以正規化概念分析建構中文資料資訊檢索之研究

— 以資訊管理學報為例

黃錦法
雲林科技大學資管所
huangcf@mis.yuntech.edu.tw

王耀德
雲林科技大學資管所
g9923722@mis.yuntech.edu.tw

摘要

近年來科技的日新月異，隨著數位資料急遽成長，大量的數位資料增進使用者在搜尋資訊上的便利，但使用者卻也面臨龐大的搜尋結果，過多的資訊量造成使用者在使用上十分不便。目前大部分的資訊檢索系統均有提供搜尋功能，並且提供一些簡單的瀏覽界面，但是能有效整合瀏覽與搜尋的資訊檢索系統並不常見。正規化概念分析是一種從資料集合中發現概念結構的資料分析理論，它所產生的概念點陣能夠有效整合資訊檢索系統的瀏覽與搜尋。

本研究利用正規化概念分析建構中文資料資訊檢索之概念點陣，並提供中文資料之瀏覽與搜尋的支援系統。此系統在瀏覽上提供多元化的概念點陣瀏覽模式，在搜尋上採用概念點陣由下而上(下確界搜尋到上確界)部分節點搜尋幫助使用者快速且準確尋找資料，並且以樹狀結構和相似度排序來展現搜尋結果。為了評估本系統的檢索品質及執行效能，本研究以資訊管理學報作為實驗資料。實驗結果顯示，在同樣的搜尋品質的情況下由下而上部份節點搜尋的執行時間約為由上而下全部節點搜尋的 56%~71%；同時使用 2 個或 3 個點陣概念時可以使用的關鍵字是使用單 1 個點陣概念的 1.62 倍或 2.12 倍。

關鍵詞：資訊檢索、正規化概念分析、概念點陣、資訊管理學報

壹、導論

近年來科技的日新月異，隨著數位資料急遽成長，大量的數位資料增進使用者在搜尋資訊上的便利，但使用者卻也面臨龐大的搜尋結果，過多的資訊量造成使用者在使用上十分不便。即使有資訊檢索服務的提供，但在數百億份網頁中找到確切想要的資料是十分困難，人類所輸入的查詢字串屬於自然語言，電腦並無法完全理解使用者所想要的資訊，僅能將使用者所輸入的關鍵字與網頁所定義的關鍵字或是內容做比對，以至於關鍵字無法完全切中要點。另外，目前大部分的資訊檢索系統均有提供搜尋功能，並且提供一些簡單的瀏覽界面，但是能有效整合瀏覽與搜尋的資訊檢索系統並不常見。

正規化概念分析(Formal Concept Analysis, FCA)是由 Rudolf Wille 在 1982 年所提出，這是一種從資料集合(Data Sets)中發現概念結構(Conceptual Structures)的資料分析理論，可以將表格式的資料轉換成圖形化的展現方式來導覽及使用，並已快速發展和應用到許多領域中(Ganter and Wille, 1998)。 Carpineto and Romano(2004)的研究中提到利用正規化概念分析所產生的概念點陣(Concept Lattice)可以在資訊檢索領域上提供改善查詢、整合瀏覽與搜尋及結合索引典(Thesaurus)的概念。概念點陣可以有效整合瀏覽與搜尋，在瀏覽上可以利用概念點陣的網狀結構來瀏覽資料；在搜尋上，概念點陣將相關性高的資料自動聚集在節點上，透過節點的搜尋可以快速且準確尋找資料，並且以視覺化的方式展現搜尋結果，這些都是以往資訊檢索系統所沒有的優點。

雖然概念點陣在瀏覽及搜尋上都具有優點，但也有其缺點的地方。在瀏覽上，由於概念點陣是網狀結構與一般使用者所熟悉的階層或樹狀結構的瀏覽模式不同，因此可能會產生使用者瀏覽不便的情形發生；而在搜尋上，目前的研究大多採用概念點陣由上而下搜尋(上確界搜尋到下確界)，此種搜尋方法必須掃描全部概念節點後，才能結束搜尋程序，因此會大幅降低搜尋上的效能。

另外，目前資管期刊多半以出版時間用人工進行分類，搜尋結果的優劣無法衡量，再者，由於資管期刊每年都必須新增期刊，當用人工進行分類時，難免會有遺漏或分類錯誤的情形發生。因此，若是能設計一套可自動歸類相關性的期刊，且可視覺化瀏覽及搜尋的系統，將可以減少教授與學生尋找研究資料的寶貴時間，增進研究上的效率。本研究以正規化概念分析為基礎，設計一套可讓使用者瀏覽和搜尋的中文論文期刊資訊檢索系統，以期望達到下列目的:利用部分節點由下而上搜尋幫助使用者快速且準確尋找資料、提供多元化的概念點陣瀏覽模式、期望學生與教授能透過本系統瀏覽及搜尋資管期刊，增進研究上的效率。

貳、　文獻探討

文獻探討主要分成資訊檢索瀏覽、資訊檢索搜尋、正規化概念分析、概念點陣瀏覽與搜尋等四部分。

一、　資訊檢索瀏覽

資訊檢索瀏覽方式主要分成三大類，分別為平面瀏覽、結構引導瀏覽及超文件瀏

覽，以下針對這三種瀏覽方式進行說明(Baeza-Yates and Ribeiro-Neto, 1999)。

(一) 平面瀏覽(Flat Browsing):是指使用者所瀏覽的文件是在單一(Single Dimention)或二維維度(Two Dimention)的平面上展示，使用者可以瀏覽目前文件的相關資訊。

