利用關鍵字分析建立自閉症論文知識庫

張峻彬[1]、謝愛家[2]、盛夢徽[3]
嘉南藥理大學{[1,3]資訊管理系 [2]資訊多媒體應用系}
{cpchang, qajhsieh, dream}@mail.cnu.edu.tw

摘要

本研究中利用 PDF 論文檔做為知識的主要來源，建立一個以自閉症主題的線上知識庫管理系統。系統主要由 PDF 文件上傳、文件驗證、文件內容解析、關鍵字關聯分析和主題知識查詢等五部分所組成。在系統特性上，使用者與管理者均不需具備特定領域的專有知識，論文內容的收集、分析、組織均是自動進行，這個過程主要是利用論文作者自定關鍵字來進行即時性的關聯分析與整理，而在知識的搜尋與應用上，系統中採用主題式的導引，使用者可在不受本身專業知識的限制下，自主性地定位，找出其所需的文獻與相關內容。

關鍵詞：關鍵字、知識庫、主題地圖

Abstract

In this Study, an online Knowledge Base Management System is established by using PDF documents as the source of knowledge. System consists of five parts, include: PDF file upload, validation, content analysis, keyword correlation analysis and knowledge query. In this system, users and manager do not need to have any domain knowledge in specific areas, thesis collection, because the analysis and organizations are automatically. In the search of knowledge, the system uses the Topic Map guide. Users can locate their required documents and relevant content by autonomous positioning function.

Keywords: Keyword, Knowledge Base, Topic Map

1. 前言

由於網路應用的發達，各類資訊的取得相當的容易，也正因為資訊的取得太容易，可能會陷入”過多的資訊等於沒有資訊”的困境之中，由於網際網路並未具有組織高重複性資料的特性，使得資料擷取產生困難，亦有資訊負載的問題[1]。因此在資訊超載的環境中，如何針對特定主題快速地收集、分析、組織、應用資訊變成一個很重要的課題。為了應付前述的狀況，建置一個系統化地保存、管理與分享知識的知識庫是必要的，此知識庫可以協助一般使用者取得知識、了解專門領域的概念。在此知識庫(Knowledge Base)指的是針對某一特定領域的經驗與知識，採用某種知識表示方式在資料庫中儲存、組織、管理和使用的知識集合。

可攜式文件格式(Portable Document Format, PDF)是由 Adobe 在 1993 年所公開出來的格式，目前是最被廣泛應用的電子文件格式。使用 PDF 有兩項主要的優點，一是它跨越各種作業系統平台(WINDOWS、MAC、UNIX...)均有閱讀器可讀取，另一個是它可以保存原始文件樣貌在各種作業系統平台上閱讀或列印，也正因為這兩項優點，許多的學術論文都採用 PDF 做為傳送與保存的格式。

以 Airiti Library 華藝線上圖書館而言，他所提供下載的論文均是使用 PDF 的的檔案格式。若以自閉症為主題(關鍵字)進行搜尋，可以找到 305 篇相關的學術論文(2014/8/28，圖 1)，這雖然不是相當巨大的文件量，但若是使用者想在自閉症這一主題下找到自己有興趣的資訊是相當不容易的，因為這類電子資源搜尋引擎只能針對論文題目、關鍵字及摘要內容進行搜尋，但是現實上論文作者不可能將論文的所有重要內容都放在論文題目、關鍵字或摘要中，所以不在搜尋結果中的論文並不代表論文中沒有使用者需要的訊息。而且就算是使用者將 305 篇相關的學術論文都下載，他會需要花大量的時間才有辦法(或沒有辦法)去進行整理，由於這些論文主要是以個別文件檔案作為處理單位，面對大量的檔案資料，如何運用有效的原則來組織這些論文，以便有效的管理以從中取的資訊與知識，將是一個很重要的工作。

圖 1以自閉症為關鍵字進行搜尋

面對大量的文件可以運用各式的技術，包括：將知識分類群聚（Clustering, Categorization）；將資訊摘要；將大量文章以鳥瞰式呈現，或以不同視覺效果（Visualization）呈現[2]，提供更多知識探索機會，提升其功能。由於人們對知識有效組織與呈現的需求愈來愈迫切，因此在 1999 年時出現了 ISO13250 主題地圖(Topic Map)，可將大量資訊整合，將其以眾多關鍵詞類聚，藉由類似地圖檢索的瀏覽方式呈現知識樣貌，讓使用者對感興趣的主題進行系統化地瀏覽，發揮知識導覽的功效。

本研究的目的是要利用 PDF 論文檔做為知識的主要來源，建立一個以自閉症主題的線上知識庫管理系統。在本研究中利用 45 篇自閉症的學術論文 PDF 檔為例，以這些檔案示範本研究如何建立自閉症主題的線上知識庫管理與查詢系統。

