技大學校務 FAQ 系統為例，使用者提出問題後並選擇處理單位後，系統將問題轉發該單位，由該單位處理或者於無法處理後再轉發其他可以處理之單位。由最後處理問題之單位回覆於 FAQ 介面上，給予使用者觀看。原 FAQ 系統於資料量過多時予以刪除舊資料。圖 1 為朝陽 FAQ 介面圖。

原 FAQ 系統之缺點有以下缺點：

（1）、使用者可能提出跟先前使用者提出類似之問題，但由於並無將舊資料保存下來，造成行政單位重複處理同一問題。

（2）、系統處理問題的速度會因為行政單位的處理速度而可能造成回覆過慢。

使用者提出問題時，將可能會發生一開始所選擇的單位無法處理問題之情況。以下為可能發生的情況：

（1）、使用者所選擇的第一個處理單位將此問題轉發至可以處理的單位。

（2）、第一個處理單位除了轉發問題外，並依照對於該問題的認知回覆了可能的處理方法。因此最後使用者於系統介面上可能看到不只一個單位之回覆，如圖 2 及圖 3。

（3）、使用者所提出之問題可能有多個單位可以處理。例如獎學金的問題，在行政單位以及學術單位都有不同的處理方式。

編號	標題	發表日期	回覆狀況
1	0186 聽流行音樂學英文 薛珍華	2007/8/8 下午 12:57:38 --by 9311044, 實習生	○教務處課務組
2	語言選修多樣化!!!!	2007/8/8 下午 12:41:04 --by 9321053, 侯念辰	○教務處課務組 ○外語中心
3	成績評量的方式?	2007/8/8 下午 12:04:04 --by 9322004, Nick	○教務處課務組
4	學分問題	2007/8/8 上午 11:57:52 --by 9317121, student	○企業管理系
5	自由選修的問題	2007/8/8 上午 10:40:26 --by 9321181, ~~	○教務處課務組 ○保險金融管理系
6	有關於選課的問題	2007/8/8 上午 10:34:49 --by 9321043, Nat	○保險金融管理系
7	請問教官	2007/8/8 上午 10:15:56 --by 9327057, 徐伯瑄	○軍訓室
8	通識課程	2007/8/8 上午 09:58:42 --by 9329016, 洪育景	○通識教育中心
9	新版網頁沒有看到可以點紅磚的地方~	2007/8/8 上午 12:15:10 --by 9527453, 學生	○電算中心網路組
10	如何重修通識課？	2007/8/8 上午 12:10:23 --by 9327027,	○通識教育中心

圖 1 朝陽 FAQ 介面圖

4	証照檢定補助	2007/8/8 下午 06:54:51 --by 9322020, Claire	○研發處研發組 ○學務處學生發展中心 ○研發處國際事務中心

圖 2 單位轉送問題

3	延畢事宜	2007/7/26 下午 10:50:51 --by 9211058, 唉	○學務處 ○教務處 ○學務處生輔組 ○教務處註冊組 ○教務處課務組

圖 3 多單位處理同一問題

針對以上可能的狀況，我們有以下的處理方式：

（1）、我們將以最後回覆之處理單位為依據，將該單位之回覆置於系統介面上供使用者觀看，並將該筆資料加入資料庫以及本體論架構中。

（2）、我們將每個單位之回覆置於系統介面上讓使用者觀看，但只將最後回覆單位之資料加入資料庫以及本體論中。因為最後回覆之單位為最能負責該問題之單位。以圖 2 為例，有關證照檢定補助的問題，使用者選擇了研發處研發組，但該單位並無法直接回覆使用者的問題。所以將問題轉發至學務處學生發展中心再由學務處學生發展中心轉發至研發處國際事務中心，並由研發處國際事務中心作出回覆。

（3）、如果使用者提出了一個可能有多個單位可以處理的問題，我們將每個單位可能的處理方式都予以列出，讓使用者可以詳細觀看其所需的資訊，如圖 3。在圖 3 的例子中，延畢的問題由學務處轉發至教務處，再由教務處轉發至學務處生輔組、教務處註冊組及教務處課務組。並由學務處生輔組、教務處註冊組及教務處課務組三個單位都回覆問題。

本論文以朝陽科技大學之 FAQ 系統為例，將依現有的 FAQ 系統的問題關鍵字建立本體論，並以此本體論建立查詢機制，但隨問題的擴充此本體論將動態更新。

本文其餘部分的介紹如下，我們首先在第二節介紹語意網以及探討一些本體論的相關應用。在第三節中將介紹我們的研究架構流程，以及所運用到的技術包括本體論的學習能力。在第四節，們將運用 Protégé3.2.1 實作出本體論的雛型。最後，在第五節，我們總結本文的重點並探討未來的研究發展方向。

2.文獻探討

2.1 語意網

根據 W3C[25]對語意網所下的定義「語意網是現有網路架構的延伸，也就是把資料的意義定義的更明確，能使人們更有效率地使用網路以及電腦所帶來的便利」。現有的網際網路主要使用超文件連結(Hypertext link) 來組織一個分散式的文件系統。意指藉由交互參照(Cross-references) 來連結一系列的文件，使瀏覽可以依照非順序(Nonsequential) 的方式。但是目前網路上的資訊太過龐大與繁雜，所有的資訊都是以文件的方式所呈現，無法處理並過濾使用者所需資訊，這是因為現今的網路內容大部分是為人們設計，而不是為機器閱讀所設計。因此 Berners-Lee 與 Fischetti 提出新一代的網路資訊架構「語意網」(Semantic web)的構想[9]，強調將網路上的文件有意義的結構化，建立起一個資訊分享及知識可重覆使用的網際網路空間，也就是將全球資訊網中的資料，變成電腦能理解的資料型態，希望電腦能夠了解使用者輸入字串的真正意思，藉以幫助使用者取得想要的資訊。由於語意網是現有網路架構的延伸，其最主要的功用乃於提供詳盡的語意描述語言，使得軟體能夠利用這些語意描述來進行資料搜尋與過濾，提供全方面的資訊來輔助使用者，使資訊變成機器可以理解的，而不再只是機器可以讀的。也因此有助於新科技的進步和工具服務的發展。

2.2 本體論的概念及應用

在過去數十年中，有許多關於本體論的定義被提出，其中最常被引用的是Gruber 於 1993 年所提出的定義：「本體論是對於群體共享的概念化之正式的、明確的表示形式」(An ontology is a formal, explicit specification of a shared conceptualization)[10]，其中「概念化(Conceptualization)」是指對現存的某個現象或領域的確定現象之相關概念抽象模型；「共享(Shared)」則是指本體論是一個共享的部分，