(二) 結構引導瀏覽(Structure Guided Browsing):是指使用者瀏覽已分類好的架構項目中的文件資料。

(三) 超文件瀏覽(Hypertext Browsing):提供高互動性的引導式架構幫助使用者以非順序的方式來瀏覽文件。

二、 資訊檢索搜尋

資訊檢索搜尋主要分成三大類，分別為內文比對查詢、反轉檔搜尋及特徵搜尋。以下將針對這三種搜尋方法進行說明 (邱立豐，2002)。

(一) 內文比對查詢(text pattern search):又稱全文檢索(Full Text Scanning)主要是將使用者輸入的關鍵字與文章內容做比對，若有符合的結果便回饋給使用者進行瀏覽，內文比對查詢的優點是儲存容易，因為資訊不需經過任何處理直接儲存至資料庫，但在檢索時由於是全文比對，因此當資料量過於龐大時，執行效率將會大打折扣。

(二) 反轉檔搜尋(Inverted File Search):主要是先將文件資料進行斷字斷詞處理，並且記錄各個關鍵詞存在的文件，利用文件關鍵詞做為文件的索引值，當使用者輸入關鍵字搜尋時，搜尋引擎不需直接比對文件的全文內容，只要藉由索引值找到含有此字彙的文件。此方式可以解決全文檢索的缺點，但由於需建立索引表，相對會佔用較多資料儲存的空間。

(三) 特徵搜尋(Signature Search):主要是透過疊加編碼(Superimposed Coding)的方式來建立文件特徵。特徵搜尋會先過濾不符合使用者輸入的搜尋條件的文件，再針對初步過濾後的文件進行詳細比對。

三、 正規化概念分析

正規化概念分析是將物件(Objects)及物件所擁有的屬性(Attributes)組成正規本文(Formal Context)，正規本文建立後即可導出正規概念(Formal Concept)，最後再將正規概念轉換成階層式的概念結構，這種結構一般稱之概念點陣(Concept Lattice)。

當要展現圖形化的概念點陣時，必須決定每個概念的前任者(Predecessors)和後繼者(Successors)，前任者可以由每個涵義的最大子概念得到。而後繼者則可以由每個涵義的最小父概念得到。當概念點陣形成之後，最大的子概念會放在概念點陣的最上方，稱為上確界(Supremum)，而最小的子概念會放在最下方，稱為下確界(Infimum)。圖 1 為概念點陣的範例(Tam，2004)，每個節點在概念點陣上都代表一個概念，節點上方表示的為此概念的涵義，下方則是範圍。由圖 1 可以得知{Fish},{Fred,Jess,Bob,Mel}為上確界，而{ Fish,Beef,Pork,Chicken},{}為下確界。

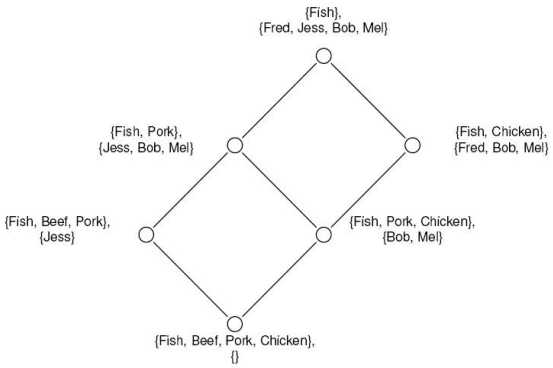


圖 1 概念點陣範例

資料來源：Tam(2004)

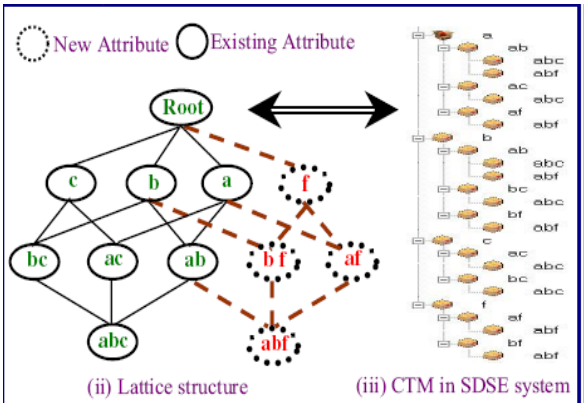


圖 2 概念點陣轉換成樹狀結構的方法

資料來源：Cho and Richards(2004)

目前有許多的正規化概念分析工具被開發出來，大部分都是採用 Java 所開發的，具有跨平台的特性。例如，ToscanaJ、Concept Explorer、及 Galicia 都是屬於此類的正規化概念分析工具。

四、 概念點陣瀏覽與搜尋

Carpineto and Romano(2004)的研究中提到利用正規化概念分析所產生的概念點陣可以有效的整合瀏覽與搜尋功能。

（一） 概念點陣瀏覽:概念點陣本身是網狀的資料結構所以可以很容易來瀏覽整個文件資料。不過一般使用者所熟悉的瀏覽架構是階層或樹狀的結構，因此有學者提出將概念點陣轉換成樹狀結構的方法(Cho and Richards，2004)。樹狀結構能讓使用者更容易的瀏覽、更快尋找到資料及可以透過不同的路徑尋找到相同的資料。概念點陣轉換成樹狀結構的方法如圖 2 所示。

（二） 概念點陣搜尋:一般在搜尋概念點陣上所採用的方法可分為概念點陣為基礎排名法、概念相似度計算及向量空間模型相似度計算等 3 大類。

1. 概念點陣為基礎排名法(Concept Lattice-Based Ranking, CLR)