2. 知識組織方式

對於知識的組織與呈現方式可概分為兩大類：由上而下及由下而上（[3]，圖 1）。在由上而下的方式中，屬於半自動化的呈現方式，此方式須事先由專門領域的人定義知識的本體(Ontology)[4]和分類(Taxonomy)[5]，再根據本體或分類去整理分類知識(Subject Categorization)。而由下而上的方式，主要是使用主題分群(Subject Clustering)和知識地圖(Knowledge Map)，可依據知識的特性自動分群。在本文中是採用由下而上的方式，利用學術論文中作者所定義的關鍵字為主題，進行知識的分群，以便架構出動態的知識地圖。

圖 2 知識的組織與呈現方式[3]

由於各個專業領域會有不同的專業術語，為能對各類的知識主題分群，特定應用領域就必須有特定的語彙典。如何快速有效地在文件資料中萃取有限的領域術語(Terminology)，一般而言，針對大量的文件資料，應用文字探勘技術可從中自動萃取出文件關鍵字，而無須再仰賴人工手動建立關鍵字。不過由於本研究所要整理的目標是學術論文，論文作者本來就會為他的文章設定了一些較能代表文章內容的關鍵字了，所以本研究中就以這些作者自定的關鍵字做為知識分群的主題。

在動態的產生知識地圖方面，本研究中是採用類似主題地圖(Topic Map)的概念來進行的。由於網路應用的發達，各類資訊透過網際網路迅速地散播，各類知識累積散佈於各處，因此對知識有效組織與呈現變得很重要，因此在 1999 年時出現了 ISO13250 主題地圖的標準，藉由類似地圖檢索的方式呈現，讓使用者對感興趣的主題進行系統化地瀏覽，發揮知識導覽的功效。

主題地圖的架構可分為主題（Topic）、關聯（Association ）和連結（Occurrence）三部分。其架構的作法可概分為三個步驟：

1.一開始是由分析資源的內容開始，找出可代表各資源之主題。
2.接下來進一步定義出各主題之間的關係，加以串聯，即形成關聯
3.最後利用為連結，便利使用者查詢資料，連接到資源的所在處。

主題地圖實現長久以來各界欲創造一格式表單以控制資訊的格式及編排的夢想，主題地圖即是利用格式表單控制資訊獲得及瀏覽的觀念，詳細敘述各種瀏覽層級，並模擬複雜的知識管理關係，提供量身訂做的路徑檢索資訊，以幫助使用者更有效率的瀏覽數位資源[6]。主題地圖讓人們在處理資源的過程中，可以用概念或主題的方式來描述物件。主題地圖的精神是把某一知識領域內的各個主題或單一主題內所有的附屬子題列出，並透過「參見」(See)和「另可參見」(Also See)的機制，將具有關聯性的主題加以連結[7]。

3. 系統建立

本知識庫管理系統主要由 PDF 文件上傳、文件驗證、文件內容解析、關鍵字關聯分析和主題知識查詢等五部分所組成。

1. PDF 文件上傳：

主要提供使用者透過網路上傳 PDF 文件、刪除與匯入等管理功能。

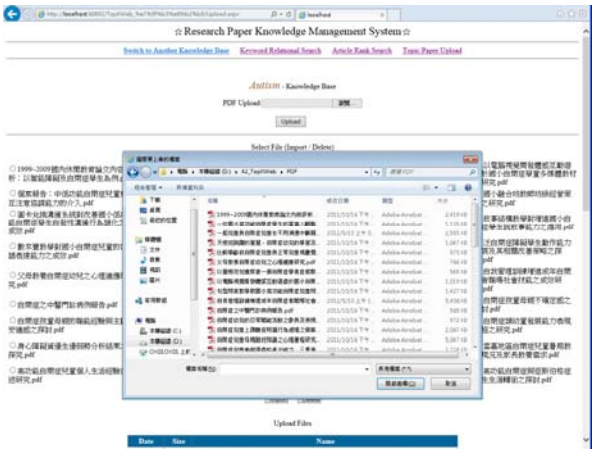

圖 3 瀏覽及上傳 PDF 論文檔

圖 4 匯入與刪除等管理功能

3.文件驗證：

在使用者匯入論文後，檢驗此論文是否符合本知識主題(主題關鍵字至少出現 10 次以上)，自動抓取論文標題與關鍵字等訊息，並於畫面上顯示論文的相關內容以供使用者確認。在使用者確定要匯入此論文後，檢驗知識庫中是否已存在此論文。

圖 5 確認與檢驗論文

4.文件內容解析：

將本文中若有原本不存在知識庫中的關鍵字加入知識庫中，對論文進行基本資訊分析，統計原生關鍵字與延伸關鍵字於本文中出現的次數。分析統計知識庫中原生關鍵字與延伸關鍵字於知識庫中出現的總次數。統計原生關鍵字與延伸關鍵字於知識庫其他論文中出現的次數。

