結合資訊檢索與分群演算法建構知識地圖

邱登裕
中華大學資訊管理學系

潘雅真
中華大學資訊管理學系

摘要

知識地圖(Knowledge Map, K-Map)是用來呈現知識分布的其中一個方法。在建立知識地圖之前，必須分析文件的內容，擷取其重要內容並建立關聯程度。本研究首先利用中研院所提供之 CKIP 中文斷詞系統，將文件內容進行斷詞及詞性標註，透過資訊檢索(Information Retrieval)、資料探勘(Data Mining)及分析等技術，將所擷取之重要特徵詞給予其代表性之權重，並進行相似度的計算與分群，最後將分析後的特徵詞與文件對應到知識地圖中的概念。

本研究實作了一個知識地圖系統，使用者可以藉由知識地圖的呈現方式，快速的找尋所需知識，並可更進一步觀察其相關知識的分布，促進組織內部知識的保留與共享，達到知識管理的目標。

關鍵字：知識地圖、資訊檢索、資料探勘

Combining Information Retrieval and Clustering Algorithm to Construct an Knowledge Map

Deng-Yiv Chiu
Department of Information Management, Chung Hua University

Ya-Chen Pan
Department of Information Management, Chung Hua University

Abstract

Knowledge map is one kind of technique to represent knowledge. Before building up a knowledge map, it is necessary to analyze related document. The two main purpose of this research are to extract the important content and to create the association level. First, CKIP system is used to identify Chinese words and to tag the morphological features. Secondly, techniques of information retrieval and data mining and analysis are used to weight importance of extracted keywords. Thirdly, similarities between documents are computed and similar documents are grouped together. Finally, extracted keyword and document are mapped to the concept of the knowledge map.

The research presents a Knowledge Map system. Through KMS, users can perform fast knowledge search and moreover they can observe the distribution of related knowledge. KMS can also be used to aid an enterprise in retaining and sharing inter-organizational knowledge to achieve the goal of knowledge management.

Keywords: Knowledge Map, Information Retrieval, Data Mining

壹、緒論

知識管理的概念漸漸為企業所重視，企業要成功的導入知識管理就必須提供組織成員一個良好且適當的介面。然而在實務上，組織內的部門必須提供企業相關的資訊，其中包含了會計部門與維修部門等。因此，如何協助組織成員在龐大的企業資訊中取得所需的相關資訊，是知識管理的首要目標。在現今企業環境下，企業內部的組織架構、部門人員、相關流程與企業規定等資料散佈在組織的每一個角落，造成組織成員在面對這麼多且龐雜的企業資訊時，必須耗費許多時間來尋找並取得其相關資訊。

一個良好的知識導引介面對現今企業內部之組織成員十分重要，透過適當的資訊科技輔助下，組織成員可以快速的取得所需之相關資訊，提升組織成員工作的效率，強化企業之整體形象與競爭力。

本研究以運用良好的知識地圖介面來提供組織成員相關且完整的企業資訊內容為目標。更具體而言，即是透過資訊檢索(Information Retrieval)的技術來擷取企業組織中各部門之特徵描述，在此基礎下結合資料探勘(Data Mining)之分析過程，建立組織成員之相關需求對應到的需求資訊，以利組織成員可更專注於工作上，全心全意為企業效力，強化企業競爭力。

一個良好的企業知識管理之管理介面，將有助於企業內部主管除了可以了解企業本身的知識分布之外，亦可清楚的看出企業在市場中的競爭力。然而，透過良善的資訊取得介面，將可以協助組織成員快速地找尋知識並確認知識的所在，以加速組織成員在其所任職之單位可以工作的更加順暢，既可提昇企業行政效率，亦可強化組織成員對企業之忠誠度。

貳、文獻探討

在此研究中，我們應用資訊檢索以及資料探勘的概念來完成整個研究的流程，其中包含了資料的前置處理動作、特徵詞的探勘、特徵詞的選取、相似度的計算、分群演算法、關聯式規則以及知識的呈現方式。除此之外，我們將所擷取到的知識，透過知識管理領域中之知識地圖的方法予以呈現。

目前在知識管理領域中，建置知識地圖是呈現企業內部知識的另一種形式。大部分探討知識地圖的文獻都以管理面的角度來說明知識地圖的管理架構(Knapp,1998)，而實際上真正建置出知識地圖之研究，其多半提出利用現有之資訊科技技術，建立其知識地圖的模式(Fu-ren Lin; Chih-ming Hsueh 2002)。由文獻中可以清楚的知道，建置知識地圖所使用的相關概念不外乎是找尋文件內容之關聯性，而系統之實際建置的情況，研究文獻仍然相當貧乏。

不過，知識地圖的建置的確可以為企業帶來很多競爭優勢，因為它可以提供使用者輸入他們的需求，以知識地圖的呈現方式快速的找出知識的所在位置(Devenport ;

Prusak 1998)，具體的呈現存在於企業的知識並以直覺化的方式呈現資訊的取得和關係，可以在不同的層級中由不同的背景提供者分享知識，以達到知識共享的目的(Vail 1999)，同時也可以讓企業高階主管可以全盤的掌握企業資產的分布情形，透過對組織本身的評估以及競爭者的評估，可為組織本身之策略作一有效的調整，為組織帶來競爭優勢(Tiwana 2000)。