概念點陣為基礎排名法是將查詢句與概念點陣的節點之間的距離當作排名依據。優點是可以解決關鍵字沒有匹配到的問題，而缺點則是當兩個文件計算出來的距離是相同時，就無法判斷出哪個文件是比較相關的。在概念點陣為基礎排名法文獻方面，Tam(2004)提出一種基於 TFIDF 概念的多值轉成單一值正規本文的方式，並且採用概念點陣為基礎排名法來跟以往的方法比較，實驗結果證實此方法可得到較高的精確率；Jun, et al.(2005)加入概念權重和使用者檔案在概念點陣為基礎排名法上，改良傳統概念點陣為基礎排名法的缺點，但在處理上比傳統的概念點陣為基礎排名法還要費時。

2. 概念相似度計算

概念相似度計算是利用概念點陣中概念的涵義或是範圍來計算概念之間的相似度。概念相似度計算已經有相關的學者提出不同的做法，例如 Formica(2006)提出基於本體論(Ontology Based)的相似度計算公式，其優點為可以應用在

本體論映對及本體論合併，此公式在計算時同時考量到兩概念的涵義及範圍；Zhao and Halang(2006)提出整合粗集理論(Rough Set Theory)和概念點陣理論(Concept Lattice Theory)的相似度計算公式，其特色為可以增進計算上的效能，此公式計算時只考量兩概念的涵義來進行計算；Wang and Liu(2008)提出基於粗集理論的相似度計算公式，其優點為可以方便計算大量的正規本文，此公式在計算時同時考量到兩概念的涵義及範圍。

3. 向量空間模型相似度計算

向量空間模型相似度計算是利用傳統資訊檢索領域的向量空間模型概念來計算概念之間的相似度。在向量空間模型相似度計算文獻方面，Zhang, et al.(2008)提出一個以正規化概念分析為基礎的兩階層式結構搜尋結果分群方法，此方法在比較概念相似度時採用 Jaccard coefficient 相似度計算公式，而在計算時是利用概念的範圍來進行計算。林群貿(2009)提出一個以正規化概念分析為基礎的本體論自動擴展機制，並藉由改良的 TFIDF 方法(Location Weight TFIDF, LTF-IDF)找出能夠代表學習元件的主要關鍵字，當作學習概念。實驗結果證實此機制能有效的擷取出學習概念，並能找出與學習概念關係相近的概念，自動地擴展本體論。

參、 研究方法

本研究以正規化概念分析的概念點陣為基礎，提出一套可以瀏覽和搜尋中文資料的系統，系統架構主要分成前置作業、關鍵字選取、概念點陣建置及瀏覽與搜尋等四大部份(圖 3)。

