摘要

隨著電子化的趨勢，各式各樣的電子資料一直在網路上迅速的繁衍與增長。雖然已有相當多研究投入文件資料搜尋，但目前似乎面臨成果不佳以及發展膠著的窘境。因此本研究致力於發展正確性、實用性及方便性的演算步驟以提昇檢索成效。

在眾多的相關研究議題中，群聚經常被利用於大型資料的歸類，但在搜尋引擎的檢索中群聚研究並未多見，故本論文將以此方向做為研究重心。我們利用關聯規則（Association Rule）挖掘共通屬性之間的關係，再藉由圖解模型（Graph Model）的架構來闡述屬性之間的關聯及強弱程度，最後透過切割圖形達到分群的效果。本作法所呈現的族群屬性與量測結果，不同於一般以核心距離偵測或向量比對的做法。我們獲得的群聚反應具有精確、易於理解及快速執行等優勢。

關鍵字：相對資訊、關聯規則、圖形切割、多重主體屬性

Abstract

With the electrifying of the Internet, all types of electronic information have been rapidly growing and increasing. Although massive amount of research has been dedicated to information searching, it seems that we are faced with the awkward situation of barely permissible results and deadlocked progress, and we are hoping to advance search utilities with higher accuracy, better practicality, and greater convenience.

In multitudinous research topics, clustering is often used to classify large-scale information, and has had exceptional results, but clustering is rarely used in search engines; therefore, this paper will discuss utilizing cluster in search technology. We apply the Association Rule to pull closer the relation between common attributes, and then use a Graph Model structure to elaborate on the association and strength of each attribute, and lastly we adopt graph segmentation to achieve classification. The category attributes and test results displayed by this approach are different from those achieved via normal distance detection, and the obtained results have advantages such as higher precision, easier understandability, and faster execution.

Keywords：Mutual Information、Association Rule、Graph Partition、Multi-Objectives

致謝

首先最感謝的是我的指導教授許玟斌老師，在我碩士班的修業期間，除了給我論文研究方面的邏輯啟發以及研究精神的指導，在表達能力、實事求是更是我所學習到最寶貴的資產，而鄭國揚老師、陳文雄老師、徐麗蘋老師、胡學誠老師，在論文口試時，也給予我許多寶貴的意見，使我的論文更完整。

其次要感謝的是所上的同學們，包括義樺、俊維、清健、真真、明富、鎮南以及其他一起努力的夥伴，在他們身上，我學習到許多的想法及目標，更感受來自他們的真摯友誼；藉由相互的鼓勵，使我能更積極及堅持的面對漫長的研究生活。

最後，我想要感謝我的家人、怡君及尚琦，是在背後默默的支持及容忍，我才能順利的完成研究所的學業，迎向下一階段的任務。

僅以此表達最深的感謝

陳仁傑 於　東海大學

2008/7/3

目錄

圖目錄

第 1 章　緒論

1.1 簡介

資料挖掘技術的目的，在於從大量的資料中透過自動化處理，挖掘出隱藏、可能、甚至是預測的資訊。目前在商業上或科學界，經常可以看到資料挖礦的技術被利用於解析或決策。

資料搜尋是挖掘技術中較常被提及的應用，其主要是探討特定條件下的母體（Population）鑑定。以文章搜尋為例，可能的應用是解析使用者付予的條件，及每一文章中回覆的反應。部份研究運用詞意聯想【1】、語意【2】或向量【6】等方法來獲取。然而特定條件在不同母體所扮演的角色往往不是必然或是重心，導致在現實的多樣性資料下，精確度面臨考驗。

以搜尋「蘋果」一詞為例，查詢結果可能是水果的「蘋果」或「蘋果日報」或「蘋果電腦」等主題的文章，這些文章的意涵應是截然不同的。因此藉由關鍵字進行搜尋時，往往發現不同主旨的文章交錯呈現。

搜尋引擎一直是面對沒有規範的大型資料庫。浩瀚的網際網路，有太多的議題、不同的看法或跨越不同的領域。而群集的效益經常被利用於發散資料，透過群聚效果，我們可以適度的將各主題文章進行比較正確的區隔。

我們試圖增加群聚的功能於文章檢索的前置階段，使其具有：

1. 快速的攝取閱讀：當使用者面對廣大的文件集合時，可隨著個人的特定條件及群聚數找尋適當文章，以利在最短的時間內有效的篩選文章；請參閱圖 1.1 虛線右側圖形。
2. 簡易的瀏覽涉略：目前搜尋引擎多以關鍵詞在每篇文章中得到的反應；但本研究認為，若能利用群聚數，應可輕易的擴張搜尋樣本。以圖 1.1 為例，雖對文件集區分成 3 群，且 a 群有 3 篇文章，此時，若藉由縮小群數為 2 群聚，將可達到廣結文章的效果。

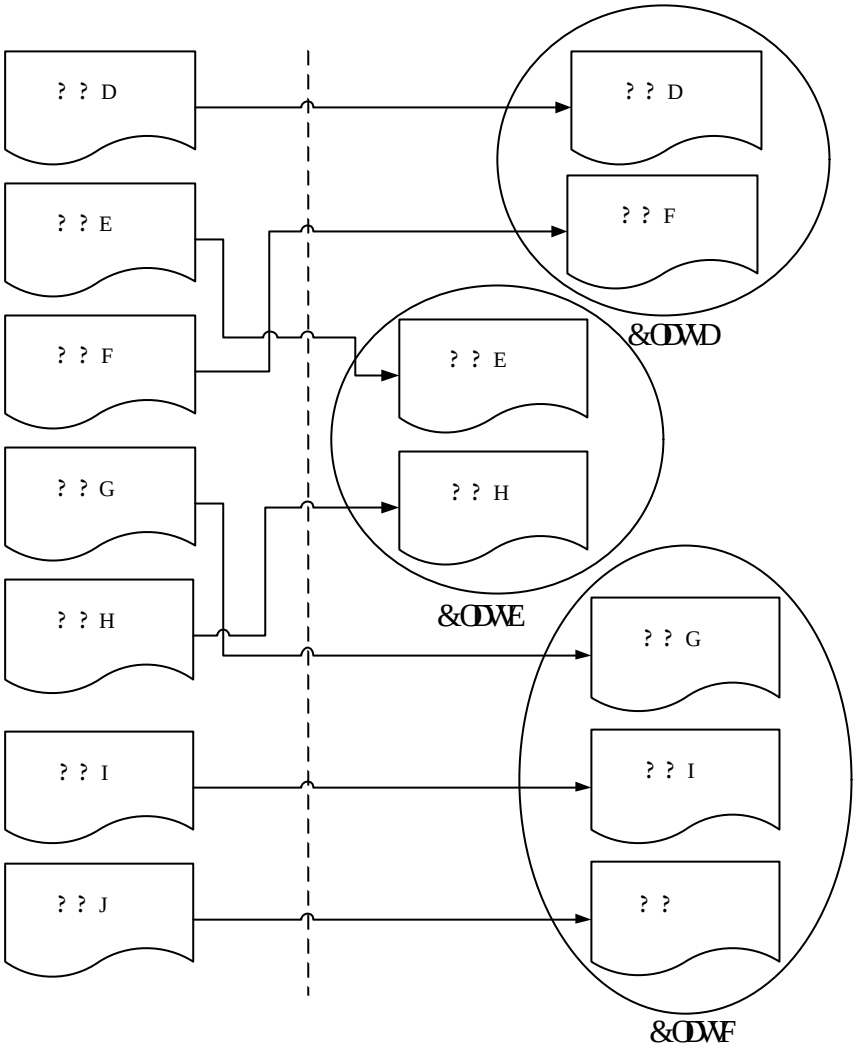
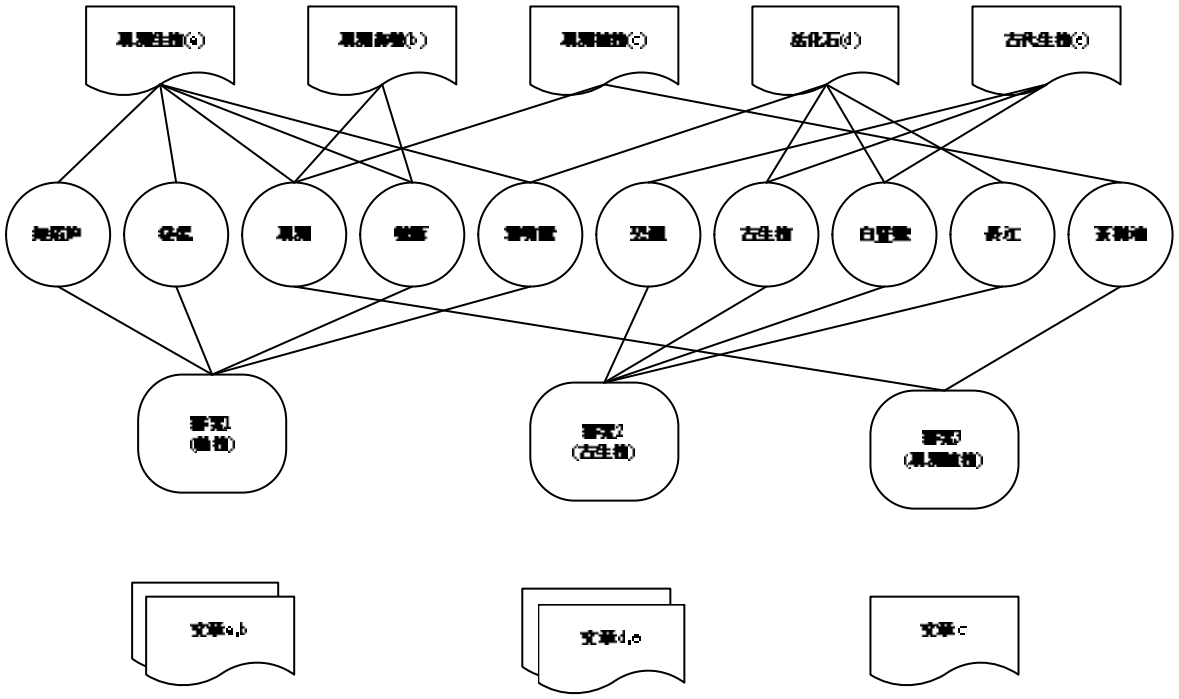


圖 1.1 群聚排序示意圖

本研究屬於硬式群聚（Hard Cluster**）不同於軟式群聚（**Soft Cluste**ring)。軟式群聚允許辨識元素可以被重複放置在多個群聚；以圖** 1.2 **為例，「澳洲」可以重複存在於群聚** 1 **及群聚** 3**，「鴉嘴獸」可以重覆於群聚** 1 **及群聚** 2**。相似基礎的軟式群聚（**Similarity-based Soft Cluste**ring**，SISC）【7，13】便是軟式群聚的實作研究。軟式群聚對群聚目的而言是客觀且可行的作法，但其若使用在文章檢索，該結果可能面臨使用者有失焦的風險。**

我們結合關聯法則及 Karypis **【**11**】的切割方法，形成各種概念群聚。我們認為透過權重切割後的硬式群聚，可幫助圖** 1.2 **認定「澳洲」詞彙對「動物」群聚不算關鍵詞，而該詞彙針對「澳洲植物」的主題卻極為重要。故客觀的關聯法則及合理的切割方法進行硬式群聚。**

我們試圖將概念群聚視為多重主體的觀念，且延伸到文章群聚。本研究兼具軟、硬式群聚的優點，企圖改善閱覽的行為模式，這將有助於提升點閱的正確性、閱覽品質，甚至提升學習成果等效益。



本論文共分為六章，除了第 1 章外，其餘各章節概要如下所述：

第 2 章　相關研究　介紹和本研究相關的資訊檢索技術。

第 3 章　知識版圖的研究與利用　本章主要是提出知識版圖的構想及建置的意義，透過切割原理解析詞彙之間的互動。

第 4 章　處理方法　為達成群聚目的，本章主要是介紹群聚處理的整個程序，包含文章處理、知識版圖的建置、三種不同切割的方法及兩種目的的計分法。

第 5 章　實驗訴求與分析　實作以上各章提出的方法，並分析及討論實驗的結果。

第 6 章　結論與未來展望　說明本論文的研究結論與後續研究發展的方向。

第 2 章　相關研究

本章主要是介紹文章分類程序會被引用到的相關研究。包含詞彙處理、詞彙萃取及超圖解模型的切割等相關研究。

2.1 中文斷詞

中文因字型結構特殊，文字可獨立或合併的運用，故句子的組成單元有字、有詞；例如，「樹」「樹木」「樹林」「林」。面對中文句子撰述時的字字相連，我們需要斷詞工具的輔助。