屬於群體而非個人；「正式的(Formal)」是指本體論是機器可以讀的、可以理解的；「明確的(Explicit)」是指本體論的概念形態及限制以明確的方式表示出來。

簡單來說，本體論就是清楚描述一個領域內的所表達的概念和與概念有關的特徵(Properties)及屬性(Attribute)，再加上屬性的限制(Constraint)和依此分類法所產生的實體(Instance)。

在諸多的研究中，我們不難發現這些本體論的應用中，大部分有一個共通點就是運用於查詢服務，將零散的資料建構成本體論後，將有助於快速且正確的查詢。本體論具有在特定領域內描述相關概念與關係的能力，其中概念是用來表示實體世界的任何事物，而關聯通常為sub-class或是has的關聯如圖4。

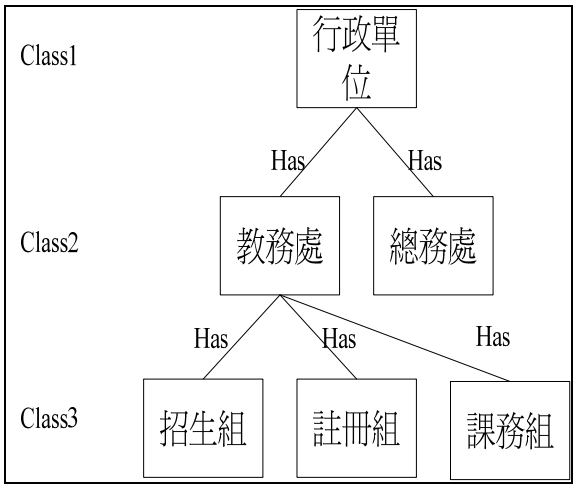

圖 4 概念關聯範例圖

由於本體論的語意(semantic)與推論(reasoning)特性，因此可利用本體論來找尋相關網頁，目前已經有一些運用於單一領域的查詢，以下介紹利用領域知識及本體論的方式來找尋相關網頁的文獻。

[4]在天氣查詢方面以提供台灣地區天氣查詢的服務為目的，並在提供天氣資訊時，運用一些自定的多媒體呈現方式，透過使用者查詢結果的回覆，讓使用者感受到所接收資訊的獨特性。

[11]是在醫學方面的應用方面，則提供一個以醫學資訊為主的資訊檢索代理人。利用醫學資訊本體論來產生查詢，在使用者查詢界面中可利用本體論來查詢關鍵字與關鍵字之間相關的資訊，並且提供特定與概念性的查詢，幫助使用者查詢醫學文獻。

[13]而在中文新聞的檢索應用方面，提供了一個可以擷取中文電子新聞摘要的代理人。可以有效的從中文電子新聞中，擷取新聞摘要，並透過圖形介面供使用者查詢。使用者可藉由此系統得到必要的資訊，而不必浪費查詢的時間。

[12]是語意客服支援服務系統以模糊型態的概念分析方法(Fuzzy Formal Concept Analysis)為基礎，發展了自動生成的本體論，加強對線上客服系統的支援。

[6]採用本體論推理(Ontological Inference)之使用者意向萃取(User Intention Extraction)、查詢擴充與概念式擷取的方法後。本體論可以幫助我們推論概念與概念之間在領域架構中的關係，並且去除可能發生語意混淆的情況。

[14]利用 Yahoo 已經分類好的運動目錄來當成本體論的撰寫，在查詢時所遇到模糊不清的情況，則計算本體論中概念與概念之間的關聯係數，以找出最接近的概念，同時也使用延伸概念來找尋相關的網頁。

[15]將使用者輸入的關鍵字，對應到本體論上，並且延伸其可能的語意概念，再使用布林函數交集的方式至各個搜尋引擎取得網頁，經由文章過濾方式將結果以相關性排序回傳給使用者。然而利用本體論所延伸出的概念，再使用布林函數交集的方式輸入搜尋引擎，會有遺失掉相關網頁的可能，且需花費較多的時間。

[16]結合傳統的 TF-IDF 方法，利用本體論及預先建立好的詞典(Thesaurus) ，來定義

網頁中概念的直接關係或間接關係，透過這些關聯來描述網頁，再利用向量模式來表達一個文件，最後使用餘弦定理計算兩向量的夾角，求出網頁間的相似度，以取得相關的網頁。

[17]結合鏈結叢集的方法，從網頁之間的鏈結描述中萃取出關鍵字，對應到本體論中以找出其含義及可能語意，稱為 Link semantic ，最後利用關鍵字在本體論中的深度來決定兩個網頁之間的相似程度，並將網頁分群，最後提供查詢的界面，以找出相關的網頁。其主要缺點為缺少潛在語意的評估，且 Link semantic 的資訊量過少，未必所有的網頁皆有充足的描述。因此，對於網頁文件仍需分析其內文，充份利用網頁特徵來找出真正的語意，才能達到搜尋相關網頁的目的。

[2]使用正規化概念分析 (Formal Concept Analysis, FCA) 建構電腦病毒知識本體，其目的是希望能藉由 Ontology 概念化(Conceptualization)、共享(Shared)、正式的(Formal)及明確的(Explicit)特性，建構出一個電腦病毒知識本體。這個電腦病毒知識本體可以協助使用者，藉由所獲知之病毒屬性特徵找出電腦病毒的解毒方式，同時經由知識本體的查詢關聯中，更可以找出各個病毒屬性之間的關聯性，提升使用者對電腦病毒有明確的整體認識。

[7]以本體論為基礎所建立出來應用在 PCDIY 領域之 FAQ 資訊整合系統，可以在分散的環境中，有效的發掘資料、擷取資料，並且整合出高品質的資料以符合使用者的需求。

[5]以知識本體來呈現使用者欲搜尋的產品領域知識，並進一步結合 JTP (Java Theorem Prover)概念推論的方式來達成不同知識本體之間具有相同概念推論的功能。傳統的 PCDIY(Personal Computer Do It Yourself)電腦零件知識本體所呈現的知識只在於硬體的分類概念而缺少應用於網路行銷的相關特徵屬性。將傳統的知識本體改善後，提出了電腦協商採購 (Negotiation for Procuring Personal Computer, NPPC)知識本體得以實現運用至網路買賣採購行為決策上的重要資訊。

[1]將本體論應用至相關文件檢索的架構被設計出來，實作一個雛型系統。系統的輸入為一份文件，而輸出為和輸入文件相關的文件；而系統處理程序主要分成若干步驟：(1)將輸入文件轉換成本體論的格式。(2)若輸入文件已存在於系統中，則直接輸出相關文件。(3)若輸入文件不存在於系統中，則進行輸入文件和已存在於系統中文件的相似度計算。其中，本研究設計兩種相似度計算方法來計算相似度，並搭配遺傳演算法來分別計算兩種相似度計算結果所對應的權重，完成最終的相似值。