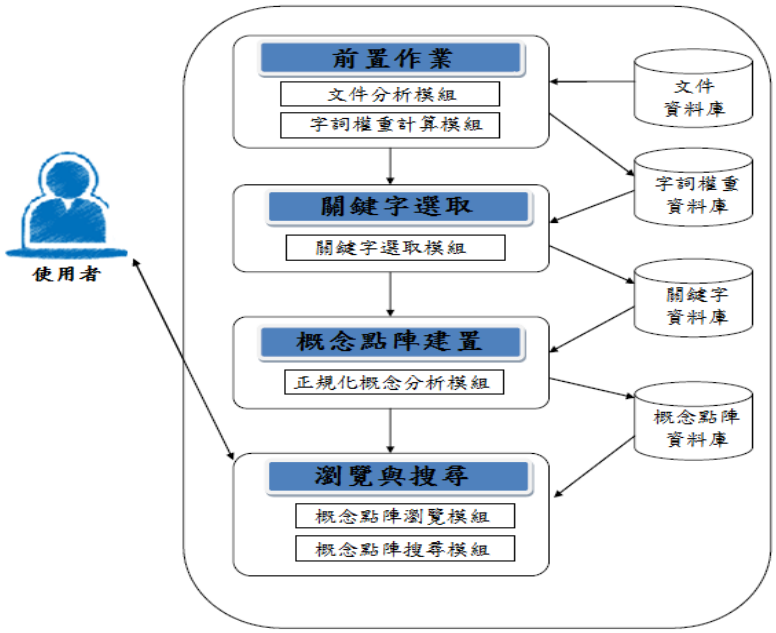


圖 3 系統架構圖

一、　前置作業

前置作業主要功能為將文件資料進行分析工作和計算關鍵字的權重。前置作業包含文件分析模組和字詞權重計算模組。

（一）文件分析模組

文件分析模組分成符號代換子模組、斷字斷詞處理子模組、詞性合併子模組、停用詞刪除子模組和同義字過濾子模組等五部分（圖 4）。

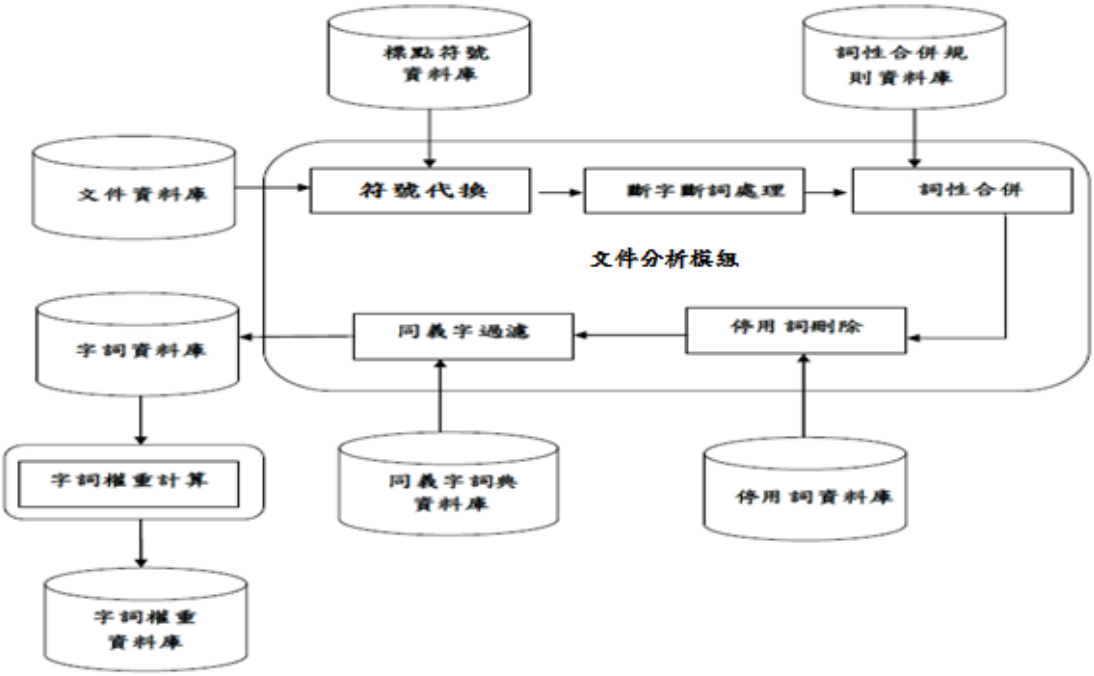


圖 4 文件分析模組

1. 符號代換子模組:主要功能為代換文件內容中 CKIP 無法判斷的符號，例如:“—”符號 CKIP 無法判別，所以事先代換成 “__”，這類的符號會影響 CKIP 進行斷字斷詞的判斷能力，因此必須先代換無法判斷的符號，以確保斷詞的品質。
2. 斷字斷詞處理子模組:主要功能為進行中文資料斷字斷詞處理。本研究選擇使用中央研究院詞庫小組所研發的中文自動斷詞系統(CKIP)來進行中文資料斷字斷詞處理，此斷詞系統可以將中文資料切割成有意義的字詞並且標註該字詞的詞性，利用字詞的詞性就可以進行後續的詞性合併及停用詞刪除等處理。
3. 詞性合併子模組:主要利用專有名詞合併規則及詞性合併規則來合併字詞。在此模組主要功能為將斷字斷詞後的一些詞性組合做合併，在經過斷字斷詞處理子模組處理後，一些專有名詞會被分開成多個字詞，而失去原本的涵義，例如「資訊管理系統」會被分開成「資訊(Na)、管理(VC)及系統(Na)」，本研究依據許正欣(2004)提出的詞性合併規則(表 1)以及參考相關資訊管理名詞彙編書籍如毛慶禎(1999)的計算機概論書目索引、余強(2007)的計算機概論名詞彙編、周宣光譯(2010)管理資訊系統和林東清(2010)資訊管理等等書籍資料整理

成資管專有名詞資料庫，將這些字詞進行詞性合併處理，降低斷字斷詞處理後所產生語意不符的問題。

表 1 詞性合併規則範例

組合規則	組合後詞性	組合範例
A + Na	A	信託(A) + 股票(Na)
Na + Na	Na	網路(Na) + 主機(Na)
Nb + Na + Nc	Nb + Na	世華(Nb) + 商業(Na) + 銀行(Nc)
Nc + Nc + Nc	Nc + Nc	中央研究院(Nc) + 語言所(Nc) + 語音實驗室(Nc)

資料來源：許正欣(2004)

4. 停用詞刪除子模組:此模組主要功能為刪除無意義的詞性及刪除有意義的詞性但無意義的字詞，主要目的為去除斷字斷詞處理子模組處理後無意義的詞性。例如，語助詞及感嘆詞，這些詞性在文章中大多是扮演修飾的用途。本研究依據江志銘(2005)所整理的停用詞(表 2)分類來刪除無意義的詞性。

表 1 詞性分類

分類	詞性	個數
名詞	Na, Nb, Nc, Ncd, Nd	5
動詞	VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL	15
量詞	Neu, Nes, Nep, Neqa, Neqb, Nf, Ng	7
副詞	Da, Dfa, Dfb, Di, Dk, D	6
停用詞	**A, Caa, Cab, Cba, Cbb, Nh, I, P, T, V_2, DE, SHI**	**12**
英文標記	FW	1

資料來源：江志銘(2005)

5. 同義字過濾子模組:此模組主要功能為利用同義字詞典將同義字進行過濾，同義字是不同的字詞可能代表相同的意義。例如，教授與老師，若未將這類的同義字進行過濾可能會造成同時存在意思相近關鍵字的問題。梁桂豪(2007)提到利用同義字的處理可以減少概念點陣內節點的數量，本研究在此階段參考相關資訊管理名詞彙編書籍且自建的同義字詞典將同義字進行過濾，避免後續產生概念點陣過於複雜的情形。

(二) 字詞權重計算模組

字詞權重計算模組主要功能為利用 TFIDF 公式計算關鍵字的權重，以便找出中文資料中較具有代表性的關鍵字。在經過同義字過濾子模組後，雖然可以得到中文資料的關鍵字，但是這些關鍵字在中文資料的重要性不一定相同 ，因此必須利用字詞權重計算模組計算各關鍵字的權重，以便找出中文資料中具有代表性的關鍵字，在此階段本研究利用 TFIDF 公式來計算關鍵字的權重。

二、 關鍵字選取

關鍵字選取包含關鍵字選取處理模組，此模組提供使用者利用「關鍵字數」及「字

詞長度」功能來選擇所要的關鍵字，主要功能為縮小代表文件所需的關鍵字範圍，並將處理結果儲存至關鍵字資料庫。利用字詞權重計算得到關鍵字權重高低來排名選取關鍵字，可以有效的選取代表文件的關鍵字，並且可以降低概念點陣的複雜度。