中文斷詞，通常以詞庫斷詞及統計斷詞最為普遍。詞庫斷詞主要利用現有詞庫，比對輸入的文件，擷取出文件中出現在詞庫的字詞。此種作法配合現成詞庫操作，在技巧上並不困難，並可依據既有詞性做為關鍵詞篩選參考，然而因受限於詞庫收錄，導致複合詞彙、專有名詞、新生詞彙並無法完善辨識。此外，在長詞優先擷取的考量下，如果長詞中包含短詞，則短詞並無法斷出。

至於統計式斷詞，需透過大量文件或語料庫(corpus)的訓練，取得足夠的統計參數，像是詞頻或門檻值等，再予以擷取滿足參數的語彙。統計式斷詞法適合大量資料的處理，而處理方式主要是依據機率統計值來決定斷詞的位置，此種作法優點是不需學科專家定義詞彙，其處理時間相對十分快速，也可有效解決詞庫在擷取複合名詞、專有名詞及新生詞彙的問題。然而所處理之語料庫如不足或同質性過高，將影響斷詞之涵蓋面，也極易產生大批無意義的字詞，而影響斷詞之可用性。

由於統計演算法並沒有考慮語法、語意的正確性，雖容易製造斷詞；但仍礙於同形異義無法判別，而使得整個句子可能表達錯誤的意思。此外，也可能有部分關鍵詞因統計參數不足而無法被選錄。例如，「彩雷」與「彩色雷射印表機」、「網芳」與「網路芳鄰」可能面臨無法被統計方法所辨識。

2.2 斷詞工具

中文文章的語句在自然語言處理（Natural Language Processing）上，必須藉由斷詞工具的處理較能從事資訊擷取、文件分類或文章摘要等應用。斷詞系統的標準於西元 1999 年由中華民國計算語言學研討會所制定，並以 CNS14366 正式通過國家標準。斷詞工具主要是由中央研究院中文詞庫小組【3】所開發，截至 2005 年 5 月已發展出「CKIP 自動斷詞系統」與「領域詞典工具」兩種版本：

（1）CKIP 自動斷詞系統：一般又被稱為 AutoTag。由中文知識資訊處理小組 (Chinese Knowledge Information Processing Group, CKIP)於 1991 年 11 月開發完成，版本為 1.0 版，其包含自動斷詞與詞性標記兩種功能，使用約十萬目個詞典當核心詞典，採用純文字檔案格式做為文章輸入格式。

（2）領域詞典工具：此工具是繼「CKIP 自動斷詞系統」所開發完成，包含了自動斷詞、詞性標注、偵測未知詞與自訂領域詞典四大功能。其功能不同於前者，主要是用來擷取出文件中的新詞或未知詞，並依原句型結構標示該新詞的詞性。當十萬目的核心詞彙遇到無法辨識部分時，其整個結果連同癥結詞彙尚可於該未知詞編輯器內透過使用者做人工的確認或增刪修並儲存之，例如像是「IC 設計」、「綠能」等詞彙就不會出現在傳統詞庫而缺乏有效詞【4】。

藉由以上的說明得知，後者使用之詞典工具較為豐富，功能也較為強大，故我們採用領域詞典工具做為文章的斷詞工具，以力求斷詞結果達成正確性的標準。

2.3 詞彙量化

在文章探索的過程中，我們不難發現詞彙的出現頻率對文章極具影響力，故詞彙的量化極具重要性。

文章的基本組成單位是詞彙，故透過詞頻(Term Frequency, TF)來解析。詞頻的觀念起源於 Luhn【14】，該實驗發現除卻高頻與低頻者，所留下的中頻（middle-frequency）字詞，多半是比較有意義的詞彙，因而提出「關鍵字詞適度詞頻」(resolving power of significant words)的理論；且引發日後諸多學者投入自動文件處理的興趣及研究。

部份的中文檢索研究，改良 Luhn 的作法【12】，並以下述方式來處理：

1. 刪除低頻詞彙、標點符號及無意義的一元詞：中文由於文字本身可獨立表意像是「樹」及「樹木」、「雨」及「下雨」可能同義；但經由母體研究發現，正式書類習慣多以二元詞以上的詞彙表述，故多忽略一元詞。此外「及」「故」「且」等詞彙亦因不具影響，故常予以忽略。
2. 刪除高頻詞彙或無意義的詞彙：如「因為」、「所以」、「由於」等詞彙。
3. 挖掘剩餘詞彙：部分研究以交互頻率（TFIDF）【12】選擇每篇文章特定程度以上的權重詞彙做為重心【7，13】。交互頻率主要是在描述被使用的詞彙在本身的文章與其他文章被引用的次數，以做為詞彙頻繁程度的量化參考。

部份研究針對詞彙量化，有人另外採用加權的作法。其考量詞彙萃取的位置如文章標題、摘要、字體，或者針對長詞給予加重權重。

長詞另計加權是中文斷詞的特色。通常中文斷詞擷取以 2 元詞到 9 元詞彙為主，基於字數越長的詞出現的機率越少，實質代表的意義卻越重要，因此對於字詞長度均以加乘本身字數的方式，進行加權，如：「知識」出現 10 次，轉換後〔10 次*2 元詞=10 次〕，「知識管理」原出現 5 次，轉換後〔5 次*4 元詞=20 次〕，藉由加權方式，以提升長詞的詞頻權重。

2.4 群聚偵測

以行銷學的詞彙來說，根據已知的有效鑑定項目，將群體細分的作法，稱之為「區隔」。但在許多案例中，雖然我們可能會懷疑一組非常凌亂的資料，是由一些更能表現出特性的群集所構成，但我們卻不知如何將其定義。我們稱此為群聚偵測（cluster detection）。

群聚偵測經常被運用於四處分散的資料特性。在商業應用方面，其經常被使用於發現不同客戶群的特色，並且依此提供客制化的服務，或者按客戶的基本資料來預測採購的行為。其實這方面的應用在購物欄分析已被探討，典型的商店銷售數以千計不同的產品給數以千計不同的客戶，如果我們可以掌握各群聚的自身的產品項目，我們就可以利用這方面的資訊，有效的在架上空間進行目標商品的促銷。

群聚研究認為，要被衡量的每一件事情都是獨立的，以便將其描述為「維度」。如果現在有 N 個變數，我們會想像一個 N 度空間，其中每一個變數的值代表其相對於軸心的距離。一筆單一資料包含一個數值，其裡頭包含 N 個變數中的每一個變數，其可以被想像成一個向量，做為定義空間中的一個特點。

如何切割重疊、釐清可能性以視為成群聚？部份的研究利用距離或向量的方法，測量相似度【6,9,15】做為測量標準。其應用經常使用 K 平均演算法（K-means algorithm）或質心向量演算法（Centroid-Vector algorithm），這種方法的第一步是選擇需要的群聚數目 K，接著選擇 K 的「種子」（Seeds）做為群聚質心的初步臆測。每一個種子只是每一項測量數值的特定集合。每一種子可能是紀錄樣本中一個真正的主體，但也可能不是，因為這並非必要。

下一步，資料庫中的每一筆紀錄基於其最接近的種子，給予一個初步的群聚分配。接著計算新群聚的質心，然後以此新質心為基準重複上述的程序，由於新質心不會和原來的種子在同一位置，某些紀錄會從第一個群集中被移除，被分配到另一群集。經過多次重複操作，這項動作停止。每一群聚的質心就包含各項測量數據，定義出新標準後的結論。

2.5 關聯性的測量

要測量兩筆文章間的關連性，可以概分為三種，分別是兩點之間的距離、兩個向量之間的角度及共同特徵的數目。

1. 兩點的距離：測量距離的方法，最常見的就是歐幾里得距離（Euclidian Distance），我們首先要找出 X 與 Y 之中對應單元的差並加以平方，所有對應單元差平方的總和再求其平方根。

2. 兩個向量的角度：所謂的向量，就是連結我們座標系統原點到向量值所表示的那一點線段。一個向量包含有大小與方向，其中前者表示由原點到該點的

距離，但比較時通常較為重要方向。兩個向量的比較採正弦值 0~1 做為衡量指標，也就是從完全相同到完全不同的差異來表示。

3. **特徵的一致性：我們將資料的特徵透過項目筆對的方式，計算兩者相符的程度，該方法具有立意明顯的優勢。例如，沙丁魚不會飛、小貓也不會飛，在「不會飛」的特徵下，「沙丁魚」「小貓」是具有一致性的特徵群聚，這種測量方法較向量法或計分法來的明確且易懂。**

上述三種作法前兩種適合用在區間變數與實際測量，也就是母體之間差異不大的範疇，第三種則適合類別變數的測量。文章之間的接近與否，雖然可以透過量化讓使用者輕易感受數值上的差異，但在閱讀行為上並無實質上的縮短時間，若將類別變數的測量應用於文章應可提出立意明顯的效果。

第 3 章　知識版圖的研究與利用

本研究認為概念的形成是經由文章的收納，詞彙的量化，高頻的分析，才得以理解聚集的成份。故本章節概分兩個部份，第一、理解概念成份的型成；第二、將概念具體化呈現。

3.1 知識版圖

人類對認知的行為包含圖形比對、符號搜尋、記憶廣度、選擇性反應等方法，其運作方法可以是獨立，也可以相互結合。當視覺導入各式各樣的訊息時，大腦開始利用輸入的訊息在適當的「記憶廣度」裡尋找，並由當下的「選擇性反應」判斷是要連結到過去曾經出現的記憶，還是建立新的訊息。我們利用知識版圖來達成記憶廣度及選擇性反應。

知識版圖的構思在學習初期，每一詞彙的發生就像記憶單元的建立並留下印象，而且或許含有延伸觸鬚等待著新記憶單元的探訪；隨著文章本身的篇幅或收集文章的篇數我們將建立起相同或相異詞彙的關聯。其建構使得整個詞彙的延伸可能變成龐大且複雜，最後就像是有經歷學習的大腦，故我們稱此為知識版圖（Knowledge Domain）。

當你嘗試在腦中放入某項資訊時，從一開始就需確實了解其意義及角色。死記一些零碎的訊息是非常的沒效率的，如果能適當說明記憶某項訊息的重要性，就比較容易被留存下來，另外如果能在資訊中加入現實狀況及實際例子，就會變得更生動活潑而比較容易回想起來。

知識版圖像是圖形（Graph）的延伸觀念，每一個 vertex 不代表一個起點或終點，而是一個可以觸類旁通的端點，必要時透過 hyper-edges 的連結接觸到各種不同的詞彙，其目的是用來建立起知識概念間的相互網路。知識版圖在建構的過程中蘊藏幾種特性：

- 成員初期可能是少部分詞彙，隨時等待另一個詞彙的延伸，像是神經元的突觸（Synapse）。
- 知識版圖的型式亦像地圖般的城市及道路，城市與道路都有其隸屬的名稱、關係及特性；故知識版圖的詞彙意義或關聯權重大小、關聯對象都有其存在的目的。
- 詞彙表現可以是資訊共享或獨立。例如，詞彙「鯨豚」「鴉嘴獸」可以是共同的話題，或各自擁有主題。
- 當特定詞彙、特定對象及一定水準的頻率緊密表現時，其可能隱含緊密又複雜的關係。

3.2 知識版圖的形成

為了知識版圖的架設，我們利用多維陣列空間做為記憶儲存的描繪，並以 Hyper-Graph G =（V, E）表示之。其中 V（Vertex）象徵詞彙，E（Edge）則代表詞彙關聯。關聯的作用在於透過維繫勾勒出互助或牽引的詞彙，關聯的表現可以是版圖中的任何成份，且關聯的對象可以是二個或者是多個詞彙的陳述。例如，「爬蟲類」聯想到「恐龍」進而聯想到「氣候」等強烈關聯。