[8]提出一種以本體論為基礎的文件段落擷取方法，以期能有效的協助網路學習者在網路文件上獲取正確且快速的相關知識。本研究根據學習者的語意，以多個關鍵字詞來評斷文章段落中，哪些是學習者欲知的知識內容。而本實驗結果也發現，以多個關鍵字詞所擷取的文章段落比單一字詞擷取更能符合學習者的語意。多個關鍵字詞所擷取出來的文章段落，能有效避免單一關鍵字詞本身擁有一字多義的情形發生。藉由領域本體論與自然語言處理的結合，把使用者所輸入的問題進行分析，找出文章中符合語意的段落回應給使用者，促進知識的分享與再利用。

[24]基於現行之研究背景及方法所面臨的問題，可以了解到本體論在語意網中扮演的重要角色，及其本身在建構上的困難度。作為代理之間互通的管道，本體論應該要能隨著資訊的快速變動而隨時更新，因此關於自動化建構本體論與本體論更新便成為語意網推行成功的一個重要課題。本文作者提出一個可以快速的自動產生本體論之概念，從網際網路中截取相關領域網頁文件，利用HTML(Hypertext Markup Language)的標籤選擇所需要的關鍵

字詞，藉由這些字詞在文件中之關係 建構一個屬於語意網中的語意本體。

[3]中作者旨在探討知識提取的策略與實施方式，特別是以領域本體為基礎的知識庫，其建置是以XML-based的OWL及描述邏輯為主，以植物學門的維管束植物做為領域對象，並採用Protégé、OWL Plugin以及Racer為領域本體開發平台，實際驗證知識庫的建置，而知識提取則利用OWL-QL做為查詢語言。研究結果顯示，知識提取不同於傳統的資料搜尋或查詢，特別是領域本體均以邏輯為建置內涵，因此亦須以邏輯為基礎的查核才能彰顯其效益。

在以上本體論相關應用的文獻探討方面，不難發現雖然本體論可以應用在多方面的領域中，但大多是應用在資料的檢索方面，而具有與使用者互動並且擁有學習能力且可以擴充本體論卻尚不多見。因此，本論文希望提出一個可以與使用者互動，並且藉由與使用者的互動進而達到本體論的擴充的 FAQ 系統。

2.3 本體論的呈現與儲存

雖然 XML 提供一個一致性的框架(Framework)和一組分析工具(Parser)，可應用於資料間的交換，但其缺乏提供語意的表示方法，為彌補 XML 在本體論應用上的缺陷，全球資訊網協會(W3C) 主導並結合多個元資料(Metadata) 團體所發展而成一個資源描述架構 (Resource Description Framework，RDF)[23]。RDF 是一個用來表示 World Wide Web 上有關資源訊息的一種描述語言，專門用來表示 Web 資源的元資料，如 Web 頁面的標題、作者和修改時間，Web 文件的版權和許可資訊等等。RDF 利用一套特定的術語(Triples) 來表達陳述中的各個部分，用於識別事物的部分稱為主詞(subject) ，區分陳述物件主詞的不同性質(如：作者，創建日期，語種等等)的部分稱為述詞(predicate) ，而受詞(object)則是指定給的性質的值。現行的本體論語言除了 RDF 之外還有 OWL，以這兩種為目前研究者最常使用的本體論建構語言。

3.研究架構

3.1 研究流程

一個具有學習能力，並且可以與使用者互動的自動化線上 FAQ 系統，除了可以讓使用者能快速得到回覆，更能依據新得到的資訊，對本體論加以擴充以應付更多的需求。此外也能根據本體論的架構，讓使用者能夠觀看與所提出之問題相關的問題，滿足使用者可能的需求。以下圖 5 為原 FAQ 系統之流程圖：

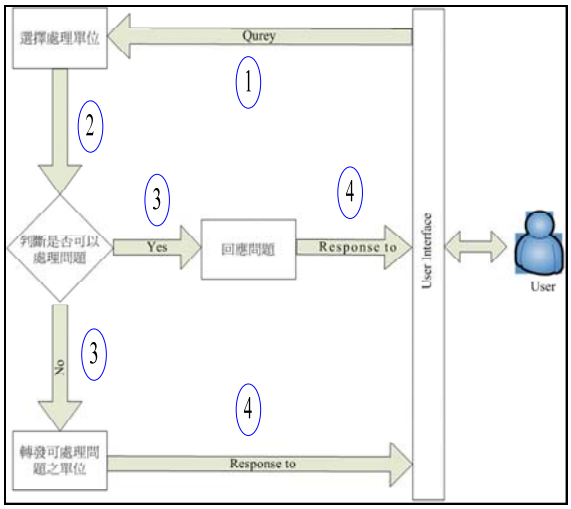

圖 5 原 FAQ 流程圖

而在我們的新 FAQ 系統中，我們將研究流程為六個步驟，以下為我們的研究流程：

(1)收集所有行政單位的網頁資料以及 FAQ 資料，建立出初步的本體論架構以及資料庫。

(2)從使用者所提出之新問題中取得關鍵概念，並且利用所取得之關鍵概念進行相似度的比對，比對結果如有相似的資料則將該筆資料回傳給使用者；如無相似之資料則進行步(3)。

(3)利用 TF-IDF 計算出關鍵概念在步驟(1)

所收集之網頁中的重要性，並利用計算後的結果來分析關鍵概念和網頁間的相關性。

(4)以和關鍵概念相關性最高的網頁所屬之行政單位為轉發問題之單位。

(5)將該關鍵概念加入本體論的架構中，並將該筆 FAQ 加入資料庫中。

(6)以人工方式去驗證新加入之概念是否正確。

以下我們對每個步驟流程詳細敘述。

(1)首先在步驟(1)中，我們收集一定數量的 FAQ 資料，利用 Protégé 建構出本體論，並且以 Access 建立資料庫。並且將學校行政組織的網頁擷取下來，將在步驟(3)中使用。

以下圖 6 為我們所建構出來部份本體論的架構階層關係。實線部分為 has_subclass 的關係，虛線部分為 has_instance 的關係；矩形 class，圓形為 instance。在圖 6 中，每一個 class 都可能有自己的 subclass 或者 instance。以教務處為例，其下有四個單位，我們將這四個單位視為其 subclass；並且教務處負責學期課表這樣業務，我們將其視為一個 instance。