三、概念點陣建置

概念點陣建置主要有正規化概念分析模組，正規化概念分析模組主要功能為利用正規化概念分析演算法產生概念點陣，並將產生後的概念點陣儲存至概念點陣資料庫，以便瀏覽與搜尋使用(圖 5)。正規化概念分析模組可分成建立正規本文子模組、正規化概念分析演算法子模組及概念點陣格式轉換子模組等三部分。

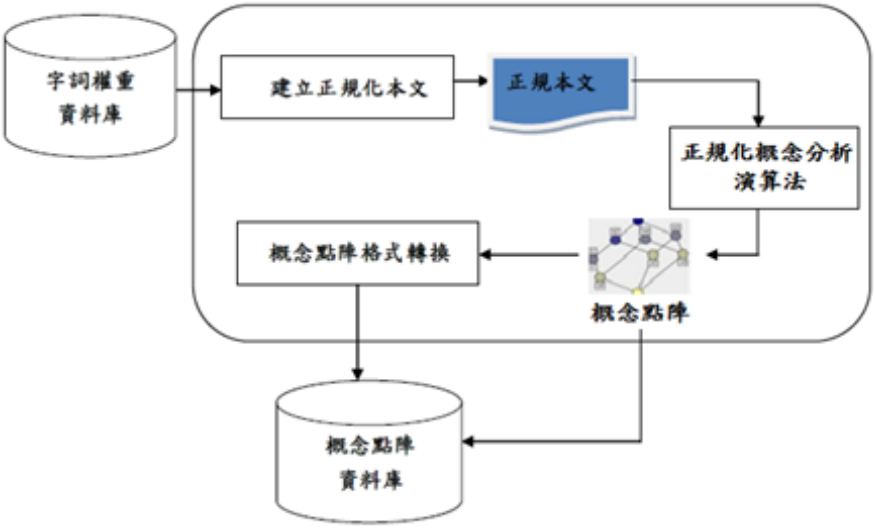


圖 1 正規化概念分析模組

(一)建立正規本文子模組

建立正規本文子模組主要功能為將文件與關鍵字組成正規本文，建立正規本文後就可以利用正規化概念分析演算法子模組產生概念點陣。本研究將傳統正規本文的物件以「文件」來取代，而屬性則由「關鍵字」來取代，最後將文件與關鍵字組成正規本文。

(二)正規化概念分析演算法子模組

正規化概念分析演算法子模組的主要功能為將已建好的正規本文利用正規化概念分析演算法產生概念點陣。本研究使用的正規化概念分析工具為 Galicia ，Galicia 為 Java 所開發的開放原始碼工具，並且提供多種建立概念點陣的演算法可供使用。

(三)概念點陣格式轉換子模組

概念點陣格式轉換子模組主要功能為將已產生的概念點陣 XML 格式轉換成概念點陣資料庫格式，並且將轉換後的概念點陣資料庫格式資料儲存至概念點陣資料庫，以便瀏覽與搜尋使用(圖 6)。概念點陣資料庫格式包含概念點陣各個節點的 ID、範圍、涵義、父節點及子節點。

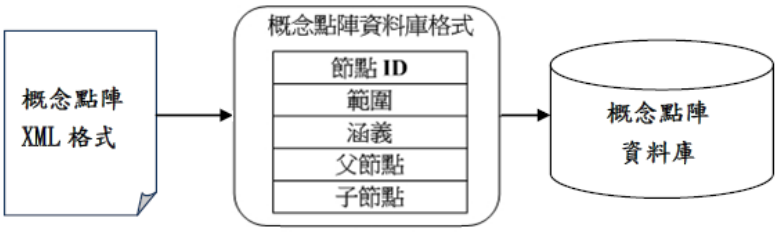


圖 2 概念點陣格式轉換子模組

四、瀏覽與搜尋

瀏覽與搜尋主要功能為提供瀏覽和搜尋概念點陣的功能。瀏覽與搜尋分成概念點陣瀏覽模組及概念點陣搜尋模組。

(一)　概念點陣瀏覽模組

此模組主要功能為將產生後的概念點陣提供給使用者瀏覽，本研究提供網狀及樹狀兩種概念點陣瀏覽模式。

1. 網狀概念點陣瀏覽模式:網狀概念點陣具有圖形化展現資料的優點，使用者可以更容易讀取與理解資料的內容，但缺點是網狀結構與一般使用者所熟悉的階層或樹狀結構的瀏覽模式不同。
2. 樹狀概念點陣瀏覽模式:樹狀概念點陣的優點是與使用者以往所熟悉的瀏覽模式相同，而缺點則是無法像網狀概念點陣具有較佳的視覺化效果。

(二) 概念點陣搜尋模組

概念點陣搜尋模組主要功能為將使用者輸入的關鍵字與已產生的概念點陣進行比對，然後回傳符合的結果給使用者。概念點陣的搜尋方式可細分成有兩種(1)由上而下(上確界搜尋到下確界)全部節點搜尋、(2)由下而上(下確界搜尋到上確界)部份節點搜尋。對於搜尋結果的展現，本研究提供樹狀結構及相似度排序兩種展現方式。