從資訊檢索的觀點來看，知識地圖可被定義為對社群有貢獻之概念，以特徵式的方式進行文件分類(Fu-ren Lin; Chih-ming Hsueh 2002)。為了協助虛擬社群達到知識管理的目標，利用知識地圖呈現其概念階層，建立其關聯程度以解決問題之流程。同時，知識地圖的維護也是相當重要的，如何將新進文件漸進式的加入至已存在之知識地圖中，可利用現有技術予以解決。

總而言之，知識地圖主要是找出知識之間彼此的關係，將此關係以視覺的方式呈現，對於來自不同領域使用者，促進其溝通與學習。透過此知識地圖的指引，找到所需的知識，節省尋找資料的時間，讓企業內部員工能夠更專注於企業的營運，建立一套良好的組織策略。

參、方法論

本研究之精神，就是要整合企業內部文件資源，將原本散布於各部門的文件資料經過蒐集與整理，透過現有資訊科技之應用，建立企業式知識地圖，如圖 1 所示，在此研究中可分為下列三個主要的階段：

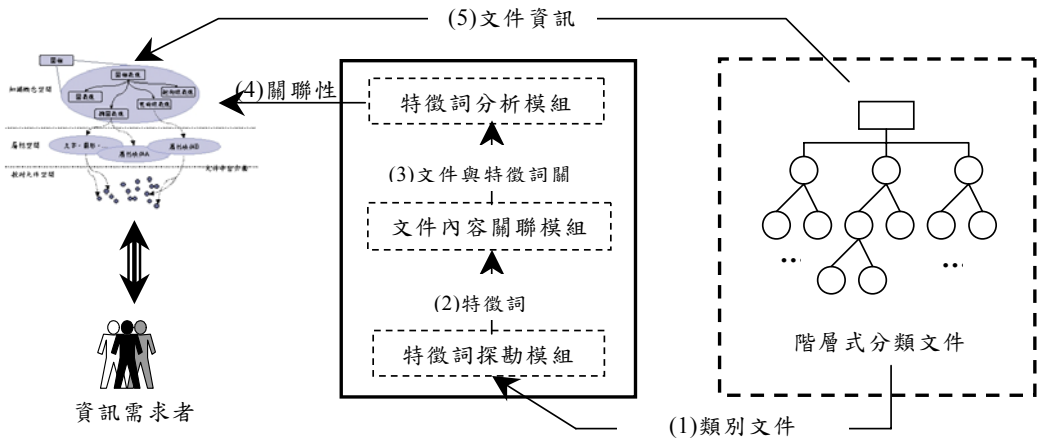

圖 1：知識地圖之資訊系統

一、特徵詞探勘模組：利用檢索技術來蒐集具有代表性之特徵詞(步驟(2))；

二、文件內容關聯模組：利用資料探勘之技術找出特徵詞間的關聯性(步驟(3))；

三、特徵詞分析模組：分析特徵詞並依據文件內容關聯與特徵詞關聯描繪知識地圖(步驟(4))。

使用者便可利用其知識地圖搜尋介面，快速精準的釐定其目標文件(步驟(5))，進而取得與目標文間相關的文件。

一、特徵詞探勘模組

本研究之企業文件主要是以中文文件為主，而所取得之階層資料為 2 或 3 層，在前置處理方面，首先將所蒐集到的資料進行中文斷詞系統斷詞處理，接著從斷詞的階層式文件中建立其詞性合併規則以找出有意義之詞彙，最後透過卡方檢定找尋其具有代表性之特徵詞彙以建立向量空間模組(Vector Space Model, VSM)之基底。

（一）斷詞處理

在此研究中，我們採用中央研究院詞庫小組所研發的 CKIP 中文自動斷詞系統，來處理本研究的中文文件資料。舉例來說：『長期大量飲酒不但可以直接傷害胃壁，而且可以引起脂肪肝、肝炎、肝纖維硬化，甚至肝硬化或肝癌等病變。』。經由 CKIP 斷詞系統處理過後，變成：『<sentence>長期(N)大量(DET)飲酒(Vi)不但(C)可以(ADV)直接(Vi)傷害(Vt)胃壁(N)，(COMMACATEGORY)</sentence><sentence>而且(C)可以(ADV)引起(Vt)脂肪肝(N)、(PAUSECATEGORY)肝炎(N)、(PAUSECATEGORY)肝(N)纖維(N)硬化(Vi)，(COMMACATEGORY)</sentence><sentence>甚至(ADV)肝硬化(N)或(C)肝癌(N)等(POST)病變(N)。(PERIODCATEGORY)</sentence>』

CKIP 中文自動斷詞系統會自動標記了斷詞過後之詞彙特性。在此，我們利用 CKIP 中文自動斷詞系統詞性的標記，擷取想要的詞性。例如我們將 N(普通名詞、專有名稱、地方詞、位置詞、時間詞)以及 Vi(動作不及物動詞、動作類及物動詞)留下，則上述例子只剩下『長期(N)飲酒(Vi)直接(Vi)胃壁(N)脂肪肝(N)肝炎(N)肝(N)纖維(N)硬化(Vi)肝硬化(N)肝癌(N)病變(N)。』等詞彙。

（二）擷取有意義的特徵詞

接下來將斷詞處理過後的詞彙作進一步的處理，透過 CKIP 中文自動斷詞系統所標記的詞彙特性，進行詞性合併的動作，以找出具有意義之特徵詞彙。此一步驟是參考目前文獻(林厚誼，蔣岳霖，周世俊，2002)整理之合併詞性規則，並依據本研究之文件內容所整理的詞性合併規則，將前述的斷詞結果進行詞性合併，以擷取有意義之特徵詞。以下是歸納出來之規則及合併順序：