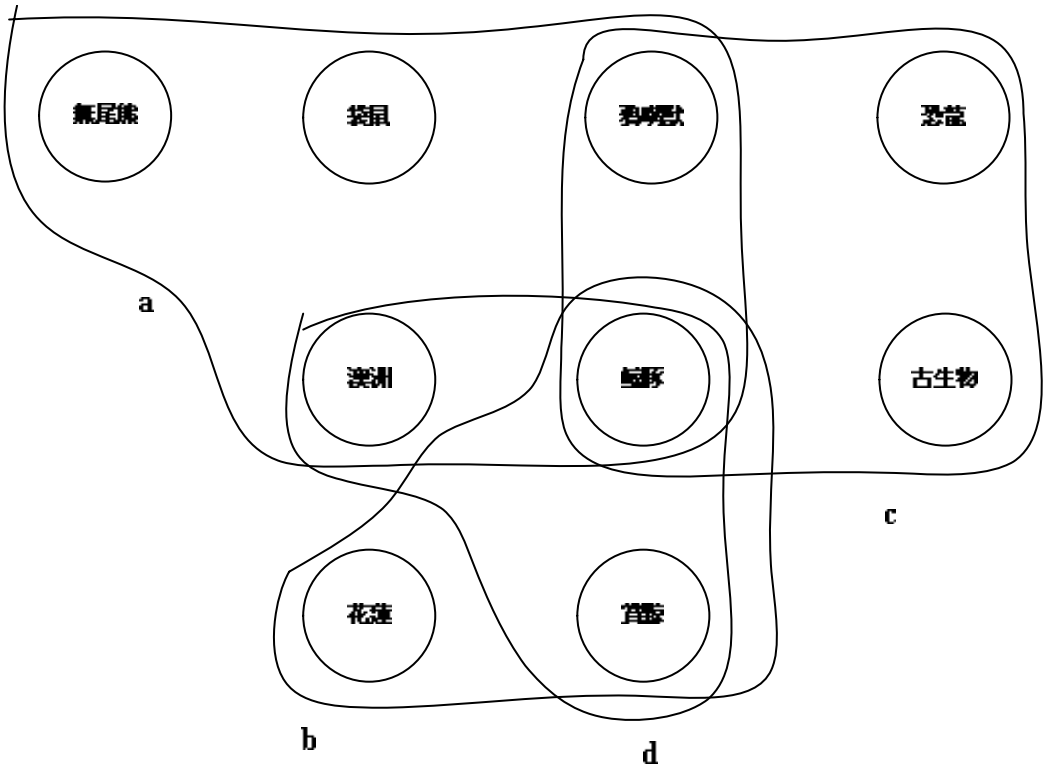
概念成員的形成初期可能會有部份詞彙重疊交錯於各領域，就像是建立起不同的可能關係及知識成份。而這種曖昧關係的情形即使在自然界也常發生，例如春天 4 月就有颱風。故詞彙的關聯可能到處是環環相扣。

為了明確的反應發生的可能性，我們經由關聯法則築起知識版圖。我們將每次發生的結果視為學習後的知識概念，其就像腦袋中的智慧，在經歷溫習及頻率的確認後，逐漸成為記憶的一部份，且隨時等待著觸發和共鳴。以圖 3.1 為例，當人類的視覺帶入「鴉嘴獸」及「恐龍」的瞬間，其會於 c 族群進行思考，甚至帶出「古生物」及「鯨豚」的突觸聯想；但若僅是面對「鯨豚」其可能會有跨越 a、b、c、d 四個族群的探索性思考。

3.3 概念成員的建立

我們企圖利用切割方法將探索性思考降到最低，故對上述所建立的知識版圖進行概念分解。所謂的切割是衡量各族群的權重是否平衡；其切割標準可分為詞彙權重、詞彙關聯及混合權重共三種，請參照圖 3.2（a）（b）（c）範例。

我們將圓圈代表不同的詞彙，並以 a、b、c 表示之，圓圈內括弧的數字代表詞彙的權重，實線代表詞彙關聯，實線上的數字代表詞彙關聯的動機程度；虛線代表分隔切割後的結果。



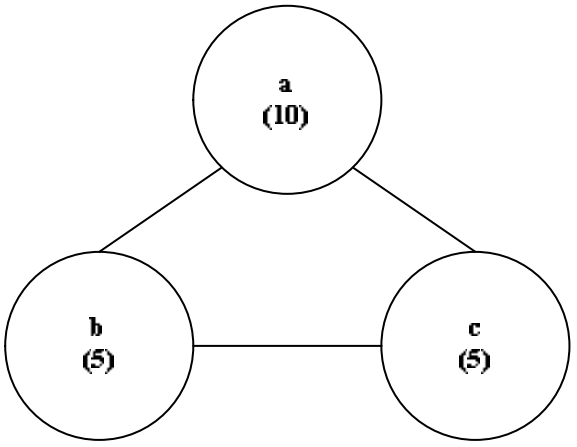
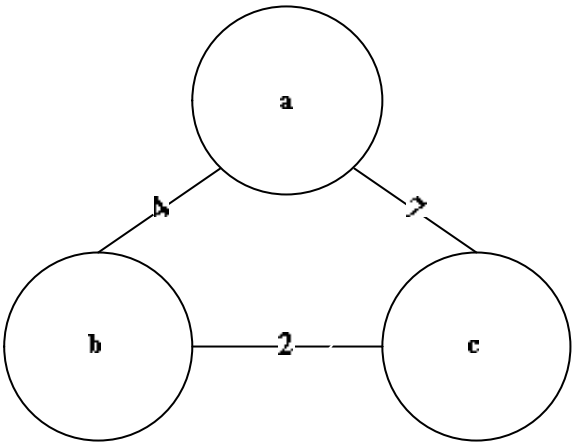


圖 3.2（a）詞彙權重進行切割



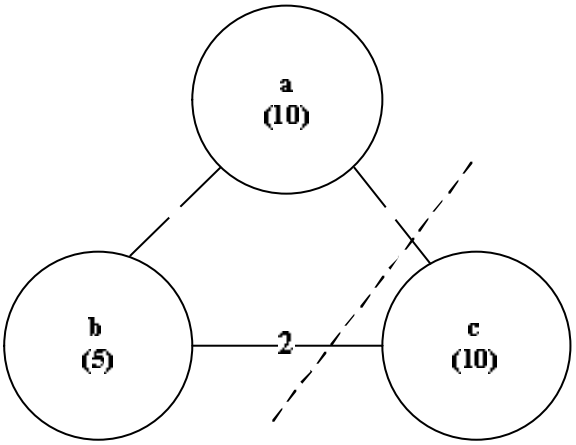


圖 3.2（c）以詞彙權重及詞彙關聯進行切割

切割的原則受限於權重及平衡的標準。以圖 3.2（c）**為例，假設欲切割成** 2 **群聚，切割過程可以有** 3 **種途徑，分別是**{ab}{c}、{ac}{b}**或**{a}{bc}**；但第** 2 **種結論的**{ac}**因為詞彙權重及關聯權重太過偏頗，故因失衡的缺點，採不選擇；而第** 3 **種結論的**{bc}**因關聯權重** 2**，不如第** 1 **種切割的**{ab}**關聯權重** 4 **來的緊密且較接近平衡，故在兼顧詞彙權重、關聯權重及平衡分佈的共同解析原則下，我們採**{ab}{c}**做為切割群聚的結果。**

切割的時間複雜度（O**）除了受限於上述條件之外，尚有切割群數。假如欲切割**

成 4 群聚，第 1 次切割我們會對切成為兩群，等到確認平衡後，再細切各自內部的群聚項目，使其成為四份群聚，故切割群數亦決定切割的時間複雜度。這就是所謂階層式的遞增切割法。

進行群聚切割，就好像是建立起群聚專屬的概念成員，其結果具有易於理解、證明及客觀的特色，較其他群聚測量方法適合於文章群聚的應用。在神經科學領域中，科學家發現思考的動作係藉由突觸串聯整個思考。突觸可分為化學性突觸（chemical synapse）和電突觸（electrical synapse）兩種，前者是指神經元之間已知的關連，其進行的是記憶性的思考；而後者是對內、外環境變化較為敏感的聯想性動作，其象徵著創造性的思考。故本文利用關聯來象徵突觸的特性，建立起研究基礎。

3.4 動態群聚計分法

當圖解模型切割成不同的組成時，我們定義該群聚項目都是重要且必須存在。圖解模型切割係依照使用者指定的目標進行切割，在最佳均衡的訴求下，得到需求的群聚個數及其項目；該結果代表著各種不同的概念及合理的概念內容。

當知識版圖被切割達成時，我們可以利用旗下各群聚所產出的詞彙元素發展出共鳴式文章群聚。其形成就好像是人類利用已知的記憶廣度與新導入的文章詞彙進行項目比對，以評核文章被隸屬的認同程度。

我們令一篇文章的詞彙項目 Ti 與各切割後的聚集詞彙元素 Ci 進行概念比對，以求得文章被觸發的程度；故可以用此來決定文章適合被放置在那一個聚集 k。群聚得分（Cluster Score）的計算公式如下：

$$ClusterScore(k) = \frac{|Ti \cap Ci|}{|Ci|}$$

上述公式中，分子表示每一文章 Ti 的所有詞彙與群聚 k 的所有隸屬項目 Ci 共同發生的表現，其結果介於 0 到群聚項目 Ci 的數量之間。故本公式亦可得知群聚得分將介於 0 到 1 之間。

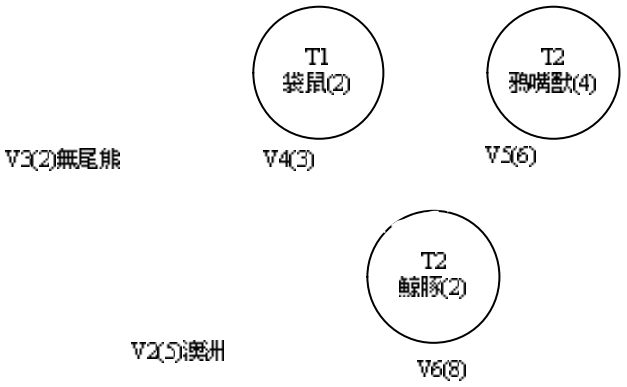
此外，由於群聚數量採變動值，故我們又稱之為動態群聚計分法。

3.5 內部排序計分法

我們發現群聚得分雖然可以使文章尋找到適切的群聚，但同一群聚內的文章之間得分經常重疊。我們試圖想像系上在申請入學篩選時，系辦有時會以像是特定的文理學科或特殊表現等成績，做為錄取資格的衡量標準，甚至透過加重計分的方法給予錄取。我們將此觀念應用於群聚內部排序。

假設文章 T1 由概念 V1、V2 所組成，文章 T2 由概念 V2、V3、V4 所組成，且 V1、V2 及 V3 都是切割後所得到的同一概念元素。此時我們對文章 T1 及 T2 所得到的群聚得分（Cluster Score）一樣都是 0.666。為了提供給資料調閱人員閱覽建議，本研究進行更細部的文章排序，以利群聚閱覽推薦。

我們將同一群聚內的所有概念，視為一個完整的個體，其組成元素各有自己的舉足輕重，透過詞頻的加入，其目的是希望藉此來突顯每一文章所佔據的比例。以圖 3.3 為例，群聚 a 分別由{澳洲(5),無尾熊(2),袋鼠（3）,鴨嘴獸（6）,鯨豚（8）}五個概念所組成，文章 T1 由{袋鼠（2）}取代，文章 T2 由{鴨嘴獸（4）,鯨豚（2）}的概念表示之，我們試圖了解 T1、T2 在群聚 a 的影響地位。



群聚a的總權重a=2+3+6+5+8=24
T1於群聚a的IRS=(2/3)/24=0.0277
T2於群聚a的IRS=((2/8)+(4/6))/24=0.0381

在圖 3.3 當中，T1 的詞彙僅佔 V4 的 2/3，且僅佔有群聚 a 的 0.027；而 T2 在群聚 a 分別佔有 V6 的 2/8 及 V5 的 4/6，故佔領群聚 a 有 0.0381 的比率。我們將此得分特性命名為內部排序計分法（Internal Rank Scoring，簡稱 IRS）。其運算公式如下：

$$\text{Internal Rank Score}(T) = \frac{\sum_{Ci \cap Ti} \frac{VwTk}{VwCk}}{\sum_{i=1}^{i} \text{Vw}Ci}$$