1.概念點陣搜尋

(1) 概念相似度:概念相似度是計算關鍵字與概念點陣的概念(節點)間的相似程度。概念相似度主要利用 Zhao and Halang(2006)所提出的概念相似度計算方法來計算(公式 1)。

$$Sim(a,\ b) = \frac{|(a \cup b)_{LA}|}{|(a \cup b)_{LA}| + \alpha |a_{LA} - b_{LA}| + (1-\alpha)|b_{LA} - a_{LA}|} \quad (1)$$

$$\alpha(a,\ b) = \begin{cases} \dfrac{a_{LA}}{a_{LA} + b_{LA}} & \text{if } a_{LA} \le b_{LA} \\ 1 - \dfrac{a_{LA}}{a_{LA} + b_{LA}} & \text{if } a_{LA} > b_{LA} \end{cases}$$

Sim(a,b)為 a、b 兩個概念(節點)的相似度，a_{LA} 為 a 節點涵義的屬性數量， b_{LA} 為 b 節點涵義的屬性數量，$(a \cup b)_{LA}$ 為同時被 a、b 兩節點涵義所擁有的屬性數量，a_{LA}-b_{LA} 為 a 節點涵義的屬性數量減掉同時屬於 b 節點涵義的屬性數量， b_{LA}-a_{LA} 為 b 節點涵義的屬性數量減掉同時屬於 a 節點涵義的屬性數量，α 值則會隨著 a_{LA} 及 b_{LA} 的大小而有所不同。

(2)文件相似度:文件相似度是計算關鍵字與文件間的相似程度。文件相似度(DocSim)之計算公式如(2)所示。

$$\text{DocSim(Q, D)} = \sqrt{\text{Max}\left(Sim\left(Q, N\right)\right) \times Sim\left(Q, D\right)} \quad (2)$$

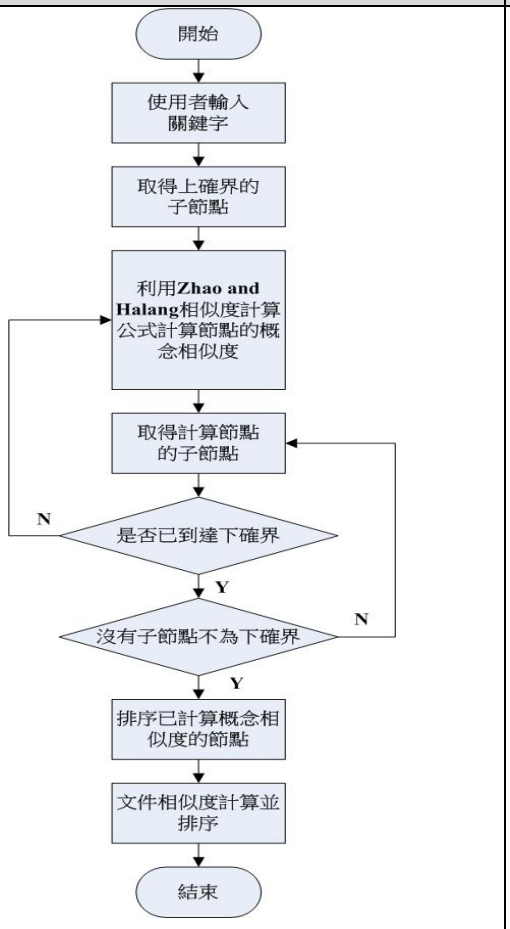
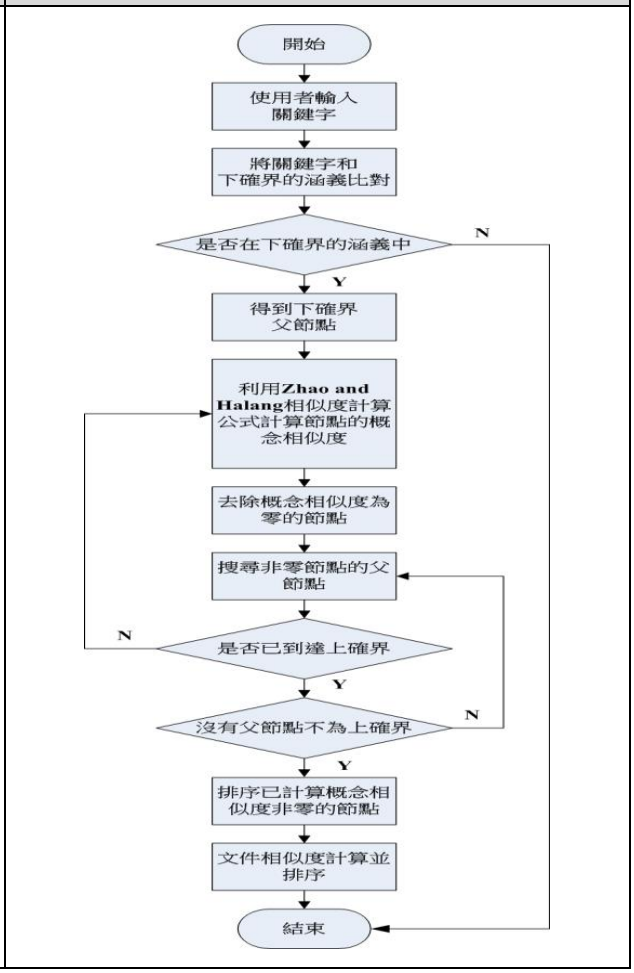
DocSim(Q, D)為關鍵字 Q 與文件 D 間的文件相似度，Max 取最大值，Sim(Q,N)為關鍵字 Q 與概念 N 間的概念相似度，Sim(Q,D)為關鍵字 Q 與文件 D 間的概念相似度。

本研究的文件相似度計算方法是當文件有多個概念相似度(文件在多個概念(節點)同時存在)時取最大值來當成概念相似度 Max((Sim(Q,N))，接著利用概念相似度計算方法來計算關鍵字與文件的關鍵字間的概念相似度(Sim(Q,D))，並將這兩項的概念相似度相乘開根號得到文件相似度 DocSim(Q, D)。