1. N+N+N+N。　例如：神經(N)　功能(N)　障礙性(N)　疾病(N)
2. N+N+N。　例如：春暉(N)　醫(N)　星球(N)
3. N+N。　例如：皮膚(N)　搔癢症(N)
4. N+Vi。　例如：長期(N)　飲酒(Vi)
5. N。　例如：中藥(N)

此 5 條規則中，必須符合(特徵詞間沒有頓號之情形)，第 4 條是參考目前文獻(林厚誼；蔣岳霖；周世俊 2002)，剩餘 4 條規則是依據本研究之文件內容需要，經由統計所加入之規則。經由此過程所擷取之特徵詞可建立有意義之特徵詞，我們可以大大的減少無意義之詞性，以利後續問題之研究。

（二）代表性之特徵詞

特徵詞探勘模組的主要功能是從各文件類別的文件中，探勘出具有代表性之特徵詞，其方式乃是從眾多文件中擷取有意義之特徵詞彙，再判斷其特徵詞彙是否具有代表性或是鑑別力。本研究是基於階層式組織架構文件來建立知識地圖，資料來源必須是階層式的文件架構，並從階層式的組織架構中來擷取較具代表性之特徵詞。

在特徵詞擷取技術方面，以卡方檢定、資料增益量及文件頻率等處理方式，比起詞彙強度、交互資訊的文件分類準確度要高(Yang ;Pedersen 1997)。其中以卡分檢定與資訊增益量用於文件分類，都有很好的效果。本研究將結合 TFIDF 以及卡分檢定變形的方法——相關係數，來鑑定詞彙是否具有代表性。若詞彙之相關係數高於設定之門檻，將此特徵詞彙設定為具有代表性之特徵詞彙，並歸納為此類別的特徵詞。

本研究利用階層式架構可以用來計算卡方檢定的係數，而卡方檢定的公式就必須建立在階層式文件分類下才可以計算，其主要特色在於將屬於此類別與不屬於此類別的文件數予以找出，以鑑別出與其它類別之差異性的特徵詞。利用此方式可以找出足以代表此類別的特徵詞，是一個非常適合為此階層式文件找出具有代表性特徵詞的方法。若單純的使用 TFIDF 方法找出特徵詞，我們將無法為某個特定的類別找出與其它不同類別之差異性特徵詞，唯有在階層式架構下才可以藉由卡方檢定之方法找出某特定類別之代表性特徵詞。

1. **相關係數**(Correlation coefficient)

相關係數為卡方檢定的一種形式，主要是鑑別一詞彙 T 對某一類別 C_j 之相關性，卡方值愈高則代表此詞彙與該類別愈相關。計算方式如公式(1)所示：

$$CC(t,C_j)=\frac{\sqrt{N}(AD-CB)}{\sqrt{(A+C)(B+D)}\sqrt{(A+B)(C+D)}} \quad (1)$$

C_j：某一類別。

t：詞彙。

N：所有類別底下之文件篇數。

A：在類別 C_j 中包含 t 之文件篇數。

B：在不屬於 C_j 類別中包含 t 之文件篇數。

C：在 C_j 類別中不包含 t 之文件篇數。

D：在不屬於 C_j 類別中不包含 t 之文件篇數。

經由卡方檢定之計算，在 C_j 類別下所有的文章之特徵詞可分成四種類型，如表 1 所示。例如，從類型I中可以看出，特徵詞 t 在 C_j 類別出現頻率高，在其兄弟類別出現

頻率低，則代表特徵詞 t 在 C_j 類別具有高度代表性。從類型II中可以看出，特徵詞 t 在 C_j 類別出現頻率高，在其兄弟類別出現頻率亦高，則代表特徵詞 t 在 C_j 類別不具代表性。

經過上述卡方檢定的過程，我們會在各類別下找出具有代表性之特徵詞彙，且具有不同強度等級之代表強度。唯有在類別 C_j 中具有區別能力($CC(t,C_j)$ 高)，且具有足夠代表性(A↑,C↓)之特徵詞彙會有較高的強度。此方法在面對階層式文件時，更能夠顯示出類別之間特徵詞之差異程度，以探勘出適切的特徵詞彙。

表 1：卡方檢定分析表

類型	類別 C_j	不屬於類別 C_j	意義	代表性程度
I	機率高 即 A↑,C↓	機率低 即 B↓,D↑	在同一類別中具有高度出現率；在兄弟類別中具有低度出現率。	$(AD-CB)$ 高，則 $CC(t,c_j)$ 高
II	機率高 即 A↑,C↓	機率高 即 B↑,D↓	在同一類別中具有高度出現率；在兄弟類別中具有高度出現率。	$(AD-CB)$ 低，則 $CC(t,c_j)$ 低
III	機率低 即 A↓,C↑	機率低 即 B↓,D↑	在同一類別中具有低度出現率；在兄弟類別中具有低度出現率。	$(AD-CB)$ 低，則 $CC(t,c_j)$ 低
IV	機率低 即 A↓,C↑	機率高 即 B↑,D↓	在同一類別中具有低度出現率；在兄弟類別中具有高度出現率。	$(AD-CB)$ 低，則 $CC(t,c_j)$ 低

