本體論支援之研究學者資訊推薦器

楊勝源
聖約翰科技大學
電腦與通訊工程系
ysy@mail.sju.edu.tw

羅謝逸
聖約翰科技大學
電腦與通訊工程系
96312018@student.sju.edu.tw

張佑任
聖約翰科技大學
電腦與通訊工程系
96312017@student.sju.edu.tw

摘要

網際網路蓬勃發展的時代裡，為使資訊需求者都能迅速精確地搜尋並擷取出有用的資訊，使用者運用網頁分類索引與搜尋引擎等工具，早已成為網際網路上不可或缺的重要能力。本論文結合資訊探勘工具 SPSS Clementine 與領域本體論從大量資料中採掘出有用的重要資訊，再利用 Java 研發一研究學者資訊推薦器 OntoRecommander，針對研究學者推薦適切的重要資訊。初步實驗得知本推薦器的信度與效度均達合理的程度，進而驗證本論文提出相關技術的可行性。

關鍵詞：本體論、資訊推薦器、SPSS Clementine、資訊分類器。

Abstract

In this quick developed and shifting era of Internet, how to make use of webpage indexing structure or search engines, which let information demanders fast and precisely search and extract out advantage information, has become extremely important capability in users on the Web. This paper combined a data mining tool SPSS Clementine with the domain ontology to mine out usefully important information from huge datum, and then to employ Java to develop an information recommender for scholars---OntoRecommander that can recommend suitably important information to scholars. The preliminary experiment outcomes proved the reliability and validation of the recommender achieving the regular-level outcomes of information recommendation, and accordingly proved the feasibility of the related techniques proposed in this paper.

Keywords: Ontology, Information Recommender, SPSS Clementine, Information Classifier.

1. 緒論

搜尋工具的使用早已經成為當代網際網路使用者不可或缺的資訊素養。但在資訊爆炸的時代中，不僅資訊的量是個大麻煩，如何取得適切的資訊更是個重要課題。Google 的出現適度地提供兼顧量與質的資訊查詢方式：關鍵字查詢。然而，僅使用少量關鍵字的查詢，卻往往傳回令人窒息的龐大查詢結果。再者，如此冗長的查詢結果及排列方式，不僅讓使用者必須費時費工的仔細瀏覽後，才能挑選出有用的資訊。造成這樣結果的主要原因：或許是使用者無法清楚表示出自己的查詢企圖或輸入的關鍵字不夠完整；再加上所輸入關鍵字在各種不同領域間均有諸多同名異義字。當資訊系統無法完整且精準地判斷出使用者的查詢需求，以及沒有針對特定領域分別處理的情況下，終究導致系統蒐集過多橫跨數個領域的龐雜查詢結果[5]。本體論（Ontology）的出現就是為了解決上述問題的有力工具。

資料探勘（或資料採礦、資料採掘，Data Mining）則利用統計分析、資訊分類或機械學習（Machine Learning）等相關技術，從大量資料中找出資訊重要的原因、關係或其潛在的規則模式，提供資訊系統決策的重要依據。通常可分兩種基本模式，一種是由上而下來驗證想法是否成立？（假設檢定）可利用數理統計相關技術來進行假設檢定；另一則為由下而上從大量資料中挖掘出其潛在的事實（知識探索），故可利用統計方法或類神經網路等資訊分析技術來進行。常見的資料探勘方法有描述（description）、估計（estimation）、分群（clustering）、分類（classification）、預測（prediction）及關聯（association）等六種[1]。各種方法均有其適用的相關技術或演算法。本論文則結合分類與相關關聯法則進行資訊探勘，提供後端系統進一步處理的重要依據。

要做好資料探勘必須先有一個嚴謹且完整的資料探勘流程。CRISP-DM（CRoss-Industry Standard Process for Data Mining）被公認為進行資料探勘的標準流程，共分成：商業理解、資料理解、資料預備、塑模、評估及部署等六大階段。整個流程能讓資訊需求者在每階段中有條理且準確地進行資訊探勘工作。因此，若能有一輕鬆易學且功能強大的資料探勘工具，便能讓探勘過程更有效率並達事半功倍之效。SPSS Clementine 正具備