其中，*Vw* 代表詞彙權重，分子代表 $Ci \cap Ti$ 時，每個特定群聚項目與文章項目共同引發的權重總和；而分母代表群聚 a 項下的總和權重。

第 4 章.　處理方法

面對各種領域的文章主體，我們大膽的假設文章搜尋的精確是有困難的。我們認為文章的表述是自然語言的一種，其必然存在著詞彙界定的困擾；例如「美國總統」不僅能出現在政治議題的新聞，還可能有藝文活動或社會頭版等文章。

文章當中的概念經常是高度相關的一群詞彙所組成；人們經常會利用文章來闡述他們想要表達的精神、看法或認知。故每位作者一定是先有概念才會有文章的產出，為了完整的表達概念的精神，每位作者必須是善用各種可能的詞彙來形容、反襯或補述，以見證其概念存在的理由，就像是獨立的一個小宇宙。

我們以多重主體屬性（Multi-Objective）的觀念建立起不同族群的重要成份。其可以使得每個母體都能透過領域區隔來達成文章歸屬。

本研究試圖將文章收集後以圖 4.1 的流程進行文章整理。首先利用關聯規則挖掘所有詞彙的關係，再經由超圖解模型的架構闡述整個關鍵詞彙的交互作用；並切割該模型使其形成易於理解的概念族群及文章群聚。

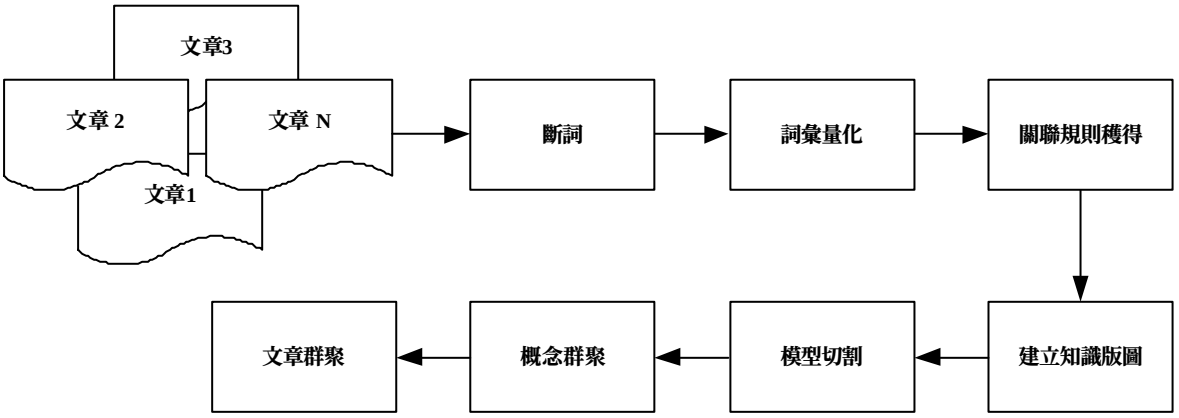


圖 4.1 超圖解模型的文章分類流程

4.1 詞彙量化

文章的關鍵元素並不需要保留所有的詞彙。文章字句是由各種屬性所組成，經常

包含各種主詞、副詞、形容詞等屬性，甚至贅詞、代名詞或作者個人的寫作習慣夾雜，故我們根據章節 2.1 及 2.2 介紹的方法，建立起詞彙並付予詞彙權重。透過交互頻率（Term Frequency Invert Document Frequency, TFIDF）【10,12】其幫助我們了解詞彙被使用或認可的程度，且可以適用於詞彙權重的建立。

詞彙的權重量化，多以詞頻與逆向文件頻率的內積(TF * IDF)得知。計算公式如下：

（1） Term Frequency（TF）：計算單字在某文件的出現頻率

$$TF_{ij} = \frac{n_j}{n_{all}}$$

代表詞彙 j 在文件 i 的出現頻率，其中

n_j：表示詞彙 j 在文件 i 的出現次數。

n_{all}：表示文件 i 內所有具意義的總詞頻。

（2） Inverse Document Frequency（IDF）：計算逆向文件頻率

$$IDF_j = \log_2 \frac{N}{df_j}$$

N：代表所有文件的總數

df_j：代表詞彙 j 有出現過的文章總數

當這兩者相乘之後，即代表修正過後的關鍵詞 T_j，在文件 d_i 的加權(weight)，如下式所述：

$$W_{ij} = TF_{ij} * IDF_j$$

上述公式被應用於各文章詞彙當中。例如有一個關鍵詞「道瓊指數」在一篇文章出現 10 次，而此篇文章共有 100 個關鍵詞，所有文件集合共有 10000 篇文章，而「道瓊指數」一詞在 10000 篇文章內，曾出現在 5 篇文章

TF=10/100=0.1

IDF=log(10000/5)　11

詞彙「道瓊指數」加權值=0.1*11=1.1

4.2 詞彙處理及關聯量化

斷詞系統雖然可以萃取出專屬於中文環境的詞彙字句，但我們認為詞彙的運用，除了以交互頻率的門檻值（Threshold）之外，尚有不同的作法。關於交互頻率的篩選，我們採較慎重的關聯法則【8】，以免錯失重要資訊。關聯法則的目的主要是為了挖掘詞彙之間的潛在關係。就好像文章中有人習慣直接談論「金磚四國」，有人習慣談論「中國」「印度」的發展，但亦有人泛指「新興市場」，上述各名詞將可能透過關聯法則被挖掘，進而有助於概念成份的完整。

在詞彙處理過程中，我們經常發現異詞同義的現象，例如「車輛」、「汽車」或「車子」可以視為同一詞彙。透過交互頻率及關聯法則，我們亦可引發詞彙之間的關係，其避免少數存在的重要詞彙遭受忽略。關聯法則可幫助我們探索詞彙之間更深入的連帶關係（Pattern Relationship）或多重主體屬性。

透過高頻項目集(Frequent Item Set)及信心指數（Confidence）的建立，我們就能獲得詞彙 A 聯想到詞彙 B 的存在強度，其幫助我們得到的詞彙不但有效且客觀。信心強度 Confidence 的計算公式如下：

Confidence（A→B）= Prob（A B）/ Prob（A）

我們透過高頻項目集(Frequent Item Set)及信心指數（Confidence）的成果，做為挖掘詞彙的關聯及權重量化。本研究將每篇文章及其項下的詞彙視為市場購物欄分析（Market Basket Analysis）的延伸，因為每篇文章具有明顯的、有用的和無法解釋的組織特性。詞彙群集在關聯法則的幫助下，可以達成下列優勢：

1. 具有歸類詞彙項目的效果。
2. 製造相關資料項目的可能。
3. 藉由沒興趣的項目來刪除原始規則內的資料。

關聯法則雖然可尋找到極細緻的大型高頻項目集(Large Frequent Item Set)，但我們並不需要如此的費時及需求。如同章節 4.2 所提及，故原則上我們僅取到 large 2 到 5 的 Large Frequent Sequences 來建置知識版圖；其具備快速執行、避免重要但少數的詞彙被疏漏，及不影響群聚效果。

Apriori 演算法【16】為研究關聯式法則的一個最具代表性的演算法之一。利用循序漸進的方式，找出資料庫中項目的關係，當資料中最多個項目共同出現，且其出現的次數為最高次的群落，此群落的組合便是這些資料中的主要規則。

執行步驟如下：

(1) 訂定最小支持度與最小信賴度。
(2) Apriori 演算法將所有資料項皆視為候選物項，由計算每個候選物項的出現次數，並且將出現次數大於或等於最小信賴度的候選物項留下，則第一次篩選完成。
(3) 由前一個步驟篩選出的候選物項計算關聯組合的出現次數，用同樣的方法剔除弱關聯性的候選物項組合。直到資料中，候選物項集合的最大數量次數計算出來為止。
(4) 最後留下的強關聯性候選物項組合即為整個資料關聯中的主要連結。

自此，Apriori【17】演算法用物項之間共同出現次數頻率計算，找出互相的關聯，尤其是強關聯。不過也因為這樣的計算方法，相對的失去了也許挺重要的關鍵連結，更無法清楚解釋為何會是這樣的組合!？

4.3 關聯整併

自然語言的模糊且易變，使我們無法一一釐訂各種詞彙的互動關係。這種情形通常發生在難以界定的模糊定位，像是「新興市場」可以聯想到國家「俄羅斯」、「巴西」或區域「拉丁美洲」、「亞洲市場」，甚至是新種詞彙「BRIC」。為了定義詞彙間互相糾葛的同類概念，我們以 edge 勾勒出彼此合理存的權重。

面對章節 4.2 的關聯重疊時，權重的計算將以平均處理之。例如，A，B，C 的關聯法則若共有六種，且其信心指數分別如下{A} ⇒ {B C}：0.8，{A B} ⇒ {C}：0.4，

{A C}⇒{B}：0.6，{B}⇒{A C}：0.4，{B C}⇒{A}：0.8，且{C}⇒{A B}：0.6，則定義 A，B，C 的 Edge 權重為 0.6＝（$\frac{0.8+0.4+0.6+0.4+0.8+0.6}{6}$）。

關聯整併的目的，是為了消彌聯想的順序或方向，其產生使得 Edge 的形成是在最低支撐值的情況下，檢視詞彙群聚的關聯。

4.4 切割方法

Karypis【11】提出一種針對多維空間的超圖解模型群聚探究。在該模型下，各元件被視為頂點（Vertex），且元件們可以經由 Hyper-Edge 串聯相近似的關聯；並透過 HMETIS【5】的切割後，其每一個群聚均足以視為一個完整的群集。Karypis 等人將該架構應用於超大型積體電路的繪圖技術，以應用於元件配置；其目的在於盡可能的最大化內部連結，又預期以最少的連結來平均分散實體裝置於各電路層板（Layer）。同樣的該方法也可應用於硬碟資料的擺放，其目的是在減少資料讀取時，時間被消耗於跨越磁區或抓取尋找。

這是一種多層次的切割方法，通常是以最小權重及最佳平衡做為切點；其透過反覆計算找出一定程度的合理化，使得每一群聚保留良好的交互程度，而群聚之間盡可能達到關係薄弱，甚至有明顯的反差效果。

4.5 切割標準的衡量

在切割過程中，我們利用圖解切割的方法找出各分群中最佳的組合。從一開始到切割成 k 群聚，我們皆使用下述公式做為剔除不適當成員的衡量標準，其中 e 代表關聯，C 代表一個群聚，群聚適合性（Fitness）的計算公式如下：

$$fitness\ (C) = \frac{\sum_{e \subseteq C} Weight\ (e)}{\sum_{|e \cap C|>0} Weight\ (e)}$$

這是測量 Edge 及其重疊的 edge 在群聚 C 的結果，如果 fitness 是大於標準值則表示該重疊的 Edge 應該被納入該群聚 C；反之則表示該重疊的部份應該被剔除或屬於

另外的群聚 C。

另外，每次經 fitness 後所保留下的群聚 C 尚須檢驗，並將沒有緊密結合的 Vertex 過濾。關連性（Connectivity）的計算公式如下：

$$connectivity(v,C)=\frac{\left|\{e|e\subseteq C, v\in e\}\right|}{\left|\{e|e\subseteq C\}\right|}$$

Connectivity 如果是大於標準值則表示該群集內的 Vertex 適合保留，使得群聚成員具有說服力。

4.6 切割型態的建構

我們將知識版圖透過空間描述檔（Space Descript File，SDFile）來勾勒其結構，以做為切割的對象。空間描述檔具有敘述各種權重的能力，包括詞彙（Vertex）權重、關聯（Edge）權重或前兩者權重的混合共分三種。