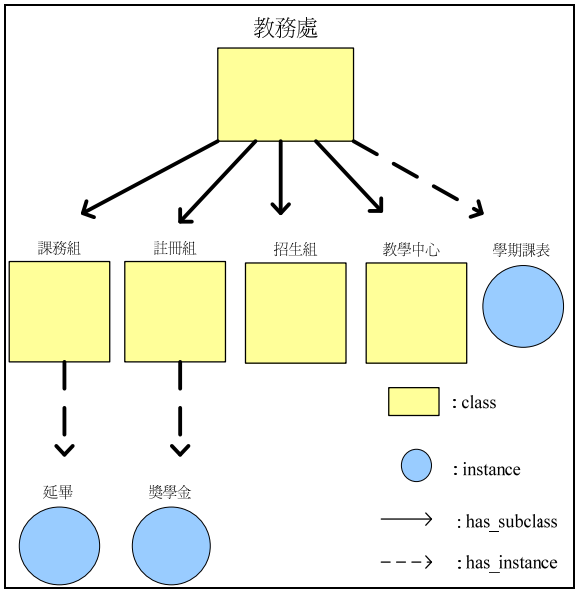

圖 6 部分階層架構圖

在步驟(1)中，所取得的 FAQ 資料以及其關鍵概念將在步驟(2)中的相似度計算中被使用。在這裡我們將所有取得的關鍵概念，分別統計出其在資料庫中出現的次數方便未來的相似度計算。

(2)步驟(2)中，我們利用 CKIP 斷詞系統（如圖 7 所示），處理使用者所提出來的問題並取得關鍵概念。在本研究中我們主要是要取得斷詞後所得到的名詞詞類。再利用相似度比對的方法去判斷該筆資料，是否已經存在於先前所建構成的資料庫中。在相似度的計算中，我們使用內積(Inner Product)的方法來計算新筆 FAQ 資料與資料庫中 FAQ 資料的相似度。我們利用步驟(1)中，所統計出關鍵概念的出現次數以及新筆 FAQ 所取得之關鍵概念已內積的方法計算。計算出資料庫中的 FAQ 資料與新筆 FAQ 的相似度，如為資料庫中所沒有與新 FAQ 相似之資料則我們進行步驟(3)、(4)及(5)，讓本體論做學習的動作，達成本體論的學習以及擴充；如資料庫中有相似之資料則我們利用本體論的架構當索引到資料庫中將該筆資料找尋出來並且回傳給使用者。圖 8 所示，即為我們的系統流程及架構。

圖 7 CKIP 自動斷詞系統

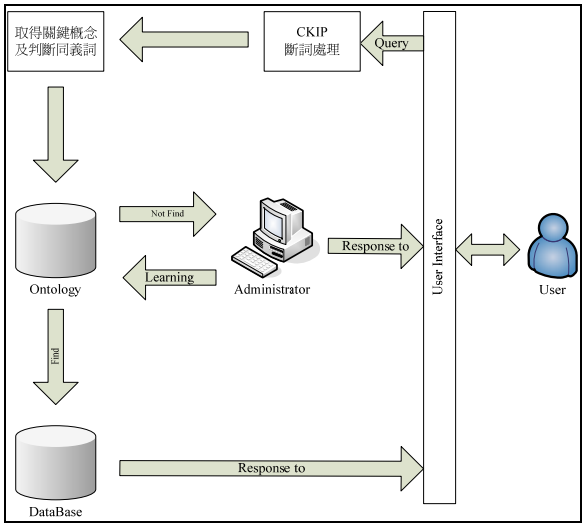

圖 8 查詢流程圖

(3)在步驟(3)中，我們利用由步驟(2)所取得之關鍵概念以及步驟(1)中所收集之網頁導入 TF-IDF 計算出關鍵概念在網頁中的重要性，並利用計算後的結果來分析關鍵概念和網頁間的相關性。

(4)在此次步驟中，我們由步驟三所得到的結果來判斷該問題應該由哪一行政單位來處理。我們以和關鍵概念相關性最高的網頁所屬之行政單位為轉發問題之單位，將此問題轉送該單位處理。

(5)由步驟四所轉送之單位回覆問題後，我們將該關鍵概念加入本體論的架構中，並將該筆 FAQ 加入資料庫中。而該關鍵概念所應該加入之階層架構即為該行政單位之子階層。

(6)在完成本體論之學習以及擴充後，我們採用人工方式去驗證新加入之概念是否正確。我們以點閱該行政單位之網頁的方式，查看是否該問題確實由此行政單位處理，來判斷本體論是否有正確的學習及擴充新一筆的 FAQ 資料。

3.2 研究技術及方法

3.2.1 本體論的建構

使用本體論語言來建構本體論的方式可以分為手動建構、半自動建構，但是都有其不同的限制與使用方法。手動的本體論建構方法比較偏向建構者的主觀意識，爭議也較多；而半自動的本體論建構方法主要分為四大類型，分別以字典、文字分群、關連式法則、知識庫為基礎的建構方法；而全自動的本體論建構方法仍處在發展中，但全自動的建構方法在資訊爆炸的網路時代是必行的趨勢[24]。

近年來，有相當多的工具被提出來，例如：由史丹佛大學所研究出來的 Protégé[26]，由馬里蘭大學 MIND 研究室所研發的 SWOOP[18]，以及一些半自動的本體論建構工具 OntoTrack[19] 、 OntoSeek[20] 、 OntoEdit[21]等。

在本研究中，我們將使用 Protégé 來建立本體論，Protégé 支援 RDF(Resource Description Framework)資源描述架構，RDF 可以描述網頁資源和其它資源之間的關聯性，並且利用類別、屬性、領域、範圍來描述資料。Protégé 是一個由史丹福大學所開發的自由軟體，主要是以知識庫的架構為基礎來編輯本體論，分為 Protégé-Frames 和 Protégé-OWL 兩大主要模型來建構本體論，利用 Protégé 來建構本體論，可以運用的本體論語言包括 RDF、OWL、XML Schema，Protege 是以 JAVA 程式語言所寫成，加強了本體論應用環境的可擴充性，並且提供了快速編寫本體論原形時的彈性。它具備以下優點：

(1). 簡單易學
(2). 免費軟體
(3). 支援多種知識本體
(4). 支援多種儲存格式
(5). 支援多種資料型態
(6). 提供完整應用程式界面
(7). 提供外掛的Plug-in 程式

在Protégé中包含許多本體論的定義；描述一個領域中的概念(concept)稱為class，描述每一個概念的屬性(attribute)與性質(property)稱為slot，描述屬性的限制(restriction)稱為facet，

類別(class)所產生的實例稱為instance。上下層類別之間可以有繼承(Inheritance)關係，subclass可以繼承superclass 所定義的slot，也可以使用slot 描述類別與類別之間的關係(relationship)，包含類別與類別的實例，是一組完整的知識概念。利用Protégé建構的本體論可以說明階層中每一個類別(Class)所包含的屬性(Slot)、類別和類別之間的關係(relationship)、描述屬性的限制(restriction)的面向(facet)、屬性的型態(data type)。