2. TFIDF(Term_Frequency * Inverse_Document_Frequency)

在計算文件間相似度之前，首先要為每一份文件建立向量基底(Base)。此研究利用 TF(Term Frequency)與 IDF(Inverse Document Frequency)的觀念擷取適合特徵詞做為文件之向量基底，找出其敘述性與代表性的關係。表示式如公式(2)和公式(3)所示：

$$D_i=\left(w_{i0},w_{i1},w_{i2},...,w_{ij}\right) \tag{2}$$

D_i：第 i 篇文件。

w_{ij}：代表特徵詞 t_j 出現在文件 d_i 之權重，其計算方式如公式(3)。

$$w_{ij}=t_jf*\log id_if \tag{3}$$

t_jf：詞彙 t_j 在文件 d_i 中出現的次數。

id_if：(所有文件篇數)/(詞彙 t_j 在所有文件中出現過的篇數)。

在此利用 TFIDF 之計算方式做為詞彙出現在文件的權重，會加上 log 只是想要把 id_jf 值降低，避免計算出來的值過大。

二、文件內容關聯模組

在第二階段中，其主要目的在於找出文件間內容之關聯性。在進行文件特徵詞探勘之前置處理之後，接下來就是利用所探勘出來的特徵詞計算文件間內容關聯性。此部份將以向量空間模組進行相似度計算，找出有關聯之文件，最後利用向量空間模組之相似度之值來進行分群。

以向量空間模組的方式對少量的資料作分類，不但快速、有效而且分類的效果也很好。但是如果採用此方法應用於大量資料時，將會形成高維度的向量基底，使得計算量非常的大。在此研究中，為了找出更精準之相似度文件，我們仍然採用前置處理過後之所有特徵詞進行文件之相似度計算。首先，先建立文件之向量，當計算文件在文件基底(亦即特徵詞)之權重後，利用餘弦夾角(Cosine)來測量文件的相似度，以進行文件分群，若兩文件所出現之特徵詞大致相似，則代表它們越相似，即可將兩份文件分在同一群。如公式(4)所示，可用來計算文件 d_1 與文件 d_2 的概念相似度。

$$\cos\left(\overline{d_1},\overline{d_2}\right)=\frac{\overline{d_1}\cdot\overline{d_2}}{\left|\overline{d_1}\right|\times\left|\overline{d_2}\right|} \tag{4}$$

文件分群之演算法如下所示。首先先將每一筆文件資料視為一個群組 (步驟一)，然後在兩兩群聚之間，找出距離最近的兩個群組 (步驟二)，距離計算方式即是利用公式(4)找出群聚間，兩兩文件之間相似度的總和。之後將相似度總和最大的兩個群組合成一個群組，然後一直重複步驟二、步驟三的步驟，直到最終的群數少於所設定的群數，即可停止。

步驟一：將每筆資料視為一個群聚 C_i；

步驟二：找出所有群聚間，距離最接近的兩個群聚 C_i、C_j；

步驟三：合併 C_i、C_j 成為一個新的群聚；

步驟四：如果目前的群聚數目多於預期的群聚數目，則反覆步驟二至步驟四，直到群聚數目少於預期的群數。

進行文件分群後，我們將以資訊檢索最常用之衡量方法，為每個分群評估其效能，分別有準確度(P_{s,p,C_i})、回覆率(R_{s,p,C_i})以及 F-measure($F_{s,p}$)。其計算方式如公式(5)(6)(7)所示。

$$P_{s,p,C_i}=\frac{N_{p,C_i}\cap N_{s,C_i}}{N_{s,C_i}} \tag{5}$$

$$R_{s,p,C_i}=\frac{N_{p,C_i}\cap N_{s,C_i}}{N_{p,C_i}} \tag{6}$$

$$F_{s,p}=\frac{2P_{s,p,C_i}R_{s,p,C_i}}{P_{s,p,C_i}+R_{s,p,C_i}} \tag{7}$$

N_{s,C_i}：被系統分類至 C_i 類別之文件篇數。

N_{p,C_i}：被人工分類至 C_i 類別之文件篇數。

$N_{s,C_i} \cap N_{p,C_i}$：共同被人工分類至 C_i 類別之文件與被系統定義為 C_i 類別之文件的文件篇數。

$F_{s,p}$：F-measure 為調和平均數，組合了準確度以及回覆率兩種指標。

三、特徵詞分析模組

建構知識地圖的第三個階段就是利用第一階段所建立的特徵詞庫，找出其特徵詞關聯性。而此階段的方法必須建立一個基本的假設：「若多個特徵詞同時出現在一份文件中，則它們必定存在其關聯性。」。

首先，將目前文件內所擷取之特徵詞放入演算法之搜尋引擎中，與其它文件之特徵詞比對並計算其關聯強度，利用 DHP 演算法進行比對，找出最佳之特徵詞集合，並計算其關聯強度，其系統流程如圖 2 所示。