5.主題知識查詢：

主要提供關鍵字於知識庫中的相關訊息；依關鍵字查詢相關論文；單一論文資訊展示，包括本文、關鍵字類型與出現資訊、與此論文具關聯性的其他論文顯示；單一論文相關查詢、關鍵字查詢、關聯性查詢等相關功能。

4. 主題知識查詢

知識庫管理系統在經過 PDF 文件上傳、文件驗證、文件內容解析與關鍵字關聯分析就可提供使用者進行主題知識查詢。進入後有區分為兩種搜尋方式：關鍵字關聯搜尋(Keyword Relational Search)，論文排序搜尋(Article Rank Search)。有別於一般查詢系統會讓使用者輸入想查詢的字串，再根據此字串去找出相關的訊息，在這個自閉症主題關聯搜尋系統中，完全不需要使用者輸入任何資訊，而是由使用者去選擇他有興趣的主題，由這個主題透過關聯分析的導引去找出使用者想看的論文。

圖 6 主功能畫面

以關鍵字關聯搜尋來說，由頁面(圖 7)上可以了解目前在自閉症知識庫中有幾篇論文(目前有 45 篇跟自閉症有關的中文論文)，這些論文總共有幾個原生與延伸關鍵字(目前有 111 個)。在學術論文中，作者都會對他自己的文章定義五個左右，最能代表這篇論文的關鍵字，像這種原作者定義的關鍵字在系統中稱做"原生"關鍵字(Original Keyword)，而在 A 論文中作者定義到的關鍵字，若是出現在 B 論文之中，但是並沒有被 B 論文作者定義為關鍵字，像這樣的關鍵字在系統中稱為 B 論文的"延伸"關鍵字(Extened Keyword)。

圖 7 關鍵字關聯搜尋畫面

在關鍵字關聯搜尋中會依關鍵字出現於論文篇數的多寡排序，列出關鍵字供使用者選擇。在畫面的右側有表列出各關鍵字於知識庫中出現的總次數，以及出現的論文篇數。關聯搜尋功能中，一開始"列出包含選定關鍵字的論文"是沒辦法按的，一定要選擇 1 至 3 關鍵字，這個按鈕功能才可以使用。

圖 8 選擇一個關鍵字(問題行為)啟用搜尋功能

以選擇一個關鍵字(問題行為)而例，可以知道知識庫有幾篇論文中有出現了這個關鍵字，關鍵字出現了幾次，以及在這篇論文中這個關鍵字屬於何種類型。

圖 9 選擇一個關鍵字的搜尋結果畫面

以兩個關鍵字為例，可由知識庫中找出包含這兩個特定關鍵字的相關論文。

圖 9 選擇二個關鍵字的搜尋結果畫面

以三個關鍵字為例，可由知識庫中找出包含這三個特定關鍵字的相關論文。

圖 10 選擇三個關鍵字的搜尋結果畫面

不管是用一個、兩個或三個關鍵字去搜尋，都可以文章列表前的 View 按鈕去看特定文章的內文。

圖 11 特定文章內文與相關資訊

若想看看與本論文關聯性比較高的論文有那一些，可利用觀看本文關聯列表(圖 11 中左邊的列表)，此列表主要是針對本篇論文中的原生與延伸關鍵字依出現的次數(Sort by Count)、知識庫中出現的總次數(Sort by TotalCount)或出現的論文篇數來排序(Sort by PaperCount)，以前三高的關鍵字來找出關聯度較高的文章。Sort by Count，Sort by TotalCount，Sort by PaperCount 三個功能可以直接點擊切換呈現的方式。

若是對圖 11 中”Top 3 Keywords in This Research Paper”標題底下所指具有”相互注意協調能力 & 自閉症 & 自閉症兒童”三個關鍵字的相關文章有興趣，可點擊前面+號將它打開，出現文章項目，點擊文章名稱，就可觀看其內容與關聯文章(Sort by Count)。

圖 12 具有相似關鍵字的高關聯論文(Sort by Count)

依不同的排序方式，可能會找出不同的關聯文章。

圖 13 具有相似關鍵字的高關聯論文(Sort by TotalCount)

而在觀看本文關聯列表(圖 11 中左邊的列表),可用於搜尋有相同原生或延伸關鍵字的相關文章。例如按"高功能自閉症兒童"前的 View 按鈕，就會列出相關的 10 篇論文，供使用者選擇觀看文章相關內容