上述探勘流程與功能的有效工具[1]。

綜言之，本論文的主題在於應用本體論（或知識本體，Ontology）技術設計出相關研究學者重要資訊的知識本體；搭配 MS SQL Server 資料庫建立出一研究學者資訊相關關鍵字後端知識本體分享平台；應用 SPSS Clementine 作為大量資訊分類與推薦實現之探勘工具；最後，利用 Java 語言建構本研究學者資訊推薦器 OntoRecommander（Ontology-supported Recommender）。換言之，本系統引進研究學者重要資訊知識本體提供分類比對及關聯分析，不僅排除因人為主觀因素所產生的資訊錯誤分析問題，還能使資訊分類統計關聯結果更加強韌，達成兼顧快速、有效支援研究學者重要資訊的推薦系統。

2. 背景知識與相關開發技術

2.1 本體論

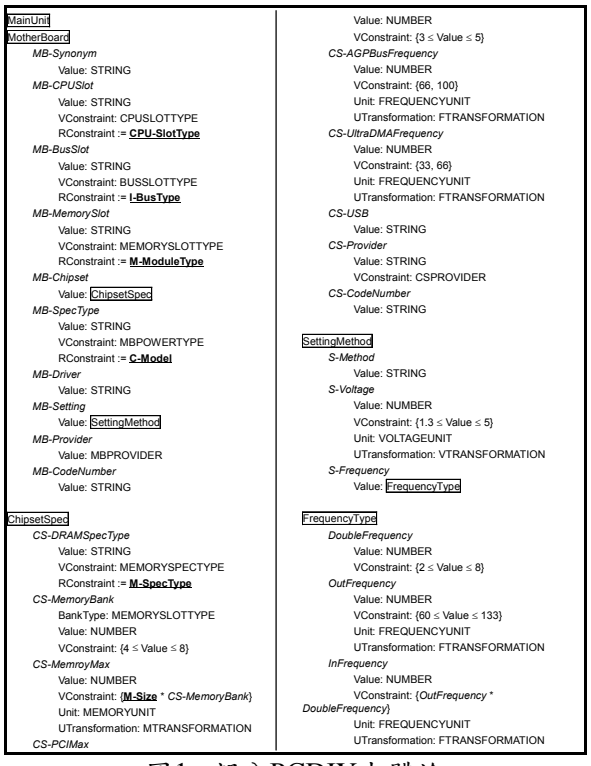

圖1. 部分PCDIY本體論

本體論原本是哲學領域中的論點，主要探討生命或現實事物的知識本質；在人工智慧領域方面，主要是用來定義一個領域的知識內容，表達知識，解決溝通、共用分享等問題；資訊科學領域中，對於電子商務（E-commerce）及知識管理（Knowledge Management）等研究發展有很大的幫助[4,5]。本體論能提供完整的語意模型，意指在特定領域中，有關於該領域相關的各式物件、物件屬性與物件彼此間統一架構下的基礎知識，具有共用與重複使用的特性。透過本體論來描述知識內容的架構，可以完整地呈現一個特定領域的知識核心，自動地瞭解相關領域資訊、溝通及存取，甚或更進一步推論出新的知識與結果，對於資訊系統的建立與維護，是個非常有力的工具[4]。圖 1 就是一個 PCDIY 知識領域的本體論[4]，主要是定義電腦主機（Mainunit）相關的基本知識，顯示部分概念的階層關係及其相關特徵。

2.2 相關探勘工具及開發技術

本系統利用 Java 語言開發相關前端使用者介面程式，後端則以 MS SQL Server 做為本體論資料庫分享平台。MS SQL Server 是目前最常被使用的一種關聯式資料庫管理系統。SQL（Structured Query Language）則是一種常見的關聯式資料庫查詢語言，用來取得資料庫中的資料。