在本研究中我們所使用的版本是在 2006 年 12 月所更新的 3.2.1 版。圖 9 即為 Protégé 的系統介面圖。

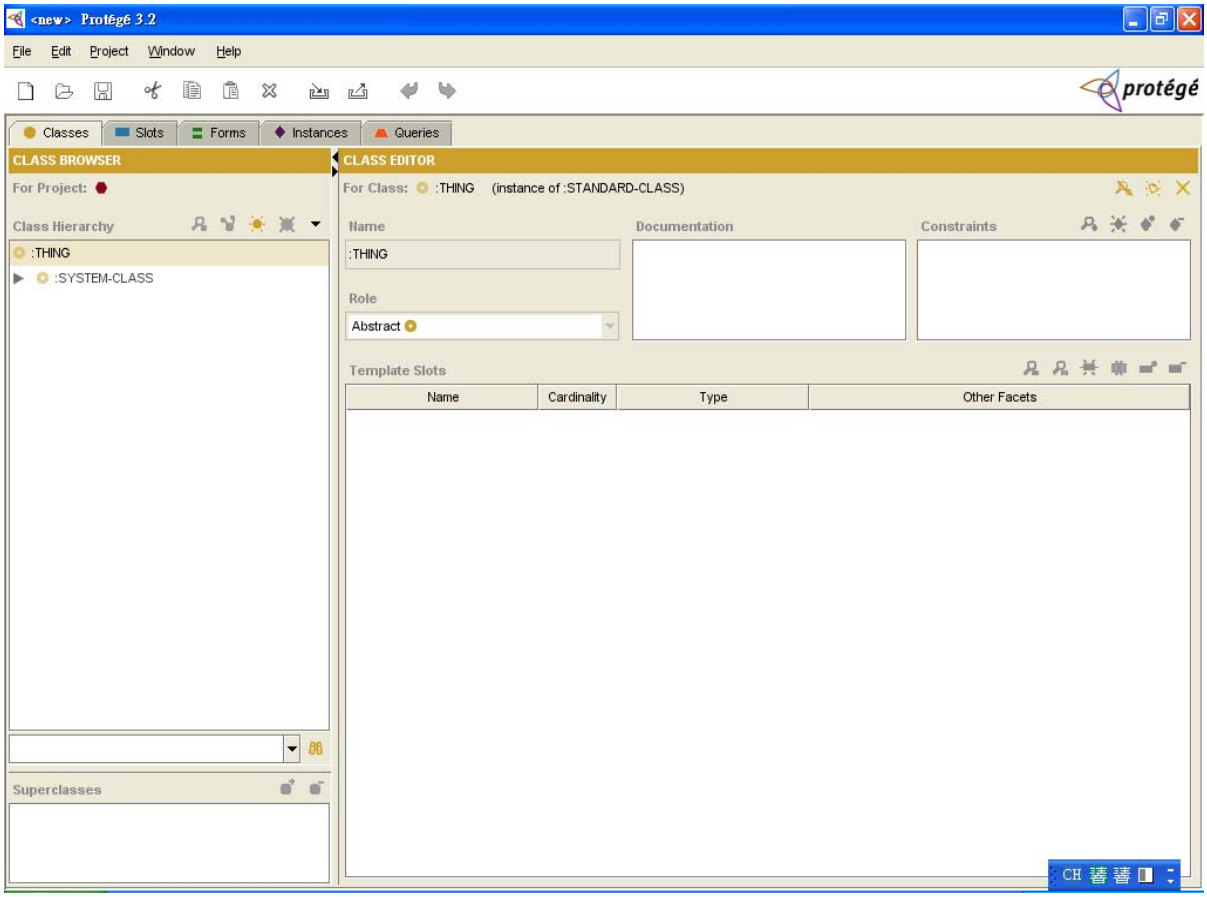

圖 9 Protégé 系統介面

3.2.2 取得關鍵概念

在關鍵字萃取之前，必須先進行文件分析的動作，例如中文斷詞方式的選擇，英文詞彙的動詞時態變化，名詞複數型態，以及刪除無法用來區分文件差異的常用字等等。文件分析的各步驟都會影響詞彙重要性的判斷，因此文件分析階段是非常重要的步驟[22]。

在本研究中我們使用 CKIP 中文斷詞系統，來取得關鍵概念。詞是最小有意義且可以自由使用的語言單位。任何處理語言的系統都必須先能分辨詞才能進行進一步的處理，例如語言分析、資訊抽取。因此中文自動分詞的工作成了語言處理不可或缺的技術。

由於中文詞集是一個開放集合，不存在任何一個詞典或方法可以列出所有的中文詞。當處理不同領域的文件時，領域相關的特殊詞彙或專有名詞，常常造成分詞系統因為參考詞彙的不足而產生錯誤的切割。為了解決這個問題，最有效的方法是補充領域詞典加強詞彙的搜集。因此新的詞彙或關鍵詞的自動抽取成為

分詞的先期準備步驟。領域關鍵詞彙多出現在該領域的文件中而少出現在其它領域，因此抽取關鍵詞時多利用此特性。高頻的關鍵詞比較容易抽取，少數低頻的新詞不容易事先搜集，必須線上辨識。構詞律、詞素、詞彙及詞彙共現訊息，為線上新詞辨識依據。

而中文的詞彙是由一個或兩個以上的字所組成，中文的詞彙與詞彙之間是連續無間隔的，所以電腦很難判斷出哪些字可組成一個詞。常見的斷詞法有，字典式斷詞法、統計式斷詞法、關係法則運算及文字與文件的相關性，每種斷詞法都有其優缺點。目前中央研究院詞庫小組提供的 CKIP 線上斷詞服務，便是以字典式斷詞法為基礎的斷詞方式，因此在處理文字分析上，我們採用 CKIP 線上斷詞服務。

在經過斷詞後，要刪除常用字如：中文的代名詞(你、我、他)、副詞(也、了、的)等等，這些字在資料分析上是不具有特殊意義的。這類單字統稱為常用字。必須將常用字從斷詞後的結果拿掉後，才能利用剩下有意義的詞彙，再作進一步分析。在本研究中，我們所需要取得的關鍵字，是經過分析斷詞後取的名詞。

3.2.3TD-IDF(Term Frequency-Inverse Document Frequency)

TF 的概念是 Salton & McGill(1983)提出，而 IDF 的概念則是 Spark Jones (1972)所提出來的，提出該架構的理由，是因為每個文件中的詞佔整篇文件的重要性，其實是不太相同的。因此產生了 TF 和 IDF 的概念，這兩者的結合，就可以衡量詞的顯著值(重要性)。其概念包含如下：