(3) 概念點陣搜尋方式

概念點陣的搜尋有由上到下全部節點搜尋和由下到上部份節點搜尋等兩種搜尋方式(表 3)。

表 3 概念點陣搜尋方式

由上到下全部節點搜尋	由下到上部份節點搜尋
	

資料來源:本研究整理

(4)搜尋結果展現

為了展現搜尋結果，本研究提供樹狀結構及相似度排序兩種展現方式。

a. 樹狀結構展現方式

樹狀結構展現方式主要是利用樹狀概念點陣來展現搜尋的結果。在概念點陣的搜尋過程中會利用概念相似度找出最相似的概念節點，然後在樹狀概念點陣中標記出，使用者可以由最相似的概念開始瀏覽。

b. 相似度排序展現方式

相似度排序展現方式主要是利用文件相似度排序來展現搜尋的結果。在概念點陣的搜尋過程中會找出相似度較高的文件，然後依相似度高低排序文件，並列出排名及相似度，使用者可以點選文件直接瀏覽。

肆、系統實作與評估

一、系統實作環境

本研究之支援系統是在 Microsoft Windows 作業系統上開發，無論是 Windows 系列或是 Windows Server 系列皆可以安裝。支援系統主要使用 Tomcat 當作 JSP 容器(JSP Container)，並搭配 Java、JSP、JavaScript 及 Ajax 技術與 Microsoft Office Access 資料庫來實作；概念點陣則是以正規化概念分析工具 Galicia 所提供的演算法來產生。

二、實驗資料

中文資料種類有很多種，本研究以台灣中文資管期刊做為研究對象，採用中文期刊電子服務(簡稱 CEPS 系統)中社會科學類的資訊管理學報為實驗資料，收集從 2000 年 7 月至 2010 年 10 月年間共計 11 年學報，只取出中文的期刊，共計 345 篇論文。論文資料以標頭、摘要及論文作者所設的關鍵字分成共三個資料庫，論文資料在經過文件分析模組後會產生標頭、摘要及關鍵字等 3 個字詞權重資料庫，經概念點陣建置後會產生標頭、摘要及關鍵字等 3 個概念點陣。

三、評估準則

本研究為了評估檢索方法的檢索品質及執行效能，採用精確率(Precision)、召回率(Recall) 、F指標(F-measure)、節點掃描率、文件掃描率及執行時間等6項作為評估準則。公式(3)~公式(7)分別是精確率(Precision)、召回率(Recall) 、F指標(F-measure)、節點掃描率及文件掃描率的公式。

$$\text{Precision}=\frac{\text{系統檢出相關文件之數目}}{\text{系統檢出之文件數目}} \quad (3)$$

$$\text{Recall}=\frac{\text{系統檢出相關文件之數目}}{\text{資料庫中相關文件之總數目}} \quad (4)$$

$$\text{F - measure} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2}{\frac{\text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{節點掃描率} = \frac{\text{比對節點數目}}{\text{概念點陣全部節點數目}} \quad (6)$$

$$\text{文件掃描率} = \frac{\text{比對文件數目}}{\text{資料庫全部文件數目}} \quad (7)$$

四、實驗評估

本研究實驗評估分成二大部分，第一部分為概念點陣搜尋方式對效能的影響，第二部分為概念點陣個數與可使用關鍵字數的關係。

（一） 概念點陣搜尋方式對效能的影響

[實驗目的]

概念點陣由上而下全部節點搜尋和由下而上部份節點搜尋對效能的影響。

[實驗方法]

1. 為了評估概念點陣搜尋方法對效能的影響，以概念點陣由上而下全部節點搜尋和由下而上部份節點搜尋兩種檢索方法進行搜尋，採用十次實驗結果平均值進行評估。評估方式是以關鍵字檢索結果為標準，來進行精確率、召回率、F 指標、節點掃描率、文件掃描率及執行時間等 6 項指標的評估。
2. 本次實驗是將標頭、關鍵字及摘要等 3 個概念點陣分別做兩種搜尋實驗。

[實驗結果]

概念點陣搜尋實驗結果如表 4 所示，由下而上部份節點搜尋與由上而下全部節點搜尋的精確率、召回率及 F 指標皆為 100%；部份節點搜尋的節點掃描率比全部節點搜尋少了約 20%~29%倍、文件掃描率則兩者皆相同；部份節點搜尋的執行時間約為全部節點搜尋的 56%~71%。

（二） 概念點陣個數與可使用關鍵字數的關係

概念點陣個數與可使用關鍵字數的關係如表5所示，使用單1個點陣概念時可以使用的關鍵字數平均為1149個；使用任2個點陣概念時可以使用的關鍵字數平均為1866個；同時使用3個點陣概念時可以使用的關鍵字數為2435個。同時使用2個點陣概念時可以使用的關鍵字是單1個點陣概念的1.62倍；同時使用3個點陣概念時可以使用的關鍵字是單1個點陣概念的2.12倍。

表 4 概念點陣搜尋實驗結果

評估準則 / 檢索方式	精確率	召回率	F 指標	節點掃描率	文件掃描率	時間平均
標頭概念點陣						
全部節點搜尋	100%	100%	100%	100%	9.75%	0.51 秒
部份節點搜尋	100%	100%	100%	79.63%	9.75%	0.36 秒
關鍵字概念點陣						
全部節點搜尋	100%	100%	100%	100%	9.63%	0.55 秒
部份節點搜尋	100%	100%	100%	71.46%	9.63%	0.37 秒
摘要概念點陣						
全部節點搜尋	100%	100%	100%	100%	8.78%	0.55 秒
部份節點搜尋	100%	100%	100%	71.96%	8.78%	0.30 秒