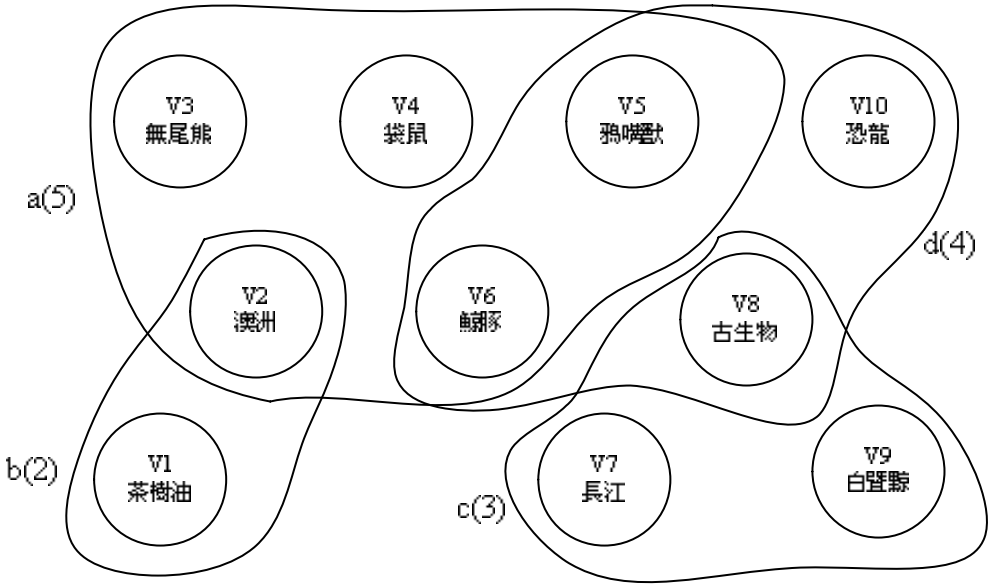
（1）以關聯為空間描述檔（Edge SDFile）的超圖解模型：本型態以詞彙關聯的最少頻率做為權重。以圖 4.2(a)為例，該圖分別以四個群聚 a、b、c、d 做為關聯空間描述檔的主軸，其具有以下幾個特徵：

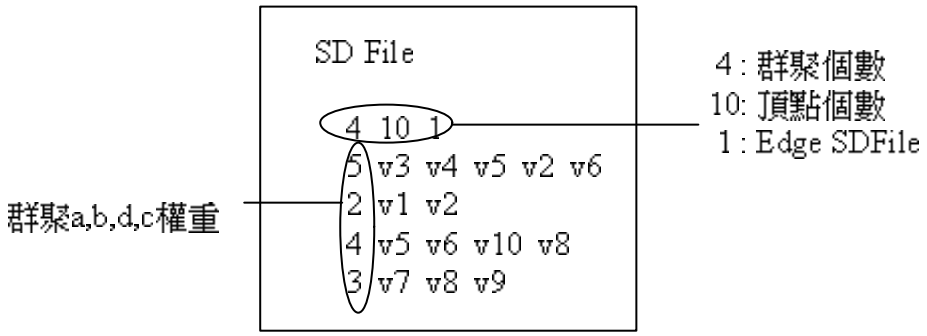
1. 該檔案的第一行必須有 3 個整數，包含詞彙關聯個數、詞彙個數以及 Edge SDFile 的識別記號「1」。
2. 該檔案自第 2 行起到第 5 行，分別代表不同的群聚。每一群聚必須是獨立一行；且該行的第一位代表該群聚的權重。

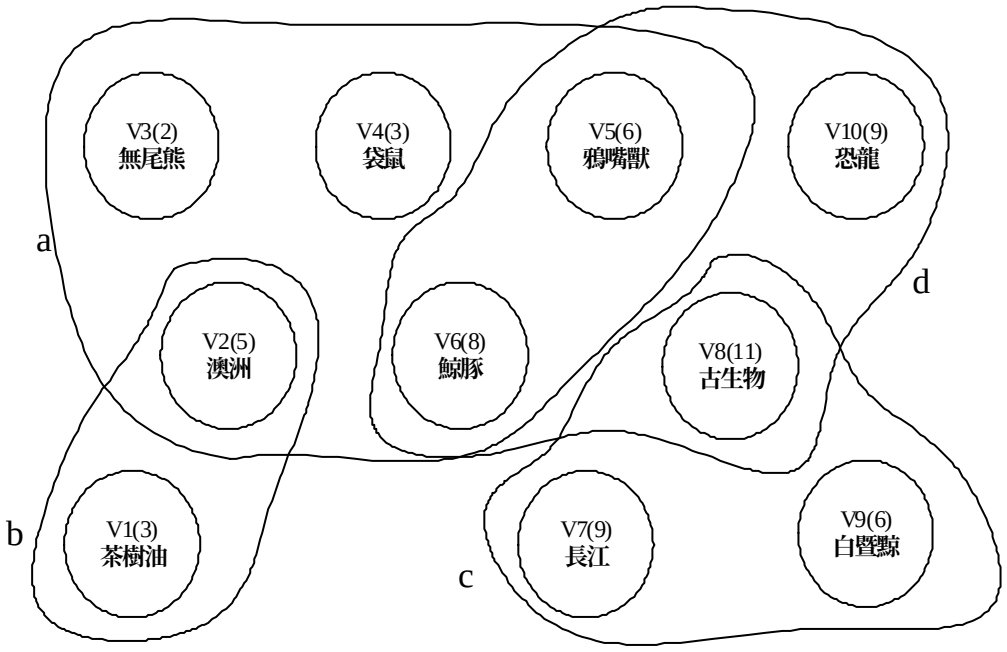
（2）以詞彙權重為空間描述檔（Vertex SDFile）的超圖解模型：本型態以各詞彙的交互頻率做為詞彙空間描述檔的主軸。此外，關聯權重雖不引用，但仍須維持群聚的組成關係。以圖 4.2(b)為例，詞彙「澳洲」以 V2 表示其權重為 5；詞彙「無尾熊」以 V3 代表，其權重為 2。該文字檔特徵如下：

1. SD File 的第一行必須有 3 個整數，分別是詞彙關聯的個數 4，詞彙的個數 10，以及 Vertex SDFile 的識別記號 10。
2. 自第 2 到 5 行分別代表不同的群聚。每一群聚個別獨立一行，不得換行。自

第 6 到 15 行分別依序代表不同的詞彙權重。







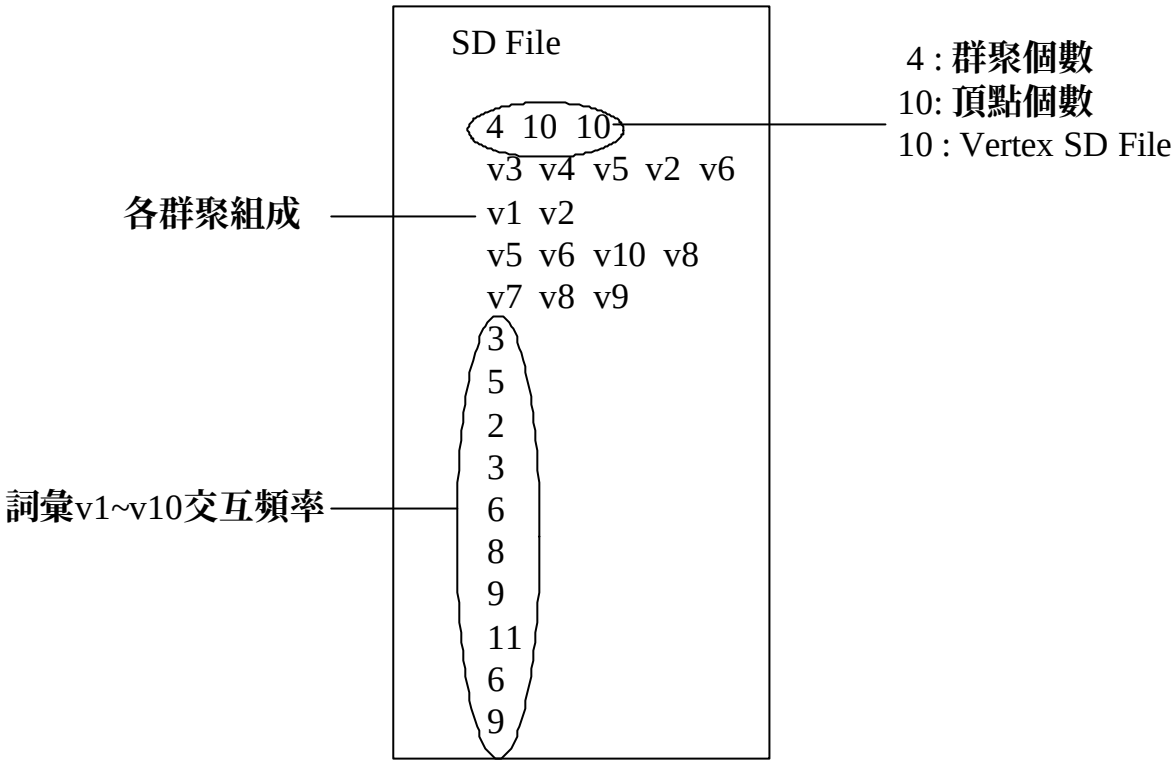
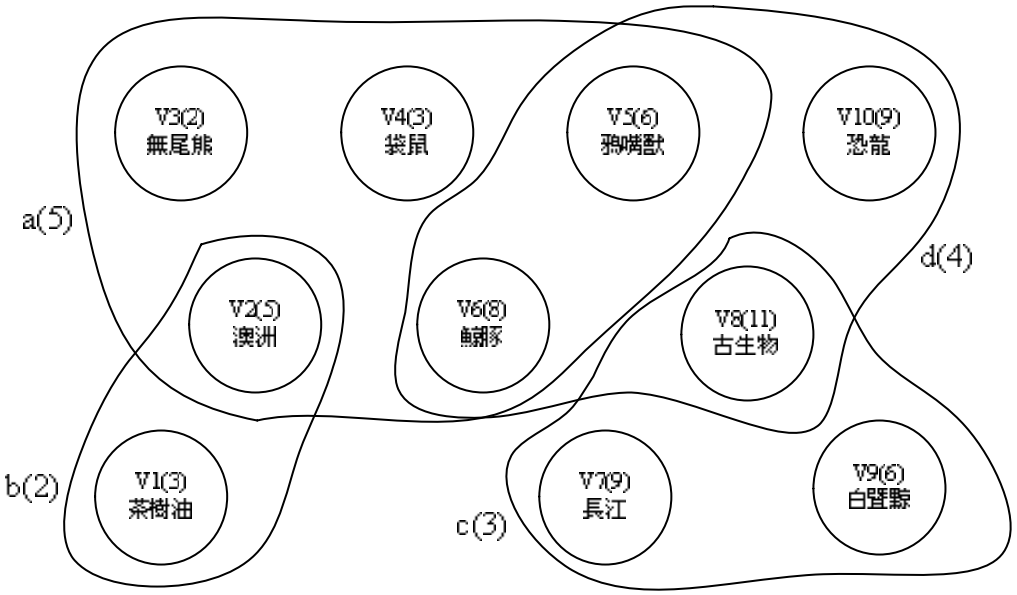
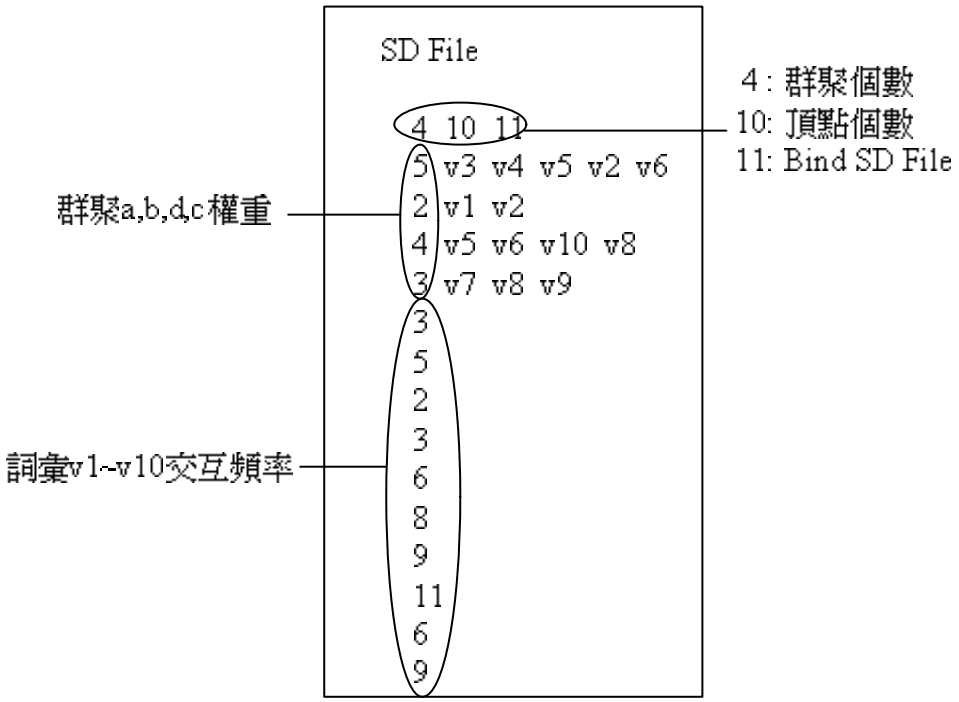