(1)Term Frequency(TF)：是指該詞彙出現在該文件中的相對頻率，也就是該詞彙在該文件中出現的頻率越高，表示該詞彙就越重要，足以成為此文件的關鍵字。

(2)Inverse Document Frequency(IDF)：DF 是指文件頻率，表示整個文件集合中，有多少文件出現這個詞彙。IDF 是指文件頻率的倒數。

(3)TF-IDF 之計算公式如下：

$$W_{i,j} = TF_{i,j} \times IDF_i = \frac{tf_{i,j}}{MAX(tf_{i,j})} \times \log(\frac{N}{n_i})$$

$W_{i,j}$：某詞彙 i 在文件 j 中的重要性，即權重。

$TF_{i,j}$：代表某詞彙 i 在文件 j 中的相對頻率，即某詞彙 i 在文件 j 中的絕對頻率$(tf_{i,j})$除以該文件中出現次數最多的詞彙頻率 $MAX(tf_{i,j})$。

IDF_i：代表詞彙 i 的反文件頻率，也就是詞彙 i 在所有文件集合中所出現頻率的倒數。

N：所有文件的總數。

n_i：出現詞彙 i 的文件數。

3.2.4 相似度計算

在本研究中我們使用內積(Inner Product)的方法來計算使用者所提出之問題與資料庫內所存在 Q&A 資料之相似度。首先我們將每筆 FAQ 問題以向量方式(Character Vector)表示:

$$CV(Q)=\{(K,W)\}$$

其中 Q 為 FAQ 問題，CV(Q)維問題以向量方式表示，K 為問題中關鍵字 W 為此關鍵字權重。

$$D_i = \sum_{k=1}^{n} (W_k \times W_{ik})$$

D_i ：資料庫中第 i 筆 FAQ

W_k ：新筆 FAQ 問句中關鍵字的權重

W_{ik} ：資料庫中關鍵字的權重

3.2.5 本體論的學習擴充

在諸多的研究中，我們發現在本體論的應用中，主要是以查詢服務為主。將特定領域的資料蒐集後，經過領域專家的分析後，建構出本體論，希望藉由本體論的架構而有助於快速且正確的查詢資料。而為了強化本體論的能力，我們進一步擴充本體論為擁有學習擴充的能力。希望藉由與使用者的互動而可以擴充本體論的架構。

我們將問題區分為單一問題以及二種以上問題。Q_{NEW} 為從使用者提出之問題中取得的關鍵概念，$Q_I \{Q_I \mid Q_I \in Q$，Q 是所有本體論內的 concept$\}$，以下為單一問題的處理方式：

Algorithm

Problem: Single concept

Input: Q_{NEW}

Output: FAQ 的答案

Step1: 利用 TF-IDF 計算出 Q_{NEW} 與所有行政單位網頁的權重

Step2: 選取權重最高的網頁，將該 FAQ 問題傳送至此網頁所屬之行政單位

Step3: 輸出由行政單位回覆的問題並且加入本體論以及資料庫中

Step4: 結束流程

將使用者所提出的問題進行斷詞取得關鍵概念 Q_{NEW}，開始本體論的學習擴充。本體論之學習擴充流程如圖 10。

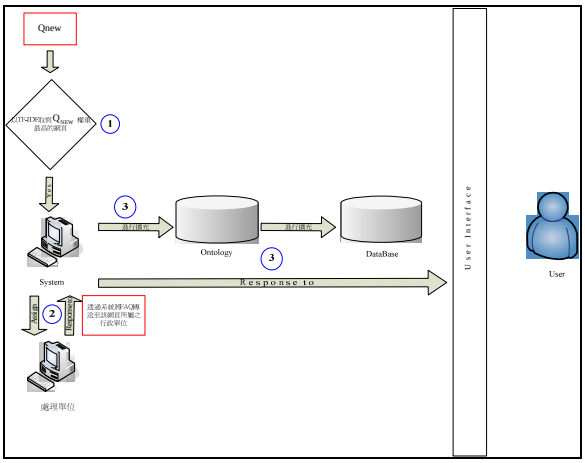

圖 10 本體論學習擴充

首先，我們收集所有行政單位的網頁資料，針對收集回來的網頁資料以及所取得的關鍵概念，利用 TF-IDF 來計算出每一個關鍵概念的權重，以權重的高低來代表關鍵概念在文件中的重要性，利用計算後的結果來分析關鍵概念和網頁間的相關性。以和關鍵概念相關性最高的網頁為主，將關鍵概念所對應到的使用者提問，轉發至相關的行政單位去處理該問題。

當相關的處理單位回覆問題後，將這個新概念新增至相關處理單位的概念階層下，此概念階層為系統依照先前 TF-IDF 所計算出相關性最高之網頁所屬的單位，並將回覆的答案新增至資料庫中。

以圖 11 為例，虛線所圈出來之 instance 為所欲加入之新概念。使用者提出了轉學考的問題，經過 CKIP 的斷詞以後，我們得到了關鍵概念 Q_{NEW}（轉學考），並用 Q_{NEW} 與所有的關鍵概念作比對。經過比對後我們發現，$Q_{NEW} \neq Q_I$，所以我們利用 TF-IDF 去計算出 Q_{NEW} 與所有網頁相關性的高低，並由系統自動轉發給相關性最高的處理單位的網頁。在此次的例子我們可以發現與轉學考相關性最高的網頁為教務處招生組的網頁。於是由系統轉發問題，並在招生組回覆問題後，將 Q_{NEW}（轉學考）新增至教務處=>招生組之下的如圖

11。在將問題的回覆新增至資料庫中，完成本體論的學習擴充。

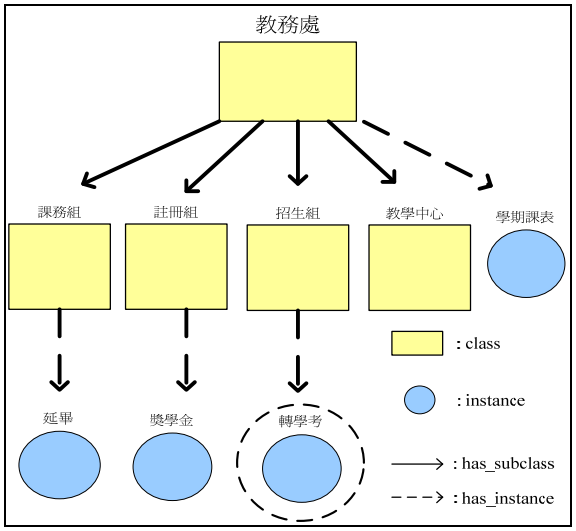

圖 11 新增概念（轉學考）

再者，由於現實的環境中使用者可能會有多個問題想提出，因此我們再進一步考慮由使用者所提出之單一個問題中，可能會包含有多個關鍵概念的狀況。假設 N（Q_{NEW}）代表問題的概念數，以下為多各關鍵概念的處理方式：