資訊探勘則採用電腦科學領域最常使用的 SPSS 統計軟體為基本工具，省去相關複雜的數學推導過程，僅強調統計學在電腦科學中的應用。SPSS 軟體系列中的 Clementine 軟體，便是以 CRISP-DM 為基準，所開發出來的資料探勘工具，採取獨特的工作串流（work stream）方式進行資料探勘，能讓資訊需求者輕鬆、快速且有效地完成資訊探勘工作。圖 2 即為 SPSS Clementine 工具介面，下方選項版中有各式各樣的基本資料串流功能，提供中央靠左資料流程區建立串流進行相關資料探勘處理。

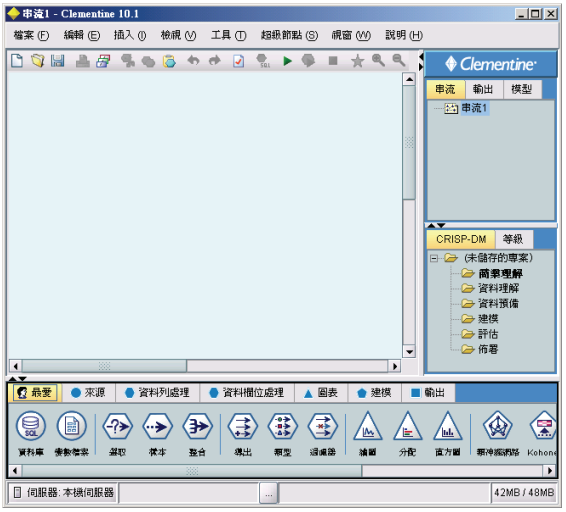

圖 2. SPSS Clementine 工具介面

3. 系統架構及處理流程

3.1 建構本體論資料庫

本論文之本體論資料庫來源是由中華民國人工智慧學會與中華民國模糊學會的歷屆

理事長、監事及理事及各大專院校人工智慧領域相關教授，分別以授課或研究領域中出現的專業關鍵字，作為本論文之本體論關鍵字的建置基底。關鍵字的權重就等於該關鍵字出現在各領域教授網頁中的累計次數，例如：人工智慧領域中，10 位教授出現 9 次人工智慧，就將人工智慧權重設為 0.9；而嵌入式系統出現 6 次，就將其權重設為 0.6，其餘關鍵字權重以此類推，部份本體論資料庫的內容詳見圖 3。

ADMIN-SEVER.crawler - dbo.officek | 物件總管詳細資料

領域關鍵字	關鍵字權重	領域名稱	相關授課	相關連結	授課領域	學術活動
artificial-intelligence	0.9	1<人工智慧>	代理人技術	http://ailab2.et....	<AI授課範圍>	中華民國人工...
ai	0.9	1<人工智慧>	人工智慧	http://ailab2.et....	<AI授課範圍>	中華民國資訊...
natural-language	0.700000000000...	1<人工智慧>	知識系統	http://ailab2.et....	<AI授課範圍>	國際電機電子...
machine-vision	0.700000000000...	1<人工智慧>	人工智慧導論	http://ailab2.et....	<AI授課範圍>	國際電機電子...
robotic-application	0.700000000000...	1<人工智慧>	資訊概論	http://ailab2.et....	<AI授課範圍>	the-institute-of-...
knowledge-base...	0.6	1<人工智慧>	資料結構	http://ailab2.et....	<AI授課範圍>	ieee
intelligent-comp...	0.6	1<人工智慧>	程式語言	http://ailab2.et....	<AI授課範圍>	ieee-computer-s...
object-oriented	0.6	1<人工智慧>	計算機科學導論	http://ailab2.et....	<AI授課範圍>	中華民國計算...
人工智慧	0.9	1<人工智慧>	計算機程式設...	http://ailab2.et....	<AI授課範圍>	中華民國人工...
機器視覺	0.700000000000...	1<人工智慧>	計算機演算法...	http://ailab2.et....	<AI授課範圍>	中華民國模糊...
human-compute...	0.6	1<人工智慧>	JAVA程式設計	http://ailab2.et....	<AI授課範圍>	國際電機電子...
機器學習	0.700000000000...	1<人工智慧>	微積分	NULL	<AI授課範圍>	中國工程師學會
自然語言	0.700000000000...	1<人工智慧>	計算學習理論...	NULL	<AI授課範圍>	NULL
機器人應用	0.6	1<人工智慧>	資料探勘	http://www.csie...	<AI授課範圍>	NULL
知識庫管理	0.6	1<人工智慧>	文件分類與網...	NULL	<AI授課範圍>	NULL
智慧型資料庫...	0.6	1<人工智慧>	非線性規劃	NULL	<AI授課範圍>	NULL
人機互動	0.6	1<人工智慧>	計算機程式	NULL	<AI授課範圍>	NULL
物件導向	0.6	1<人工智慧>	作業系統	NULL	<AI授課範圍>	NULL
man-machine-int...	0.8	1<人工智慧>	資料庫系統	NULL	<AI授課範圍>	NULL
人機介面	0.8	1<人工智慧>	資料庫系統設計	NULL	<AI授課範圍>	NULL
genetic-algorithms	0.9	1<人工智慧>	物件導向程式...	NULL	<AI授課範圍>	NULL
基因演算法	0.9	1<人工智慧>	程式設計語言	NULL	<AI授課範圍>	NULL
遺傳演算法	0.4	1<人工智慧>	programming-lan...	NULL	<AI授課範圍>	NULL
genetic-algorithm	0.8	1<人工智慧>	object-oriented-...	NULL	<AI授課範圍>	NULL
智慧代理人	0.9	1<人工智慧>	database-syste...	NULL	<AI授課範圍>	NULL
intelligent-agent	0.9	1<人工智慧>	database-systems	NULL	<AI授課範圍>	NULL