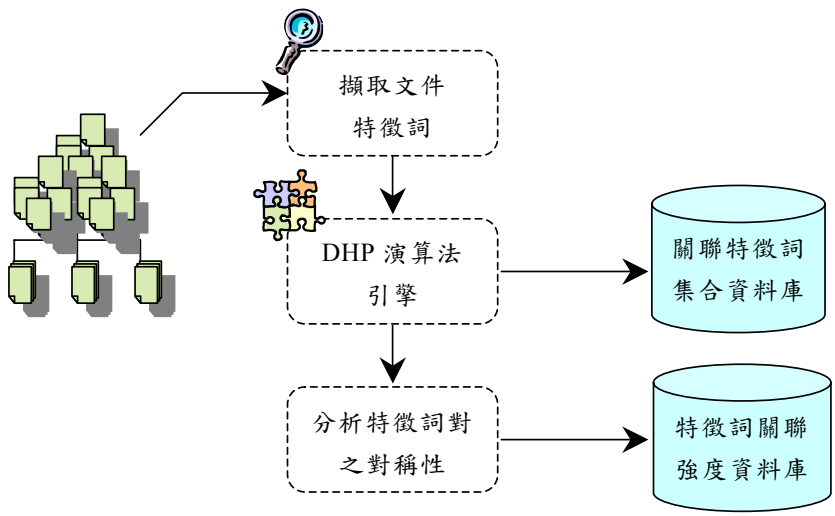

圖 2：特徵詞截取架構

（一）DHP 關聯分析法

我們使用 Apriori 演算法所衍生出來的 DHP 演算法找出特徵詞之關聯性，與 Apriori 不同的是，DHP 演算法加入了 hash table 的架構，改善以往在關連法則挖掘的效率，演算法如表 2 所示。

表 2：DHP 之演算法

輸入：(1)信賴值之門檻值 α
(2)支持度之門檻值 S
(3)資料庫中每一篇文章所擁有之特徵詞 f
輸出：滿足信賴值與支持度之特徵詞集合
步驟：
針對每一篇文章 D_i：
(1)$F \leftarrow \{f \mid f$ 之信賴值 $\geqq \alpha$ 以及 f 之支持度 $\geqq S\}$
(2)針對每篇文章 D_i 之特徵詞：
(2.1)產生 $\left(f_j, f_k\right), j \neq k$
(2.2)利用 Hash Function 建立其 Hash Table
h({x y})=((order of x)*10+(order of y))Mod 7
(2.3)產生 F-Set$\{(f_j, f_k)$,j≠k $\mid (f_j, f_k)$符合信賴值與支持度, $f_j, f_k \in F\}$
(3)針對 F-Set 內之特徵詞：
(3.1)產生候選特徵詞彙集
(3.2)產生符合信賴值與支持度之特徵詞集合

分析特徵詞關聯的主要目的是從資料庫中之所有文章中，找出具有關聯性之特徵詞對。將每一篇文章符合信賴值與門檻值之特徵詞集合起來(步驟(1))，接下來將每一個特徵詞一一配對(步驟(2.1))，計算特徵詞對的出現機率，並建立其 hash table(步驟(2.2))，若特徵詞對的出現機率高於設定之信賴值與門檻值(步驟(2.3))，則將此特徵詞對歸納為具有關聯性之特徵詞並記錄下來。

（二）分析特徵詞對之對稱性

在計算特徵詞對之關聯性之後，即取得兩兩特徵詞之間的關聯程度，然後再針對特徵詞對之關係做進一步的分析，找出其關聯之方向性。

設定兩兩特徵詞間的關聯程度門檻值可以發現，其特徵詞對之關係可分為兩種：高度關聯性(關聯程度高於門檻值)和低度關聯性(關聯程度低於門檻值)。再從一組關聯(考慮兩兩特徵詞集合的大小)來分析，其特徵詞的文件集合大小不一，產生了關聯的相對對稱問題(孫振凱，2002)，可歸納為兩種，分別是對稱式的關聯問題與非對稱式的關聯問題。我們將針對此兩種問題分別說明如下。

1. 對稱式的關聯性問題

從圖 3 中的示意圖所示，可以很清楚的看到，具有特徵詞 A 與具有特徵詞 B 的文章數量差異不大，所以 A∩B 不管是對特徵詞 A 或者是特徵詞 B 來說，其重要性是對稱的。但重要性的程度則要視 A∩B 交集的大小比例而定，其特徵詞 A 與特徵詞 B 的關聯程度如公式(8)所示。