圖 4.2（b）詞彙強度的空間描述檔

（3）以混合權重為空間描述檔（Bind SDFile）的超圖解模型：本型態將各詞彙權重及關聯權重進行合成，以圖 4.2(c)為例。該文字檔特性如下：

1. SD File **的第一行必須有 3 個整數, 分別是詞彙關聯的個數 4, 詞彙的個數 10, 以及混合權重空間描述檔的識別記號 11。**
2. **自第 2 到 5 行分別代表不同的群聚，且各行的第一個數值代表該群聚的關聯權重。**
3. **自第 6 到 15 行分別依序代表不同的詞彙權重。**


V3(2)
無尾熊
V4(3)
袋鼠
V5(6)
鴨嘴獸
V10(9)
恐龍
a(5)
d(4)
V2(5)
澳洲
V6(8)
鯨豚
V8(11)
古生物
b(2)
V1(3)
茶樹油
c(3)
V7(9)
長江
V9(6)
白蛙黯


SD File
4 10 11
5 v3 v4 v5 v2 v6
2 v1 v2
4 v5 v6 v10 v8
3 v7 v8 v9
3
5
2
3
6
8
9
11
6
9
4 : 群聚個數
10: 頂點個數
11: Bind SD File
群聚a,b,d,c權重
詞彙v1~v10交互頻率

況，故僅擇前者當中的 shmetis 做為工具。

我們將高頻詞彙的關聯結果透過 HGraphFile 做為知識版圖的描述，並以空間描述檔（Space Descript File，SDFile）做為 shmetis 的輸入檔案。輸入檔案的定義 HGraphFile 分成三種不同的切割的標準，其型態已於章節 4.6 述之。

我們選擇以此程式為切割工具，最主要的原因是因為該切割具有允許群聚頂點不均等的特性，較貼近實務上的群聚切割。

Shmetis 可採用下述方式，於命令提示列底下直接執行，其切割（Partition）方法為：

shmetis HGrapgFile Nparts UBfactor

HGraphFile：知識版圖的輸入檔案

Nparts：表示預期切割的成份

UBfactor：此參數被用於指定切割期間時，可允許不平衡（UnBalance factor）的比率，該參數必須為介於 1 到 49 之間的整數；如果以切割成 2 部份而言，其結果將會使得每次切割後的頂點數量介於 $(50-b)n/100$ 到 $(50+b)n/100$ 之間；其中 n 為空間描述檔的頂點數量，b 為 UBfactor。

透過上述作法，如欲切割成 2 群聚，且令 b 等於 5，則每一群聚產生出來的個數將介於 $0.45n$ 到 $0.55n$ 之間的頂點個數；同樣的，如果是要遞迴式切割成 4 群聚，則上述的每一群聚數將變成為 $0.45^2 n$ 到 $0.55^2 n$ 之間，也就是，每群聚有 $0.2n$ 到 $0.3n$ 之間的頂點個數。

在切割結束時，我們將切割後的結果存放於文字檔，並以 SDFile.part.Nparts 做為檔名。其中 SDfile 為知識版圖的檔名，Nparts 為切割群聚數，我們將章節 4.6 的三種空間描述檔，欲切割成 3 群聚，且 UBfactor 設定為 3 時，切割後的結果將如圖 4.3 所示。

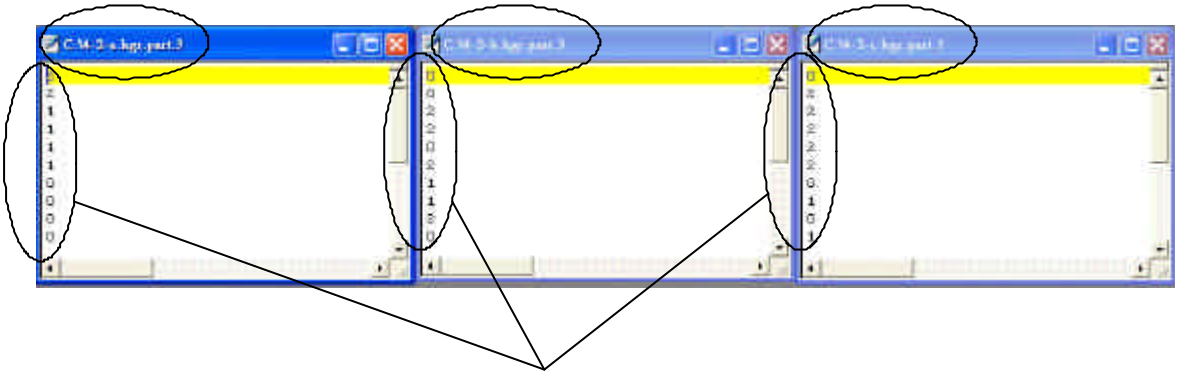
我們定義同一群聚及其項下的成分代表同一概念。以圖 4.2（a）（b）（c）為模式的切割結果，我們得到圖 4.3 的三種不同群聚結果：

1. 關聯強度切割：群聚 0={V7,V8,V9,V10}，群聚 1={V3,V4,V5,V6}，群聚 2={V1,V2}
2. 詞彙權重切割：群聚 0={V1,V2,V5,}，群聚 1={V7,V8}，群聚

2={V3,V4,V6,V9}

3. **綜合權重切割：群聚** 0={V1,V7,V9}，**群聚** 1={V8,V10}，**群聚** 2={V2,V3,V4,V5,V6}

關聯強度切割結果 **詞彙權重切割結果** **綜合權重切割結果**



詞彙自V1**到**V10 **歸屬坐落群聚**0,1,2**三群聚**

圖 4.3 針對圖 4.2 的三種切割原理進行群聚歸屬

第 5 章　實驗及分析

5.1 系統需求

本系統考慮了一般使用者與系統管理者兩種不同身分的角色，其可以簡單的運作文章群聚系統。

系統管理者根據文章收集後的資料，僅處理關聯調整及群聚微調兩件工作。關聯調整主要是在排除詞彙萃取所無法排除的漏網詞彙以及進行必要的強迫關聯；而群聚微調（UBfactor）則是針對詞彙內容的成份，視其必要進行調整，以彌補切割可能的疏漏。

系統管理者可以設定系統要管理的文章，並可檢視文章分類的結果。管理者必須配合文章的特性，調整相似度參數，使一般使用者得到最佳的結果。請參考圖 5.1 的使用者介面。

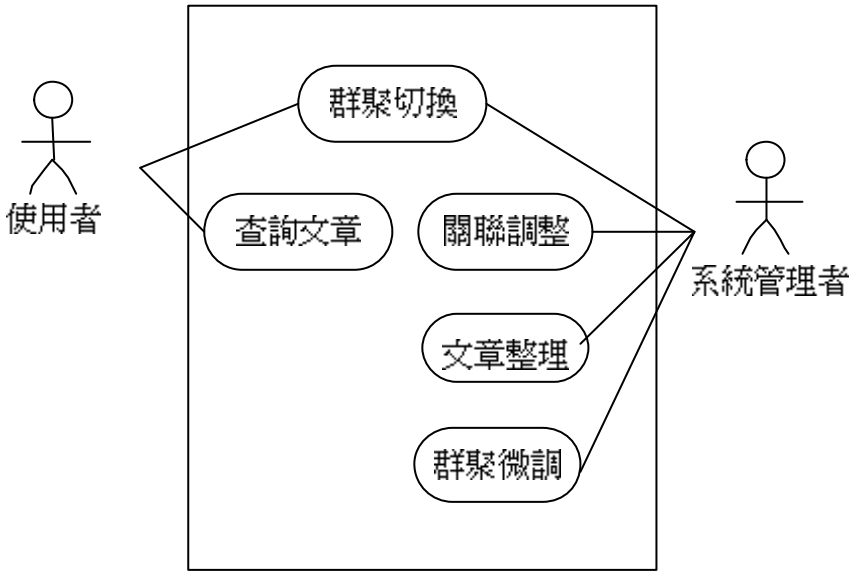


圖 5.1　使用者介面概述

一般使用者可以輸入任意的關鍵詞及指定群數，使系統依據切割後的共鳴式群聚法，提供各類的文章給使用者。

5.2 實驗過程

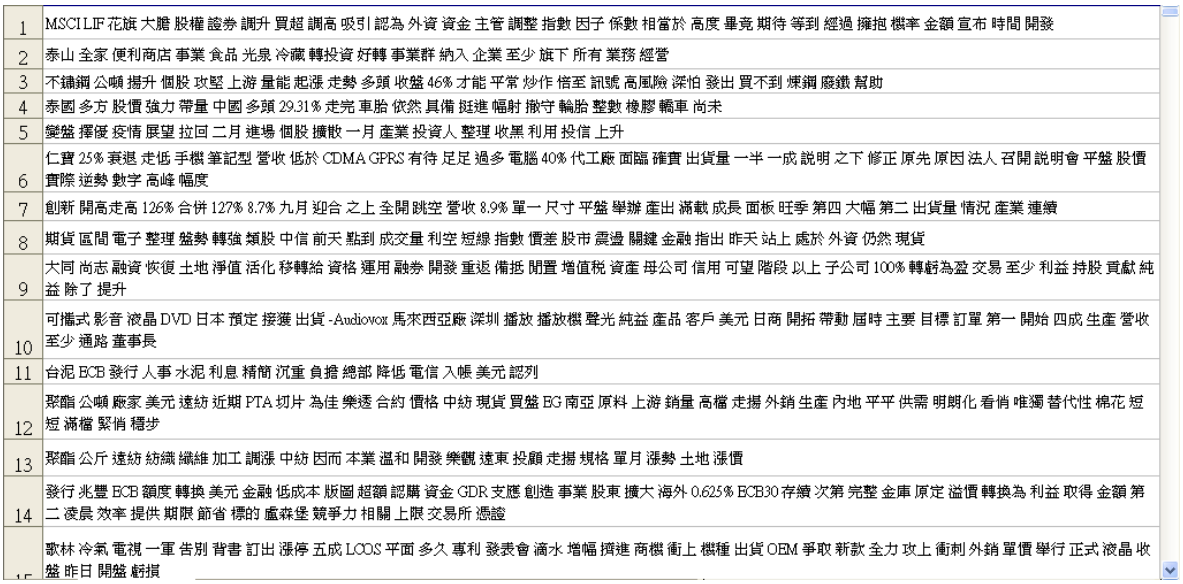
我們將收集後的 300 篇文章進行斷詞，並對斷詞後的詞彙進行過濾，共得到 18,920 個 2 元詞以上的有效詞彙。

此外，我們針對上述的有效詞進行詞彙量化。透過章節 4.1 的作法，我們將詞彙在該篇本文的被使用（TF）及在其他文章的被利用（IDF）進行計算，以做為詞彙的權重。以圖 5.2 為例，「MSCI」在 001.txt 共出現 6 次，且 001.txt 內共有 100 個有效詞，故其 TF 為 0.06，而 IDF 則是以該「MSCI」在 300 篇文章中的出現的對數。