Algorithm

Problem: Multiple concept

Input: Q_{NEW}

Output: FAQ 問題的答案

Step1: 利用 TF-IDF 分別計算出每個 Q_{NEW} 與所有行政單位網頁的權重

Step2: 分別選取與每個 Q_{NEW} 相關權重最高的網頁，將 FAQ 問題分別傳送至選取網頁所屬之行政單位

Step3: 輸出由行政單位回覆的問題並且將每筆 FAQ 資料加入本體論以及資料庫中

Step4: 結束流程

首先判斷 Q_{NEW} 的數量是否多於一，如果 Q_{NEW} 數量超過一個，則針對每一個 Q_{NEW} 利用 TF-IDF 來計算出與所有網頁的權重，以權重的高低來代表關鍵概念在文件中的重要性，利用計算後的結果來分析關鍵概念和網頁間的相關性。以和關鍵概念相關性最高的網頁為主，將關鍵概念所對應到的使用者提問，轉發至相關的行政單位去處理該問題。

當相關的處理單位回覆問題後，將這個新概念新增至相關處理單位的概念階層下，此概念階層為系統依照先前 TF-IDF 所計算出相關性最高之網頁所屬的單位，並將回覆的答案新增至資料庫中。

最後為提供使用者可能需要的相關資訊。於本體論架構，將位於同一 class(處理單位)下所存在的 instance(關鍵概念)提供給使用者，讓使用者可以選擇是否願意觀看該處理單位所解答過之問題，以獲得更多可能所需要之相關資訊。

4.實驗環境以及初步實驗結果

在本研究中，我們以朝陽科技大學的學生 FAQ 系統為實作對象。目前收集了八百筆 FAQ 資料，在 Pentium4 2.4G、512MB 的記憶體下的硬體環境，搭配 Protégé3.2.1 來進行實作。

在前置作業中我們使用 CKIP 中文斷詞系統將所收集的 FAQ 資料進行斷詞的步驟。將使用者所提出的問題進行斷詞後，依照所取得的關鍵字以及處理單位，利用 Protégé3.2.1 建立出我們的本體論架構。並藉由所蒐集的 FAQ 資料來進行同義詞的建立。

首先依照學校的行政組織階層作為本體論的基本架構，建立出本體論的概念階層架構。再將取得的關鍵概念加入所建立好的階層架構如圖 13。將我們將蒐集的八百筆資料，以其中二百筆 FAQ 的關鍵詞以人工的方式建立出本體論的實體（instance）雛型架構。再

利用剩餘的六百筆資料以學習的方式將之加入本體論架構中，並由人工方式檢視其所加入後之架構是否正確。圖 12 即為我們目前以六百筆資料學習後的正確性分析，座標軸 X 代表累積第 N 百筆學習資料，座標軸 Y 代表正確性的百分比。而圖 14 為本系統利用所蒐集的部份 FAQ 資料所建立之本體論雛型架構。

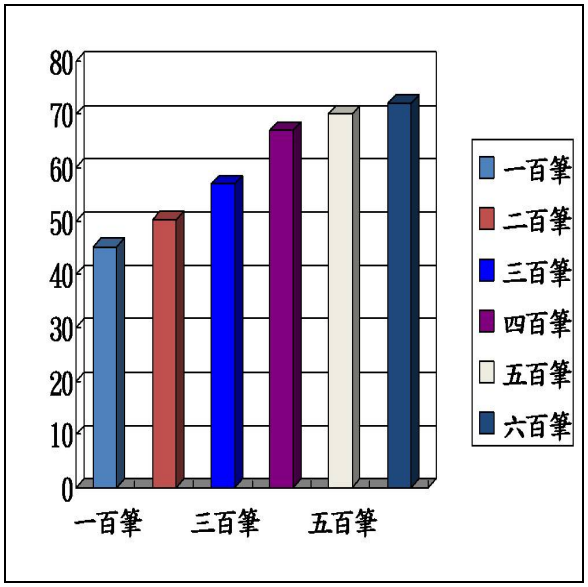

圖 12 正確性分析

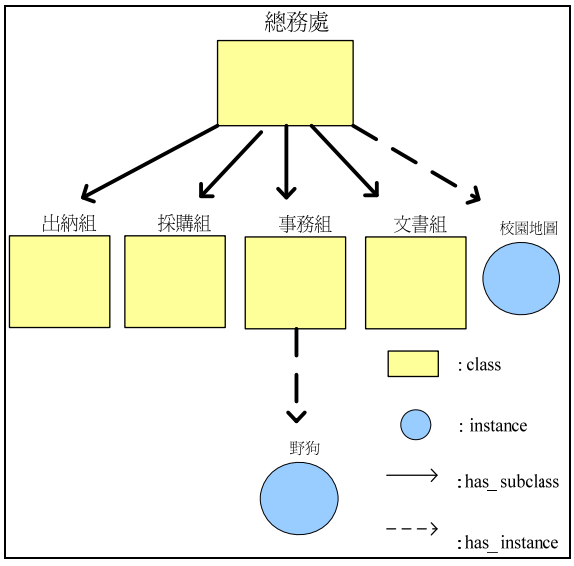

圖 13 本體論加入關鍵概念

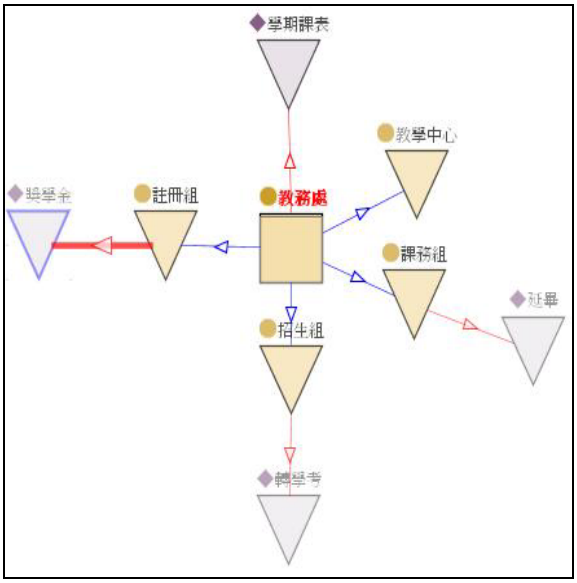

圖 14 系統本體論雛形

實驗的正確性評估我們採用人工的方式。在完成每筆本體論之學習以及擴充後，我們以實際點閱網頁的方式，查看是否該筆資料確實由網頁所屬之單位來處理，如果結果符合則此筆資料為正確的學習以及擴充；反之，則為系統誤判，並以人工方式將之更正。我們以此方式判斷本體論是否有正確的學習及擴充新一筆的 FAQ 資料。