圖 14 有相同原生或延伸關鍵字的相關文章

點擊圖 14 各文章篇名前的 View 按鈕，就會出現圖 11 的特定文章內文與相關資訊。

介紹完關鍵字關聯搜尋，接下來介紹論文排序搜尋的作法，在進入論文排序搜尋的頁面後，主要分成兩種排序方式：依比例排序(Sort By Keyword Ratio)，依次數排序(Sort By Frequency)。首先說明依比例排序，在這裡的比例指的是利用本論文中原生與延伸關鍵字的總個數除以知識庫中原生與延伸關鍵字的總個數所得到的數值，這個數值愈高，表示這篇論文包含了相當多自閉症主題下的關鍵字，這篇論文應是比較偏廣泛型的文章。

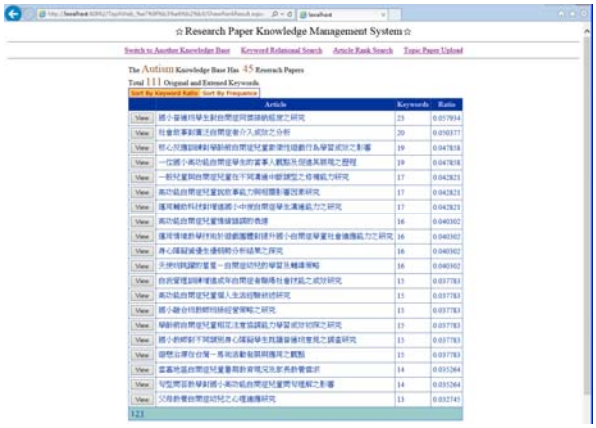

圖 15 依比例排序的畫面

依次數排序的作法，主要是針對各論文中的原生與延伸關鍵字出現的次數進行合計，以合計數去排序。一樣的按 View 按鈕，可看特定文章的內文與關聯文章。

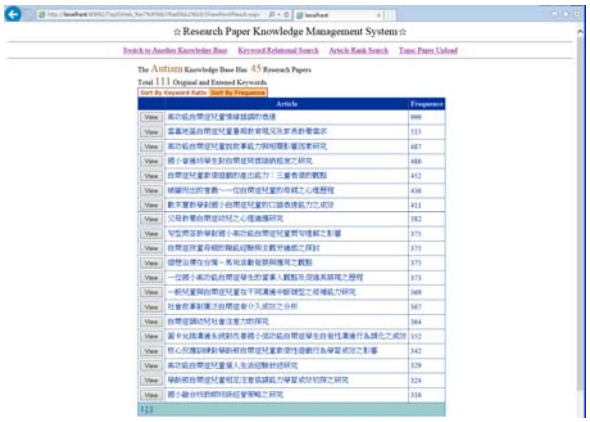

圖 16 依次數排序的畫面

5. 結論

由於網路資訊取得容易，因此常會陷入"過多的資訊等於沒有資訊"的困境之中,如何針對特定主題快速地收集、分析、組織、應用資訊變成一個很重要的課題。以學術論文來說，大都是以個別文件檔案(PDF 格式)的方式存在，面對大量的檔案資料，如何運用有效的原則來組織這些論文，以便有效的管理以從中取的資訊與知識，這是一個很重要的工作。

本研究中成功地利用 PDF 論文檔做為知識的主要來源，建立一個以自閉症主題的線上知識庫管理系統。系統主要由 PDF 文件上傳、文件驗證、文件內容解析、關鍵字關聯分析和主題知識查詢等五部分所組成。在系統特性上，使用者與管理者均不需具備特定領域的專有知識，論文內容的收集、分析、組織均是自動進行，主要是利用論文作者自定關鍵字進行即時性的關聯分析與整理，而在知識的搜尋與應用上，系統中採用主題式的導引，使用者可在不受本身專業知識的限制下，自主性地定位，找出其所需的文獻與相關內容。

參考文獻

[1] Bowman, C. M., Danzig, P. B., Manber, U., & Schwartz, F., "Scalable internet resource discovery: Research problems and approaches," Communications of the ACM, 37(8), pp.98-107, 1994.

[2] Yang, C. C., Chen, H., & Hong, K., "Visualization of large category map for internet browsing," Decision Support Systems, Special Issue on Web Retrieval and Mining, 35(1), pp.89-102, 2003.

[3] 陳文華、施人英、吳壽山, "探討文字採掘技術在管理者知識地圖之應用," 中山管理評論, 2004 年中文特刊, pp.35-64, 2004.

[4] Noy, N. F., McGuinness, D. L., Ontology development 101: A guide to creating your first ontology, Stanford University, 2001.

[5] Bruno, D., Richmond, H. "The truth about taxonomies," Information Management Journal, 37(2), pp.44-51, 2003.

[6] 阮明淑, "知識經濟時代的主題分析", 21 世紀資訊科學與技術學術研討會論文集," pp.351, 2001.

[7] 溫達茂及呂春嬌, "知識組織的工具 – Topic Maps", 21 世紀資訊科學與技術學術研討會論文集, pp.389, 2001.