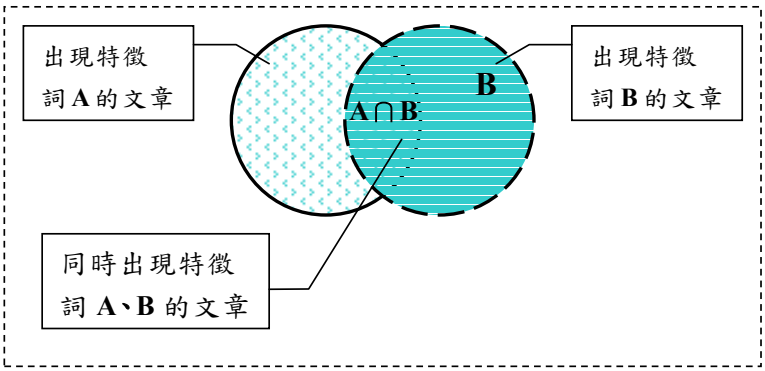

圖 3：具有 A、B 特徵詞之文章數量對稱式之關聯示意圖

$$\frac{N(A\cap B)}{N(A)} \cong \frac{N(A\cap B)}{N(B)} \tag{8}$$

N(A)：擁有特徵詞 A 之文件篇數。
N(B)：擁有特徵詞 B 之文件篇數。
N(A∩B)：擁有特徵詞 A 與特徵詞 B 之文件篇數。

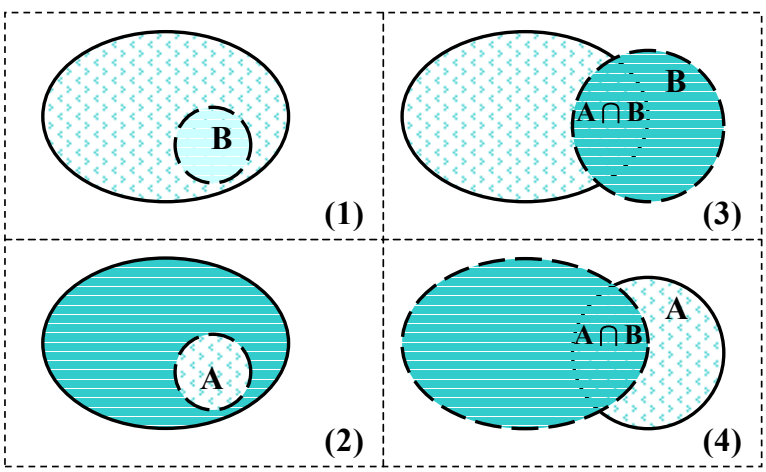

圖 4：A、B 特徵詞之文章數量非對稱式關聯示意圖

2. 非對稱式的關聯性問題

從圖 4 中的示意圖所示，可以很清楚的看到，具有特徵詞 A 與具有特徵詞 B 的文章數量差異很大，所以 A∩B 對特徵詞 A 或者是特徵詞 B 來說，其重要性是相當不對稱的。但重要性的程度則要視 A∩B 交集的大小比例而定，其特徵詞 A 與特徵詞 B 的關聯程度如公式(9)所示。

$$\frac{N(A\cap B)}{N(A)} << \frac{N(A\cap B)}{N(B)} \quad 或 \quad \frac{N(A\cap B)}{N(A)} >> \frac{N(A\cap B)}{N(B)} \tag{9}$$

N(A)：擁有特徵詞 A 之文件篇數。

N(B)：擁有特徵詞 B 之文件篇數。

N(A∩B)：擁有特徵詞 A 與特徵詞 B 之文件篇數。

從對稱式之關聯性與非對稱性之關聯性的分析中，必須還要考慮的就是特徵詞間比較的重要性關聯程度。從圖 4 來說，可以看出特徵詞 A、B 間的影響程度有三種：

(1) A 對 B 的影響程度相近於 B 對 A 的影響程度。

(2) A 對 B 的影響程度遠勝於 B 對 A 的影響程度。

(3) B 對 A 的影響程度遠勝於 A 對 B 的影響程度。

從圖 5 中可以看到以特徵詞 w0 為例之關連性矩陣，可以明確的看出有出現特徵詞 w0 的文件有 307 篇，有出現特徵詞 w1 的文件有 822 篇，與 w0 和 w2 共同相關的文件有 62 篇，這即是 A 矩陣中每一個元素之意義。

A =

	w0	w1	w2	w3	w4
w0	307	41	62	30	12
w1	41	822	123	7	35
w2	62	123	1003	1	12
w3	30	7	1	72	3
W4	12	35	12	3	102

圖 5：以特徵詞『w0』為例之關聯矩陣

接下來分析矩陣中兩兩元素之特徵詞交集的數量，藉此可以了解特徵詞間相互之重要性，轉換後的矩陣以[B]來表示其相對關聯矩陣，而矩陣[B]的數學定義如下所示：

Define [B]
bij = aij/aii *100%
其中 aij 代表特徵詞 i 與特徵詞 j 共同出現之文件篇數;
aii 代表特徵詞 i 出現之文件篇數。

以特徵詞 w0 為例，初使矩陣[A]經過公式轉換後可以得到相對矩陣[B]。由圖 6 矩陣中觀察 w1 和 w4 的關係，可得到 b14=4.3%和 b41=34.3%。由 b41=34.3%可得知，特徵詞對 w1-w4 比特徵詞對 w4-w1 佔較多的比例。

B =

	w0	w 1	w 2	w 3	w 4
w0	100%	13.4%	20.2%	9.8%	3.9%
w1	5.0%	100%	15.0%	0.9%	4.3%
w2	6.2%	12.3%	100%	0.1%	1.2%
w3	41.7%	9.7%	1.4%	100%	4.2%
w4	11.8%	34.3%	11.8%	2.9%	100%

圖 6：以特徵詞『w0』為例之相對矩陣

肆、實驗

本研究之實驗資料是從網頁 “Yahoo!”上之『健康』 類別下之階層式文件作為實驗對象，此類別下共分為 6 個類別，分別有醫藥館、舒壓館、健身館、飲食館、美容館以及兩性館，共分為 6 個子類別以及 6 個子類別又有各自的子類別。

網頁 “Yahoo!” 上的服務對象主要為一般使用網路搜尋資料的民眾，透過網頁的方式將相關資料分享給大眾，所以絕大部分皆以文字陳述來描述，約佔所有資料的 96%。本研究之實驗資料收集期間為 2004 年 12 月 1 日至 2005 年 3 月 16 日，共 992 筆『健康』相關文件。以下將針對實驗過程及實驗結果分別介紹：