圖 3. 本論文建置之部分本體論資料庫

3.2 系統架構與流程

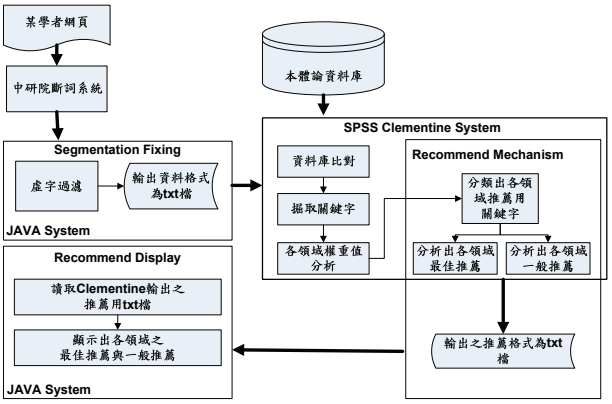

圖 4. OntoRecommender 系統架構圖

圖 4 為描繪出 OntoRecommender 系統架構與處理流程，茲將各功能細節與相關技術描述如后。

(1) 中研院斷詞系統：本系統為中研院資訊所及語言所於民國七十五年成立一個跨所合作的中文計算語言研究小組，共同合作建構中文自然語言處理的資源與研究環境。目前廣為國內外中文自然語言處理及其相關研究提供基本的研究資料與知識架構。本研究中以此系統作為前端輔助工具，將研究學者網頁內文斷詞（segmentation），並過濾大部分虛字（stop words），處理輸出格式為文字檔。

(2) Segmentation Fixing：因中研院斷詞系統會將一些領域特定專有名詞斷字錯誤，例如「臺灣科技大學」斷詞為「臺灣」、「科技」、「大學」。這對後端字詞比對精準度會有極大的錯誤影響。本方塊功能則著重於解決這些問題。功能詳細說明如下[6]：

(i) 前置作業：載入虛字庫 StopWords.txt，目前包含 1700 餘字虛字；去除網頁中無關分類之資料，包括 Tab 鍵、換行、標點符號、連續空白等；最後將文件存入字串陣列，供後續處理之用。