目前初步實驗的結果可以發現本體論內的資料量越大時，未來所學習的正確性越高，目前正確性在百分之七十。而我們希望接下能將學習的正確性提高，能達到自動學習擴充的高正確率。

在未來我們將以圖 14 之本體論架構為基礎，建構出我們的系統並讓使用者實際使用來測試其效果是否達到預期。其中，如何準確的抓取出關鍵概念並進一步提高正確性，為我們接下來的重要課題。而最終目的是達成一個以本體論架構為索引比對使用者所提出問題之關鍵概念，再對應至資料庫中找尋正確的答案，並回傳給使用者。並進一步擴充本體論為擁有學習擴充的能力，藉由與使用者的互動而可以擴充本體論的架構

5.結論與未來工作

利用本體論領域知識的概念來幫助訊息檢索，是大部分研究所著重的。根據領域專家的專業知識所建構出來的本體論，更能夠加速訊息的檢索速度。

藉由本體論的學習擴充功能，可以讓本體論不斷的新增最新資訊，慢慢擴充至可以應付大多數使用者需求，而以自動學習為主，輔以人工彌補不足的方式，可以保證本體論的高正確性，不會有概念階層間的錯誤，配合自動化的回覆系統，可以給使用者更多的便利性。

而本研究除了能夠建構一個能夠加快資訊搜尋的本體論外，還提出了使本體論具有學習能力，並且可以與使用者互動的自動化線上 FAQ 系統。期望可以達到讓使用者能快速得到所需要的資訊，更能讓系統依據新得到的資訊，對本體論加以擴充以應付更多的需求。在未來我們將把整個系統實作，希望能取代原有之 FAQ 系統帶給使用者更多的便利。

參考文獻

[1] 王亮超 (2006)，“領域本體論為基之網頁知識擷取機制設計”，***南華大學資訊管理研究所碩士論文***。

[2] 林建宏 (2006)，“正規化概念分析建構電腦病毒特徵之知識本體”，***國立雲林科技大學資訊管理系碩士班碩士論文***。

[3] 林建良 (2004)，“使用OWL-QL開發領域本體知識庫之知識提取”，***中原大學資訊管理研究所***。

[4] 林智揚 (2004)，“以知識本體為基礎的多代理人資訊系統之研究-以天氣查詢為例”，***大葉大學資訊管理學系碩士論文***。

[5] 邱英華 、陳志豪 (2007)，“基於知識本體之概念推理搜尋代理人-以搜尋電腦零件為例” ，2007年資訊科技國際研討會，朝陽科技大學。

[6] 洪奕璿 (2004)，“採用本體論推理之使用者意向萃取、查詢擴充與概念式擷取”，***國立東華大學資工程學系碩士論文***。

[7] 楊勝源、朱毓君、何正信 (2001)，“以本體論強化網路FAQ系統之解答整合能力”，***第六屆人工智慧與應用研討會***，高雄，台灣。

[8] 施亮如 (2006)，“引用本體論至相關文件檢索之研究”，***國立中央大學資訊管理研究所碩士論文***。

[9] T.Berners-Lee, J. Hendler and O.Lassila(2001),“The Semantic Web”，***Scientific American***, pp. 34-43.

[10] T.Gruber(1993), “Translation approach to portable ontology specifications,” ***Knowledge Acquisition***, vol. 5, pp. 199-220.

[11] J. Abasolo and M. Gómez (2000), “MELISA. An ontology-based agent for information retrieval in medicine,” ***ECDL 2000 Workshop on the Semantic Web Lisbon***, pp.73-82.

[12] Q.T. Tho, S.C. Hui and A.C.M. Fong (2006), “Automatic Fuzzy Ontology Generation for Semantic Help-Desk Support,” ***IEEE Transactions on Industrial Informatics***, vol. 2, no. 3, pp.155-164

[13] C.S Lee, Z.W. Jian, and L.K. Huang (2005), “A Fuzzy Ontology and Its Application to News Summarization,” ***IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics***, vol. 35, no. 5, pp. 859-880.

[14] Y. Labrou and T. Finin (1999), “Yahoo ! as an ontology- using yahoo! Categories to describe documents,” ***Proceedings of the eighth international conference on Information and knowledge management,***

Kansas City, Missouri, United States, pp. 180-187.

[15] R. H. L. Chiang, C. E. H. Cecil, V. C. Storey (2001), "A smart Web query method for semantic retrieval of Web data," ***Data & Knowledge Engineering***, vol. 38, pp. 63-84.

[16] F. Meziane and Y. Rezgui (2004), "A document management methodology based on similarity contents," ***Information Sciences—Informatics and Computer Science: An International Journal***, vol.158, pp.15-36.

[17] I. Varlamis, M. Vazirgiannis, M. Halkidi and B.Nguyen (2004), "THESUS, a Closer View on Web Content Management Enhanced with Link Semantics", ***IEEE Transactions on Knowledge and Data Engineering Journal***, vol. 16, pp. 685-700.

[18] A. Kalyanpur, B. Parsia, E. Sirin, B. C. Grau and J. Hendler (2006), "Swoop: A Web Ontology Editing Browser," ***Web Semantic, Services and Agents on the World Wide Web***, vol.4, pp.144-153.

[19] T. Liebig and O. Noppens (2005), "ONTOTRACK: A Semantic approach for Ontology Authoring," ***Web Semantics: Science, Services and Agents on the World Wide Web***, vol.3, pp.116-131.

[20] N. Guarino, C. Masolo and G. Vetere (1999), "OntoSeek: Content-based Access to the Web," ***IEEE Intelligent Systems***, vol.14, no.3, pp.70-90.

[21] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Stider and D. Wenke (2002), "OntoEdit: Collaborative Ontology Development for the Semantic Web," ***Proceedings of the First International Semantic Web Conference***.

[22] C. Y. Yang, J. C. Hung, C. S. Wang, M. S. Chiu and G. Yee (2005), "Applying word sense disambiguation to question answering system for E-learning," ***IEEE Computer Society***, pp.28-30.

[23] M. C. A. Klein, "XML, RDF, and Relatives (2001)," ***IEEE Intelligent Systems***, vol.16, pp.26-28.

[24] R. C. Chen, J.Y. Liang, R. H. Pan (2008),"Using Recursive ART Network to Construction Domain Ontology Based on Term Frequency and Inverse Document Frequency," ***Expert Systems with Applications***, Vol.34, pp 488-501.

[25] W3C, http://www.w3.org.

[26] Protégé, http://protege.stanford.edu/index.html.