一、實驗過程

本研究之實驗過程是依據方法論中之三個主要的模組進行實驗，分別為特徵詞探勘模組、文件內容關聯性模組以及特徵詞分析模組，以作為知識地圖繪製的依據。以下將分別描述：

（一）特徵詞探勘模組

為了找出具有代表性以及敘述性的特徵詞，首先要將文件進行斷詞處理，以便找出其特徵詞。在此，由於實驗資料來源採用中文文件，故我們利用中央研究院詞庫小組所研發的 CKIP 中文自動斷詞系統，來處理本研究的中文文件資料。

接下來將中文斷詞所得到的結果，利用方法論中所提出的特徵詞彙擷取規則，進行特徵詞之擷取，如此可以降低無意義之特徵詞彙，其詳細過程如方法論中之特徵詞擷取之描述。

由表 3 所示，可以看到每一個類別下所擷取到的特徵詞數量，其中可以發現，文件數量越多的類別下所擷取到的特徵詞彙就越多。主要原因是因為文件內容經由斷詞之後，文件內容越多所對應到的特徵詞彙擷取規則的機率就越大。

表 3：利用特徵詞擷取規則所擷取到之特徵詞數量類別比較表

類別	1_1	1_2	1_3	2	3_1	3_2	4_1	4_2_1	4_2_2	4_2_3	4_2_4	4_2_5	4_2_6	4_2_7	4_2_8	4_2_9	4_2_10	4_2_11	4_2_12	5_1	5_2	6_1	6_2_1	6_2_2	6_2_3	6_2_4
文件數	70	13	32	51	169	42	36	7	12	10	42	25	22	17	10	27	23	26	49	21	22	72	9	5	6	22
特徵詞數	169	21	17	54	132	63	62	10	10	14	28	8	11	12	13	14	11	11	33	38	30	74	0	3	1	6

此研究之資料結構為階層式文件，故在此可利用卡方檢定以找出具區別能力之特徵詞彙。此部分將藉由設定卡方門檻值，以系統來判定符合此類別的特徵詞。由圖 7 可看出，在不同卡方值限制下，每個類別符合卡方值之特徵詞個數。舉例來說，在類別 1_1 中，雖然擁有很多特徵詞彙，但其區別能力不佳。亦即卡方值越高，其區別能力越佳。

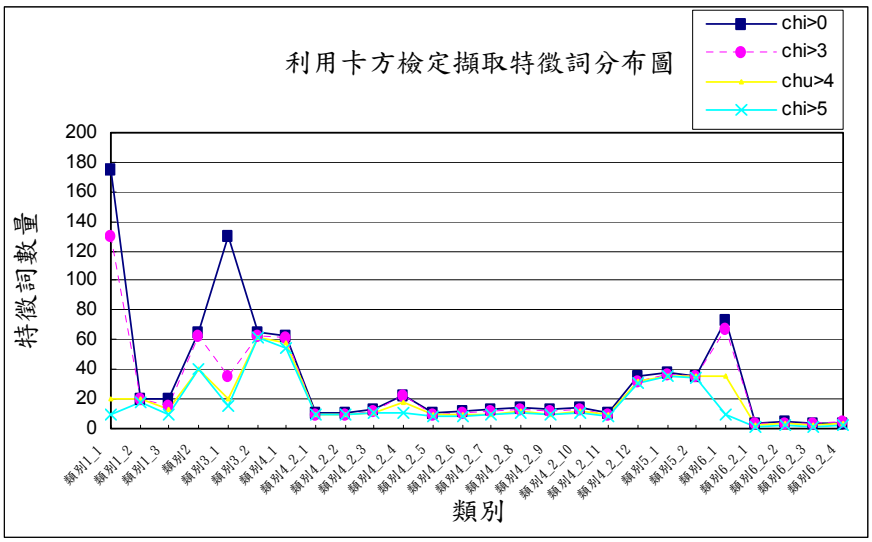

圖 7：、利用卡方檢定所找到的特徵詞數量直方圖

（二）文件內容關聯模組

在文件內容關聯性部分，所欲探討的內容為將文件內容進行文件相似度計算、特徵詞相似度計算以及文件分群等相關文件關聯性處理動作。

1. 文件相似度計算

實驗中以 992 筆文件內容進行相似度計算，每一份文件將以向量空間模組的方式進行文件與文件的相似度計算，以前置處理過後所擷取之特徵詞作為向量空間模組之基底。表 4 為文件分群的部分結果，表中分別記錄文件所屬之類別、文件_1、文件_2

以及其相似度之值。實驗結果清楚的看出，同一個類別之文件其相似度之計算結果皆較佳。由此可知，此研究之實驗資料來源仍具有些許代表性，可以代表一般階層式之文件。

不過，仍有一些文件與同類別下之文件完全沒有相似度，亦即相似度之值為 0，故為何歸為此類別之文件，判定為可能是我們所蒐集到的文件數量不足，導致沒有辦法作完整之相似度計算。

相似度計算方式為(文件_1 與文件_2 之外積)/{(文件_1 距離平方和)* (文件_2 距離平方和)}。以編號 1 為例，文件_1 距離平方和為 65.44，文件_2 距離平方和為 22.35，文件_1 與文件_2 之外積為 333.12，得到相似度＝(333.12)/(65.44*22.35)=0.23。