(ii) 斷詞：搜索文件中的每一空白字元，自首字元起，至爾後第一空白字元止，此為一單辭，逐一將文件中每一單辭取出，並存入陣列[2]。

(iii) 過濾虛字：以虛字列（stop list）存放這些字詞，作為索引字詞篩選時必須去除之對象，用以減少文件內容之雜訊，強化效能及增加準確率。

(iv) TF 統計：設置一詞頻陣列對應詞彙陣列的每一個辭彙，以陣列中第一個詞彙為基底起，比對之後的每一個辭彙，如遇重複詞彙，則於基底詞彙之對應詞頻中添加一筆紀錄，並刪除重複詞彙及詞頻紀錄，至最後一個辭彙止。最後，計算一個辭彙在整篇文章中所出現之次數，用以作為分類依據。

(3) SPSS Clementine System：輸入為 Segmentation Fixing 處理後之文字檔案，結合領域本體論資料庫作分類統計與推薦比對分析的依據。此部份處理分成兩階段：第一階段判別此研究學者網頁是否符合特定領域；第二階段則擷取出研究學者的重要資訊作為推薦的重要依據。推薦機制中，除了一般常態推薦外，更加入最佳推薦模組，例如當有一研究學者授課清單中有與其他相關學者授課清單重複時，則此重要科目將會成為「授課」推薦時的最佳推薦資訊。

(4) Recommender Display：本方塊功能為呈現出經 SPSS Clementine 探勘後的推薦結果。本系統以 Java 程式實作本推薦器的使用者介面。

4. 系統呈現與驗證

4.1 系統呈現

圖 5 顯示出經中研院斷詞系統斷詞後的

研究學者網頁內容。圖 6 左半部則為前者除去虛字及計算相關字詞詞頻後的結果。圖 7 顯示出利用 SPSS Clementine 建構出相關探勘串流的畫面。圖 8 右半部則顯示出本推薦器將 SPSS Clementine 探勘出的研究學者重要資訊，分成課程、學術活動及相關簡介按最佳推薦、AI、Fuzzy 及 NN 三個領域推薦出研究學者的重要資訊。