表 5 概念點陣個數與可使用關鍵字數的關係

單 1 個概念點陣	字數	2 個概念點陣	字數	重複字數	3 個概念點陣	字數	重複字數
標頭	852	標頭、關鍵字	1526	398	標頭、摘要、關鍵字	2435	1012
關鍵字	1072	標頭、摘要	1972	403			
摘要	1523	摘要、關鍵字	2100	495			
平均	1149	平均	1866				

伍、結論

資訊檢索系統主要提供使用者搜尋與瀏覽兩大功能，目前大部分的資訊檢索系統主要提供搜尋的功能，有提供瀏覽功能的並不多見。概念點陣可以有效的整合瀏覽與搜尋功能，本研究以正規化概念分析的概念點陣為基礎提出一套中文資料的瀏覽與搜尋系統，此系統在瀏覽上提供網狀及樹狀的概念點陣瀏覽模式，在搜尋上採用概念點陣由下而上部分節點搜尋方式僅需搜尋部分節點，並且以樹狀與網狀結構來展現搜尋結果。

為了評估本系統的檢索品質及執行效能，本研究以資訊管理學報作為實驗資料。實驗結果顯示，在同樣的搜尋品質的情況下由下而上部份節點搜尋的執行時間約為由上而下全部節點搜尋的 56%~71%；同時使用 2 個或 3 個點陣概念時可以使用的關鍵字是使用單 1 個點陣概念的 1.62 倍或 2.12 倍。

在未來的研究中，首先將實作瀏覽與搜尋模組讓使用者可以透過網頁瀏覽器來使用本研究所開發的支援系統，接著將本研究方法實際應用到其他領域的中文資料資訊檢索。

參考文獻

1. 江志銘，2005，應用問答系統技術於電腦領域論壇檢索之研究，國立雲林科技大學，碩士論文。
2. 邱立豐，2002，互動式概念查詢應用於網路文件自動摘要之效益，國立雲林科技大學，碩士論文。
3. 許正欣，2004，語意網上自動化建構本體論之研究，輔仁大學，碩士論文。
4. 梁桂豪，2007，正規概念分析法為基礎之動態知識管理，國立雲林科技大學，碩士論文。
5. 林群貿，2009，以正規概念分析為基礎之本體論自動擴展機制，國立成功大學，碩士論文。
6. Baeza-Yates, R. and Ribeiro-Neto, B., 1999, Modern information retrieval, Addison-Wesley Publishing Company.
7. Cho, W. C., and Richards, D., 2004, "Improvement of precision and recall for information retrieval in a narrow domain: Reuse of concepts by Formal Concept Analysis", IEEE/WIC/ACM International Conference on Web Intelligence, pp.370-376.
8. Formica, A., 2006, "Ontology-based concept similarity in Formal Concept Analysis", Information Sciences, vol.176, pp.2624-2641.
9. Ganter, B., and Wille, R., 1998, General Lattice Theory(2nd edition), Birkhauser Verlag.
10. Jun, T., Du, Y., and Shen, J., 2005, "Research in concept lattice based automatic document ranking", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, pp.18-21.
11. Ma Jun, 2005, Research on concept lattice and its Visualization, HENAN University, Master's degree paper.
12. Priss, U., 2000, "Lattice-based Information Retrieval", Knowledge Organization, vol.27(3), pp.132-142.
13. Salton, G., 1988, Automatic Text Processing, Addison-Wesley Publishing Company.
14. Tam, G. K. T., 2004, FOCAS - Formal Concept Analysis and Text Similarity, Monash University, Honours thesis.
15. Wang, L., and Liu, X., 2008, "A new model of evaluating concept similarity", Knowledge-Based Systems, vol.21(8), pp.842-846.
16. Wille, R., and Ganter, B., 1999, Formal Concept Analysis : Mathematical Foundations, Springer.
17. Zhang, Y., Feng, B., and Xue, Y., 2008, "A New Search Results Clustering Algorithm based on Formal Concept Analysis", 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, pp.356-360.
18. Zhao, Y., and Halang, W., 2006, "Rough concept lattice based ontology similarity measure", Proceedings of the 1st international conference on Scalable information systems.

Using Formal Concept Analysis to Construct Information Retrieval of Chinese Data
- Data Set from Journal of Information Management

Huang Jin Fa
Department of Information Management
National Yunlin University of Science & Technology
huangcf@mis4k.mis.yuntech.edu.tw

Wang Yao De
Department of Information Management
National Yunlin University of Science & Technology
a123863902@gmail.com

Abstract

During recent decades, the explosive development of the information technology has brought a remarkable advance in search engine and it has also changed our daily life. Nowadays, most information retrieval (IR) systems can only help us search information with simple interface, but the IR system that can help integrate web browsing and searching functions is very rare. Fortunately, Concept Lattice can help address the shortage and provide the browse of web tree structure.

This paper uses Formal Concept Analysis (FCA) to construct Chinese Information Retrieval, and provides a browsing and searching system in Chinese data. This system can display multi-browsing modes for Concept Lattice, and we use both bottom-up (from the Infimum to the Supremum) and top-down (from the Supremum to the Infimum) searching way to help users find information efficiently and correctly. We also use tree structure and similarity order to present our result. To evaluate the searching quality and performance, this paper uses the data form Journal of Information Management as our dataset.

Keywords: Information Retrieval, Formal Concept Analysis, Concept Lattice, Journal of Information Management.