文章	詞彙	WordFreq	TF	IDF	加權值	文章	詞彙	WordFreq	TF	IDF	加權值	文章	詞彙	WordFreq	TF	IDF	加權值
001.txt	100%	1	0.01	3.70	0.0370	001.txt	因子	1	0.01	6.91	0.0691	001.txt	擴大到	1	0.01	5.91	0.0591
001.txt	LIF	3	0.03	6.91	0.2072	001.txt	年線	1	0.01	5.91	0.0591	001.txt	爆出	1	0.01	4.59	0.0458
001.txt	MSCI	6	0.06	4.91	0.2944	001.txt	成交	1	0.01	3.59	0.0358	001.txt	證券	4	0.04	3.32	0.1329
001.txt	大量	1	0.01	3.00	0.0300	001.txt	收盤	1	0.01	3.59	0.0358	001.txt	證實	1	0.01	4.91	0.0491
001.txt	大膽	2	0.02	6.91	0.1381	001.txt	至於	1	0.01	2.59	0.0258	001.txt	觀察	1	0.01	3.70	0.0370
001.txt	干擾	1	0.01	5.91	0.0591	001.txt	行情	1	0.01	2.81	0.0281	002.txt	10%	1	0.01	4.59	0.0364
001.txt	不會	1	0.01	3.70	0.0370	001.txt	利多	1	0.01	3.17	0.0317	002.txt	11.4%	1	0.01	6.91	0.0548
001.txt	今天	1	0.01	2.32	0.0232	001.txt	吸引	2	0.02	4.32	0.0864	002.txt	4.76%	1	0.01	6.91	0.0548
001.txt	主管	2	0.02	3.59	0.0717	001.txt	完全	1	0.01	4.09	0.0409	002.txt	一成	1	0.01	5.32	0.0422
001.txt	加上	1	0.01	1.59	0.0158	001.txt	投資	1	0.01	1.59	0.0158	002.txt	一些	1	0.01	4.32	0.0343
001.txt	加權	1	0.01	5.91	0.0591	001.txt	使得	1	0.01	2.59	0.0258	002.txt	一級	1	0.01	5.32	0.0422
001.txt	可能	1	0.01	2.32	0.0232	001.txt	受到	1	0.01	2.59	0.0258	002.txt	八寶粥	1	0.01	6.91	0.0548
001.txt	可望	2	0.02	1.59	0.0317	001.txt	法人	1	0.01	1.59	0.0158	002.txt	力道	1	0.01	3.59	0.0285
001.txt	台北	1	0.01	4.59	0.0458	001.txt	股權	3	0.03	4.59	0.1375	002.txt	口味	1	0.01	6.91	0.0548
001.txt	台灣	3	0.03	2.00	0.0600	001.txt	花旗	3	0.03	5.32	0.1597	002.txt	大陸	1	0.01	2.00	0.0159
001.txt	外資	3	0.03	2.59	0.0775	001.txt	金額	2	0.02	3.32	0.0664	002.txt	大幅	1	0.01	1.59	0.0126
001.txt	市場	1	0.01	0.00	0.0000	001.txt	非常	1	0.01	5.32	0.0532	002.txt	今年	10	0.08	0.00	0.0000
001.txt	本土	1	0.01	5.91	0.0591	001.txt	係數	1	0.01	6.91	0.0691	002.txt	內部	1	0.01	4.59	0.0364
001.txt	交易	1	0.01	3.59	0.0358	001.txt	前波	1	0.01	5.91	0.0591	002.txt	分為	1	0.01	5.91	0.0469
001.txt	交易日	1	0.01	5.91	0.0591	001.txt	南韓	1	0.01	5.32	0.0532	002.txt	日前	1	0.01	3.17	0.0252

圖 5.2　詞彙量化

我們取每篇文章中特定權重的比例，做為知識版圖的成分。圖 5.3 為本研究針對每一文章萃取高頻詞彙的百分之 30 做為關聯法則的基礎之部分畫面。



1	MSCI LIF 花旗 大膽 股權 證券 調升 買超 調高 吸引 認為 外資 資金 主管 調整 指數 因子 係數 相當於 高度 畢竟 期待 等到 經過 擁抱 櫥率 金額 宣布 時間 開發
2	泰山 全家 便利商店 事業 食品 光泉 冷藏 轉投資 好轉 事業群 納入 企業 至少 旗下 所有 業務 經營
3	不鏽鋼 公噸 揚升 個股 攻堅 上游 量能 起漲 走勢 多頭 收盤 46% 才能 平常 炒作 倍至 訊號 高風險 深怕 發出 買不到 煉鋼 廢鐵 幫助
4	泰國 多方 股價 強力 帶量 中國 多頭 29.31% 走完 車胎 依然 具備 挺進 幅射 撤守 輪胎 整數 橡膠 輪車 尚未
5	變盤 擇優 疫情 展望 拉回 二月 進場 個股 擴散 一月 產業 投資人 整理 收黑 利用 投信 上升
6	仁寶 25% 衰退 走低 手機 筆記型 營收 低於 CDMA GPRS 有待 足足 過多 電腦 40% 代工廠 面臨 確實 出貨量 一半 一成 說明 之下 修正 原先 原因 法人 召開 說明會 平盤 股價 實際 逆勢 數字 高峰 幅度
7	創新 開高走高 126% 合併 127% 8.7% 九月 迎合 之上 全開 跳空 營收 8.9% 單一 尺寸 平盤 舉辦 產出 滿載 成長 面板 旺季 第四 大幅 第二 出貨量 情況 產業 連續
8	期貨 區間 電子 整理 盤勢 轉強 類股 中信 前天 點到 成交量 利空 短線 指數 價差 股市 震盪 關鍵 金融 指出 昨天 站上 處於 外資 仍然 現貨
9	大同 尚志 融資 恢復 土地 淨值 活化 移轉給 資格 運用 融券 開發 重返 備抵 閒置 增值稅 資產 母公司 信用 可望 階段 以上 子公司 100% 轉虧為盈 交易 至少 利益 持股 貢獻 純益 除了 提升
10	可攜式 影音 液晶 DVD 日本 預定 接獲 出貨 -Audiovox 馬來西亞廠 深圳 播放 播放機 聲光 純益 產品 客戶 美元 日商 開拓 帶動 屆時 主要 目標 訂單 第一 開始 四成 生產 營收 至少 通路 董事長
11	台泥 ECB 發行 人事 水泥 利息 精簡 沉重 負擔 總部 降低 電信 入帳 美元 認列
12	聚酯 公噸 廠家 美元 遠紡 近期 PTA 切片 為佳 樂透 合約 價格 中紡 現貨 買盤 EG 南亞 原料 上游 銷量 高檔 走揚 外銷 生產 內地 平平 供需 明朗化 看俏 唯獨 替代性 棉花 短短 滿檔 緊俏 穩步
13	聚酯 公斤 遠紡 紡織 纖維 加工 調漲 中紡 因而 本業 溫和 開發 樂觀 遠東 投顧 走揚 規格 單月 漲勢 土地 漲價
14	發行 兆豐 ECB 額度 轉換 美元 金融 低成本 版圖 超額 認購 資金 GDR 支應 創造 事業 股東 擴大 海外 0.625% ECB30 存續 次第 完整 金庫 原定 溢價 轉換為 利益 取得 金額 第二 凌晨 效率 提供 期限 節省 標的 盧森堡 競爭力 相關 上限 交易所 憑證
15	歌林 冷氣 電視 一軍 告別 背書 訂出 漲停 五成 LCOS 平面 多久 專利 發表會 滴水 增幅 擠進 商機 衝上 機種 出貨 OEM 爭取 新款 全力 攻上 衝刺 外銷 單價 舉行 正式 液晶 收盤 昨日 開盤 虧損

圖 5.3　詞彙組成

透過 Apriori【17】**建構詞彙，我們在 300 篇的文章中，共產生 13088 個關聯法則（rule）及 77277415 個 edge。並將結果存於如圖 5.4 所示的 edge.rul 檔案，以做為知識版圖的範圍。**

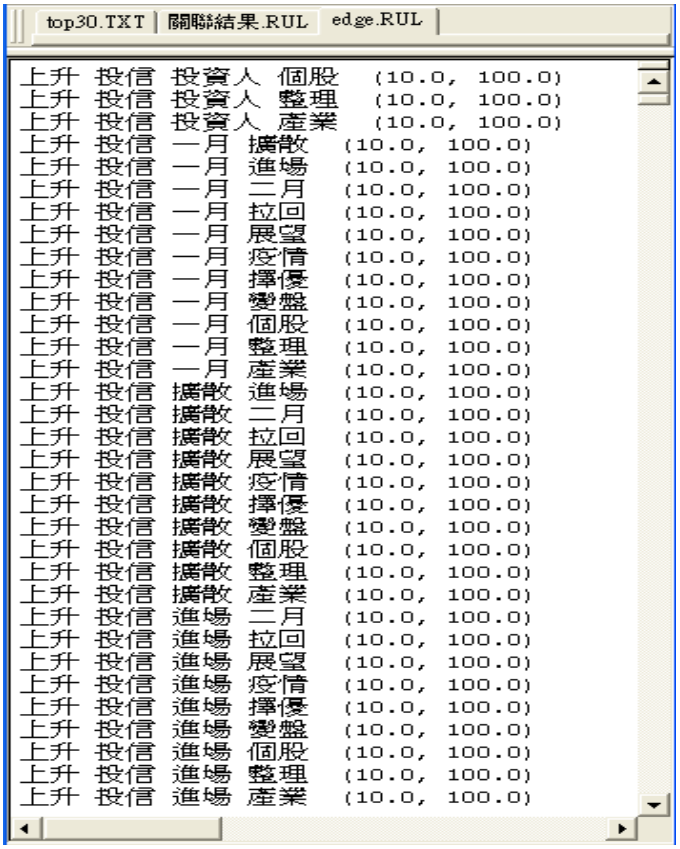


圖 5.4　hyperedge 的部分組成

上述圖 5.4 中，關聯 edge 的輸出格式中，每行皆顯示為 a b c ... (x%, y%)的輸出結果，其中 a，b，c 分別代表關聯 edge 的組成，x%代表 a，b，c 之間的平均信心指數，y%代表 a，b，c 條件下的 edge 強度。

我們保留 4 個以上的 edge 成份及以 40%的信心指數建立詞彙關聯，並且透過章節 4.3 的原理，將 8772 個詞彙及 8651 個 edge 組成知識版圖，我們建構如章節 4.6 的*.hgr 做為空間描述檔，切割過程如圖 5.5 所示。