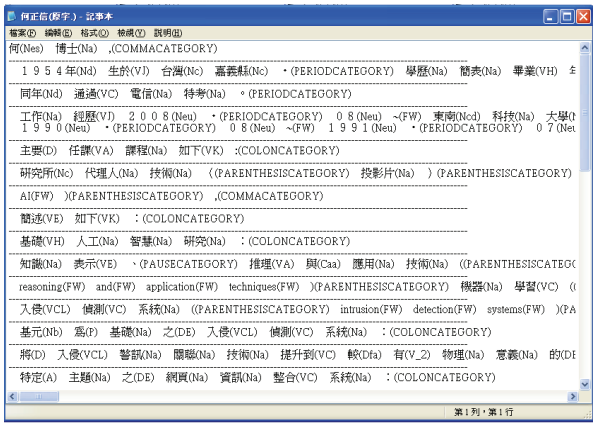

圖 5. 中研院斷詞後之學者網頁內容文字檔

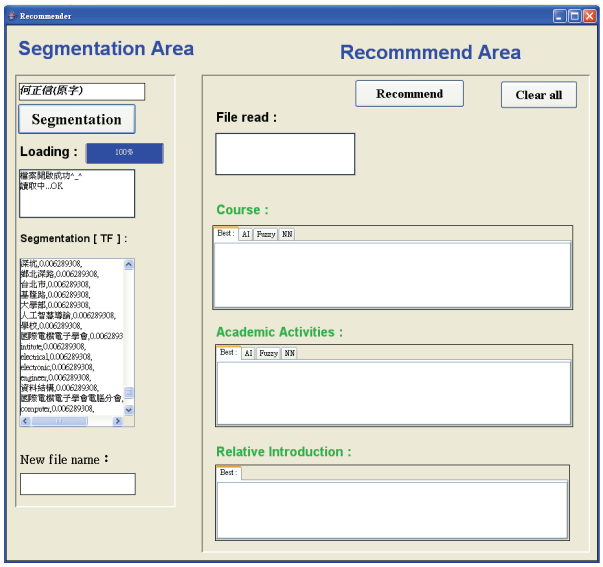

圖 6. 讀取斷詞內容去除虛字並計算相關 TF 值

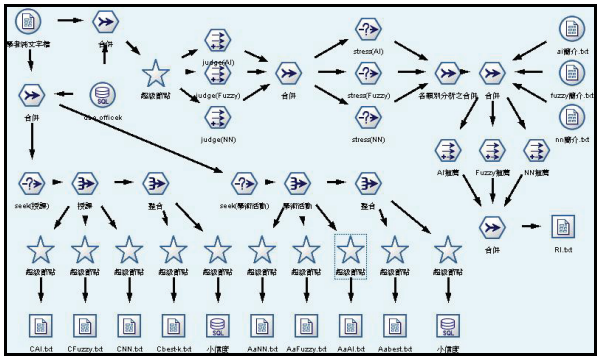

圖 7. 利用 SPSS Clementine 建構資料探勘串流畫面

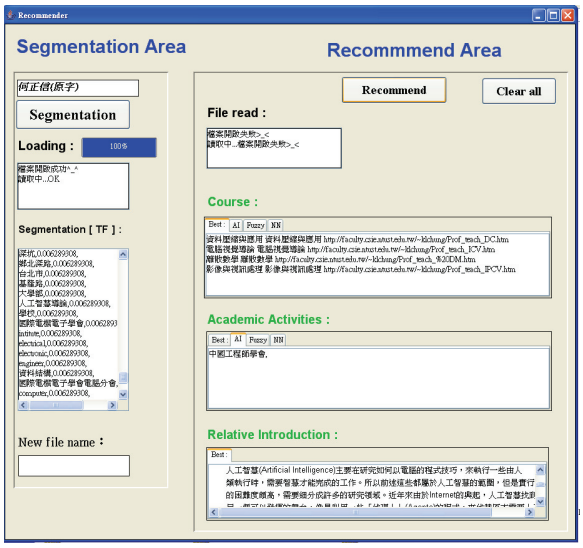

圖 8. 本推薦器的推薦結果

4.2 系統驗證法則

資訊的推薦是由一群相關資訊集合中，選出最佳資訊作為推薦首選。這與大量資料中，抽樣樣本是否能代表母體的程度，具有異曲同工之妙。然而，在抽樣調查的領域中，信度（reliability）常用以檢測抽樣系統本身的準確程度；效度（validity）則注重能否正確反映現象的屬性[3]。換言之，信度評估量度工具的穩定性；效度則注重工具本身的正確性。

(1) 信度

J.P. Peter於1979年借助數學模式表達信度與效度的意義如后。

假設一個測量工具所測得的值為X_o（通常以平均數代表）則X_o可分解為：

$$X_o=X_t+X_e \quad (1)$$

其中，X_o（observed X）表觀察值；X_t（true X）表真實值；X_e（error X）表誤差值。當假設測量所得的變異量為V_o，同理V_o亦可分解為：

$$V_o=V_t+V_e \quad (2)$$

其中，V_o（observed V）表觀察值；V_t（true V）表真實值；V_e（error V）表誤差值。則真實變異量與觀察變異量之比，即為信度：

$$r_{tt}=V_t/V_o \quad (3)$$

但從統計的角度來看，V_t很難直接估算，因此常將公式(3)移項定義信度為：

$$r_{tt}=(V_o-V_e)/V_o=1-(V_e/V_o) \quad (4)$$

換言之，信度即等於1減去「誤差變異量與觀察變異量之比」。

(3) 效度

如果把V_t再分解成$V_{co}+V_{sp}$則

$$V_o=V_{co}+V_{sp}+V_e \quad (5)$$

其中，V_{co}（correlated V）表與測量特質相關的共同變異量；V_{sp}（specific V）表與測量特質無關的個別變異量；則則效度（V_{al}）定義為：

$$V_{al}=V_{co}/V_o \quad (6)$$

(3) 信度與效度的數學關係

將$V_t=V_{co}+V_{sp}$代入公式(3)得知：

$$r_{tt}=V_t/V_o=(V_{co}+V_{sp})/V_o$$
$$=V_{co}/V_o+V_{sp}/V_o \quad \text{(代入公式(6))}$$
$$=V_{al}+V_{sp}/V_o$$

換言之，$V_{al}=r_{tt}-V_{sp}/V_o$，亦即效度應不大於信度。

4.3 數據驗證

表 1 分類資訊的信度數據驗證結果

推薦內容 / 領域 教授	授課			學術活動			專業領域分類
	V_e	V_o	r_{tt}	V_e	V_o	r_{tt}	
何正信	1	12	0.92	2	7	0.71	AI
郭大維	0	2	1	0	1	1	AI
楊勝源	0	5	1	0	1	1	AI
陳錫明	7	11	0.36	3	7	0.57	Fuzzy
許聞廉	0	1	1	2	4	0.5	AI
平均值	0.856			0.756			

表 2 分類資訊的效度數據驗證結果

推薦內容 / 領域 教授	授課			學術活動			專業領域分類
	V_{co}	V_o	V_{al}	V_{co}	V_o	V_{al}	
何正信	11	12	0.92	5	7	0.71	AI
郭大維	2	2	1	1	1	1	AI
楊勝源	5	5	1	1	1	1	AI
陳錫明	4	11	0.36	4	7	0.57	Fuzzy
許聞廉	1	1	1	2	4	0.5	AI
平均值	0.856			0.756			

本實驗相關推薦重要資訊均由領域專家來認定，包括觀察值、真實值、誤差值與相關變異量。表 1 為利用公式(4)計算得出驗證研究學者之「授課」與「學術活動」資訊的信度驗證數據，平均分別為 0.856 及 0.756；表 2 則採用公式(6)計算得出驗證研究學者之「授課」與「學術活動」資訊的效度驗證數據，平均分別為 0.856 及 0.756；最後，值得注意的是：各驗證研究學者的「專業領域」分類也都準確的呈現，更驗證出本論文所提出推薦架構的精準度與可行性。目前，對於相關分類的信度及效度的數據驗證仍在持續進行中。

5. 結論

本論文已應用本體論技術設計出相關研究學者重要資訊的知識本體；搭配 MS SQL Server 資料庫建立出一研究學者資訊相關關鍵字後端知識本體分享平台；接著應用 SPSS Clementine 作為大量資訊分類與推薦實現之探勘工具；最後，利用 Java 語言建構本研究學者資訊推薦器 OntoRecommander。換言之，本系統引進研究學者知識本體提供分類比對及關聯分析，不僅排除因人為主觀因素所產生的資訊錯誤分析問題，還能使資訊分類統計關聯結果更加強韌，達成兼顧快速、有效支援研究學者重要資訊的推薦系統。初步實驗得知本推薦器的信度與效度均達合理的程度，進而驗證本論文提出相關技術的可行性。

本推薦器 OntoRecommander 因採用開放原碼的設計哲學，更利用業界常見的資料探勘工具 SPSS Clementine，為的就是希望所提出的系統架構能廣泛的為感興趣的研究學者採用。持續改善本系統的執行效能、擴充本體論資料庫及持續研發資料探勘中介工具都是未來的重點研究工作。

致謝

作者們對聖約翰科技大學電通系在相關研發設備及環境上的支援，特此致謝。

參考文獻

[1] 宏德國際軟體諮詢股份有限公司，SPSS Clementine 中文版教育訓練講義，台北，台灣，民 95。

[2] 吳文峰，中文郵件分類器之設計及實作，碩士論文，逢甲大學資訊工程學系，民 91。

[3] 吳統雄，“態度與行為研究的信度與效度: 理論、應用、反省”，民意學術專刊，民 74，pp. 29-53。

[4] 楊勝源、吳志偉、何正信，“結合規則挖掘與行為預測來強化網路資訊查詢系統,” *第六屆人工智慧與應用研討會論文集*，高雄，台灣，2001，pp. 574-579。

[5] 楊勝源、朱毓君、何正信，“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論文集*，高雄，台灣，2001，pp. 52-57。

[6] 楊勝源、江洪鈞、吳霽時，“本體論支援之研究學者網頁分類器,” *第十九屆國際資訊管理學術研討會論文集*，南投，台灣，2008，pp. 113。