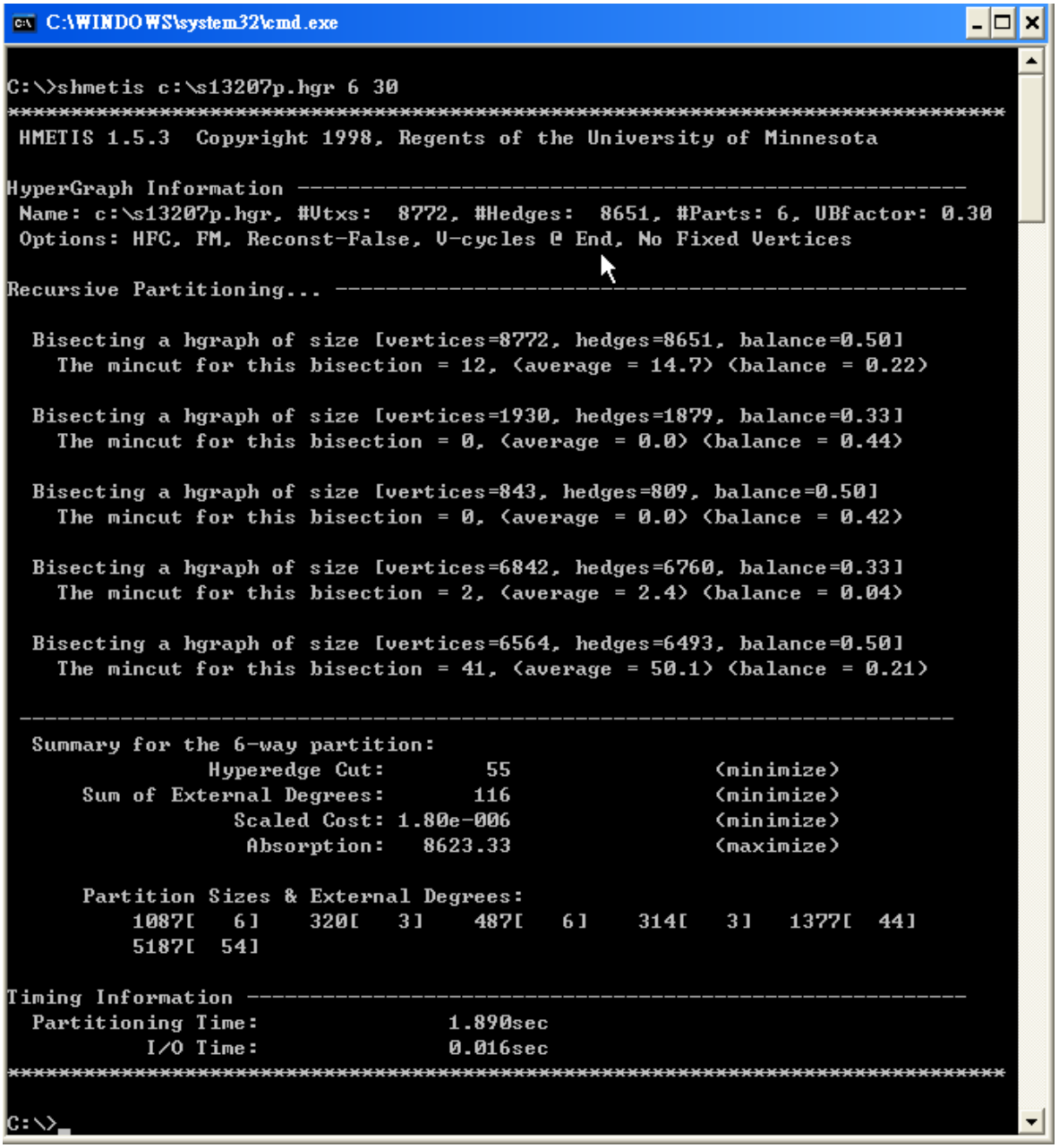


圖 5.5　實際切割時的過程畫面

5.3 實驗結果

我們以實驗檢視群聚的正確性及可能影響品質的詞彙標準。本論文透過召回率（Recall Rate）及精確率（Precision Rate）定義群聚的品質，其定義如下：

召回率=找到群聚 i 的文章總數/群聚 i 在系統的總文章數

精確率=(群聚類別被判定正確的數目+非群聚類別被判定正確的數目)/檢索過的文章總數

為了有效尋找不同類別的文章分布，本實驗以 2003 年「聯合新聞網」（http://udn.com/）「股市理財」領域之新聞文件為主要資料來源，其包含傳統產業、金融、科技等領域，我們隨意選取 300 篇文章做為實驗對象，並切割成 6 個不同的群聚。並以 5 位同學的群聚認知與實驗做結果比較。

由圖 5.6 及圖 5.7 的呈現，我們發現透過關聯權重及混合權重對群聚整理較有顯著成效。

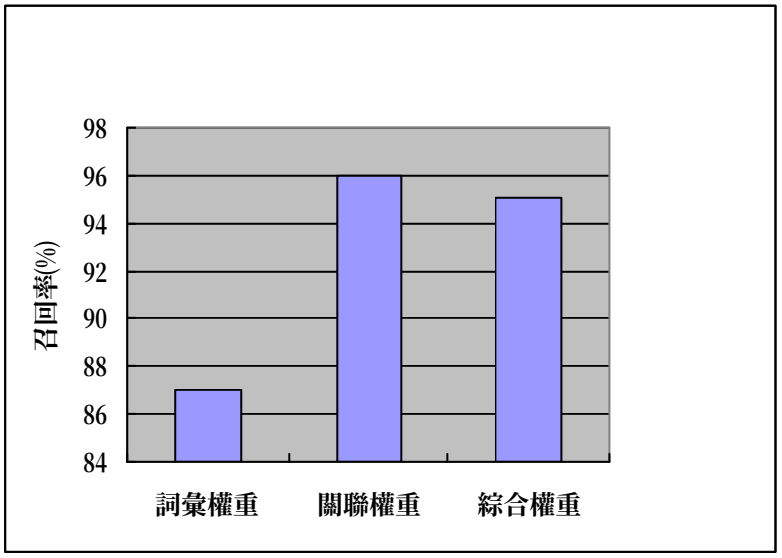


圖 5.6 三種不同切割標準的召回率比較

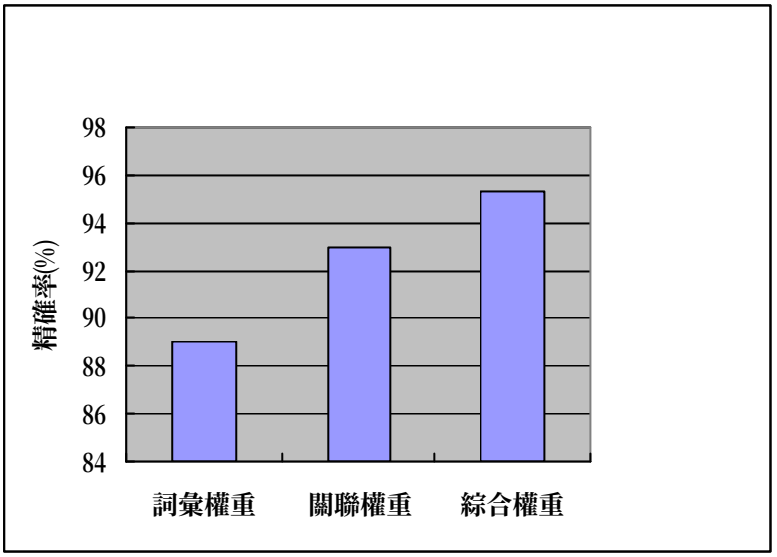


圖 5.7 三種不同切割標準的精確率比較

此外，在製作知識版圖的過程中，我們以 5 種不同的支撐值（Support）做為文章擷取比率。由圖 5.8 我們發現以每篇文章的 30%權重所做成的知識版圖，對群聚結果已可達成最佳化，特別是關鍵權重及綜合權重兩種。而超過 30%以上的詞彙權重對知識版圖並無顯著的幫助。

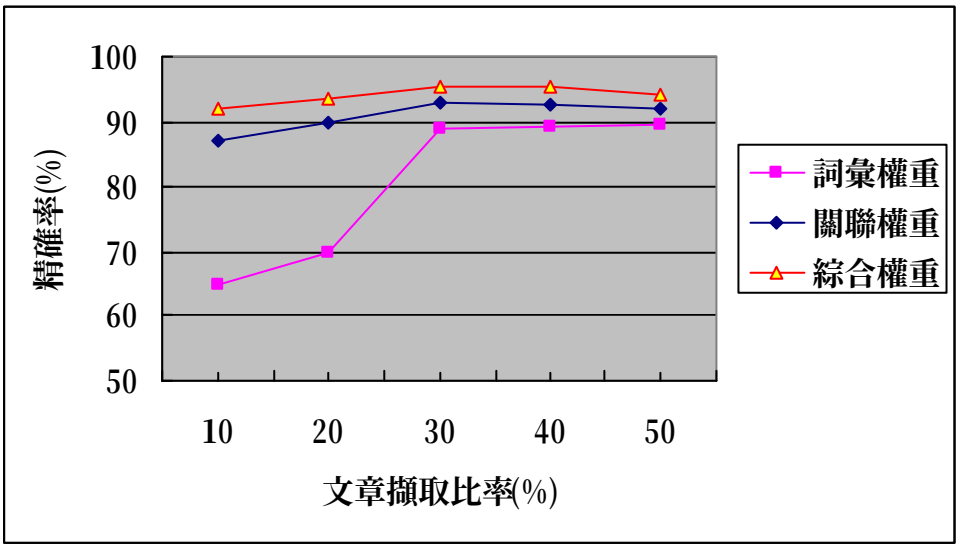


圖 5.8 不同的擷取比率對應精確率

5.4 方法比較

為了證明本研究方法的實用性，我們選擇以 k-nearest neighbor 及 centroid-vector

algorithm 做為方法的比較。這兩種演算法的優勢在於其具有簡單易懂；且可以透過現成免費的資源【18】提供我們現成實現建置的工具。

我們將每一文章視成多維陣列的組成，並將每一文章置入統一的多維陣列空間，透過 HyperGrpah Partition、K-Nearest Neighbor 及 Centroid-Vector 我們同樣切割成 15 個不同的群聚。並得到表 5.1 的結果，有關確信度(Certain)是以 5 位使用者透過人為審閱的方式及機器運算的結果，所得到的誤差結果。

表 5.1 三種不同的方法比較

Methoed	Certain	Uncertain	Total
Hypergraph Partition	94%	6%	100%
K-Nearest Neidgbor	84%	16%	100%
Centorid Vector	81%	19%	100%

第 6 章　結論與未來展望

我們利用關聯法則及切割方法，其發覺概念群聚底下的推測性表達，幫助了各類文章的歸納。該模型具有下述特點：

- **減少點閱的時間浪費，**
- **闡述每一群聚的關鍵概念，**
- **提供關鍵詞彙做為加速知識的預測建議。**

此外，透過共鳴式群聚法，我們得知其可以反映出『是否提及』，這作法就像是專利局裡頭的文章監控，隨時可以幫助我們在最少的時間找到所有的文章。故可以應用於科技發展或資訊整理。

本實驗作法上集中在概念的範圍拿捏及文章被選取的反應；日後如能結合時序與族群內的詞彙應該可以發展成學習地圖，給予求知者循序漸進的學習導引。另外，我們希望做到『提及程度』的判斷。畢竟『是否提及』和『提及程度』，對知識管理的領域有其存在的意義。

參考文獻

【1】 顏義樺，“以聯想法則概念網路為基礎之文章概念探索及相似性比對“，東海大學資訊工程與科學研究所碩士論文.2003

【2】 董振東、董強(1999)，《知網簡介》http://www.how-net.com

【3】 http://ckip.iis.sinica.edu.tw/CKIP/tool/，10 May,2004

【4】 http://ckip.iis.sinica.edu.tw/CKIP, Chinese Knowledge Information Processing Group, Academic sinica，10 May,2004

【5】 G. Karpis. hMETIS 1.5.3 http://www.cs.umn.edu/~karpis/metis/main.html,1998

【6】 Ｊoachims, T.(1998)：Text Categorization with support Vector Machines：Learning with Many Relevant Features, in：Proceedings of the 10th European Conference on Machine Learning, pp. 137-142.

【7】 King-lp Lin；Ravikumar Kondadadi, “A Similarity-Based Soft Clustering Algorithm for Documents”， Database Systems for Advanced Applications, 2001. Proceedings. Seventh International Conference on Volume , Issue , 2001 Page(s):40 – 47

【8】 Witold Pedrycz, “Associations and Rules in Data Ming: a Linkage Analysis” 2002 IEEE

【9】 Ian H. Witten, Eibe Frank “Data Mining-Practical Machine Learning Tools and Techniques with Java Implementations” Jun. 2003

【10】Hu Xiao，Wu Qinyi，and Zhong Yixin， “A Statistics Based Method of Mining Hierarchical Word Relation”.2001 IEEE

【11】G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar. “Multilevel Hypergraph Partitioning: Applications in VLSIDomain”, 34th Design Automattion Conference. Mar. 1997.

【12】Thorsten Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization”. March.1996.

【13】Alexandrin Popescul, Gary Willian Flake, Steve Lawrence, Lyle H. Ungar, C. Lee

Giles, “Clustering and Identifying Temporal Trends in Document Databases” IEEE Advances in Digital Libraries, ADL 2000, Washington, DC, May 22-24, pp. 173-182,2000

【14】Luhn, H.P, “The Automatic creation of literature abstracts. IBM Journal of Research and Development” 1958,2(2), pp.159-165.

【15】Prof. Dr. Thorsten Teichert, Marc-Andre Mittermayer, “Text Mining for Technology Monitoring” IEEE IEMC 2002

【16】Show-Jane Yen，Arbee L.P. Chen，“A Graph-Based Approach for Discoverying Various Types of Association Rules”，IEEE Trans. Knowledge and Data Eng.，pp.839-845，2001.

【17】http://www.borgelt.net/apriori.html，12.Jun,2008

【18】http://www.kdnuggets.com/，12.Jun,2008

【19】John Medina，《Brain Rules》，http://brainrules.blogspot.com/，12.Jun,2008