表 4：文件內容相似度計算之部分結果資料

編號	類別	文件_1	距離平方和	類別	文件_2	距離平方和	外積	相似度
1	3_2	001. 三大乳霜介紹.txt	65.44	3_2	004. 怎麼調配與使用晶露？.txt	22.35	333.12	0.23
2	3_2	001. 三大乳霜介紹.txt	65.44	3_2	007. 晶露的種類和功能.txt	110.46	666.25	0.09
3	3_2	001. 三大乳霜介紹.txt	65.44	3_2	011. 補充美麗的情緒能量-疲勞、筋疲力竭篇.txt	14.90	222.08	0.23
4	3_2	001. 三大乳霜介紹.txt	65.44	3_2	014. 月經失調.txt	24.34	222.08	0.14
5	3_2	001. 三大乳霜介紹.txt	65.44	3_2	023. 頭髮也要感性-夏日洗髮專題.txt	50.57	555.40	0.17
6	3_2	001. 三大乳霜介紹.txt	65.44	3_2	025. 眼睛保養.txt	86.63	444.16	0.08
7	3_2	001. 三大乳霜介紹.txt	65.44	3_2	031. 搶救問題皮膚.txt	31.92	222.08	0.11
8	3_2	001. 三大乳霜介紹.txt	65.44	3_2	036. 美顏魔法：卸妝.txt	95.20	222.08	0.04
9	3_2	001. 三大乳霜介紹.txt	65.44	3_2	037. 精油真偽經驗談.txt	24.34	222.08	0.14
10	1_1	001. 皮膚搔癢症的中醫治療.txt	31.73	1_1	012. 人體的元氣可以測量嗎？.txt	44.84	594.59	0.42

表 5：特徵詞相似度計算之部分結果資料

編號	特徵詞_1	距離平方和	特徵詞_2	距離平方和	外積	相似度
1	一氧化碳中毒	61.61	子女	21.24	1021.87	0.78
2	力銀行	21.24	外食族	21.24	451.23	1.00
3	十字花科	31.86	奶粉	21.24	676.85	1.00
4	三角眼熟	21.24	三角	21.24	451.23	1.00
5	下半身纖細	31.86	冬瓜	37.26	575.85	0.49
6	下半身纖細	31.86	竹筍	31.86	1015.28	1.00
7	下半身纖細	31.86	材料	81.74	793.32	0.30
8	下半身纖細	31.86	春暉醫星球	77.95	270.59	0.11
9	下半身纖細	31.86	香菜	27.21	613.12	0.71
10	下半身纖細	31.86	麻油	31.30	575.85	0.58

2. 特徵詞相似度計算

實驗中以前置處理過後所擷取之特徵詞進行相似度計算，每個特徵詞將以向量空間模組的方式進行特徵詞與特徵詞的相似度計算，以找出每一個特徵詞之分佈情況是否相似。表五為特徵詞分群之部分結果，表中記錄特徵詞_1、特徵詞_2 以及相似度之值。

相似度計算方式為(特徵詞_1 與特徵詞_2 之外積)/{(特徵詞_1 距離平方和)* (特徵詞_2 距離平方和)}。以編號 5 為例，特徵詞_1 距離平方和為 31.86，特徵詞_2 距離平方和為 37.26，特徵詞_1 與特徵詞_2 之外積為 575.85，得到相似度＝(575.85)/(31.86*37.26)=0.49。

從表中之數據可以知道，如果兩兩特徵詞相似度之值越高，代表此兩兩特徵詞所分佈之文章位置相似，由此可知兩兩特徵詞彼此之關聯性非常高。

3. 文件分群

在此部分我們先分為兩個部分，第一部份為針對文件相似度結果所對應到之各個類別進行分群，第二部份為針對所有的文件進行分群，透過文件相似度之值進行文件分群，若文件相似度之值為 0，則不予進行分群之演算法。由實驗過程中得出，其分群結果完全相同。

表 6：階層式文件分群類別分群數比較表

類別	1_1	1_2	1_3	2	3_1	3_2	4_1	4_2_1	4_2_2	4_2_3	4_2_4	4_2_5	4_2_6	4_2_7	4_2_8	4_2_9	4_2_10	4_2_11	4_2_12	5_1	5_2	6_1	6_2_1	6_2_2	6_2_3	6_2_4
文件數	70	13	32	51	169	42	36	7	12	10	42	25	22	17	10	27	23	26	49	21	22	72	9	5	6	22
群數	19	2	1	4	39	9	6	0	1	2	2	1	1	1	2	0	0	0	6	5	0	9	0	0	0	1

由表 6 中可以清楚的看出，某類別下之文件內容越相似，則所分的群數就會比較少。舉例來說，在屬於類別 Cls7_1 與 Cls7_2 中，進行分群時所分的群數比較多，此乃因為此類別名稱屬於新知類別，故文件內容非常不相似所導致之結果。而屬於類別 Cls3_1“美容專欄”之文件內容較相似，故所分群數就比較少。從表中可以大致看出每一個類別的分群篇數以及分的群數為何。以類別 1_1 為例，有相似度之值的文件共有 70 篇，分群結果共可分為 19 群。

（三）特徵詞分析模組

在分析特徵詞關聯性部分，所想要探討的內容為將每一篇文件所擁有之特徵詞，利用 DHP 關聯分析法、對稱關聯性以及矩陣表示法等方法進行處理。

1. DHP 關聯分析法

在此實驗部分，我們採用 DHP 演算法進行特徵詞彙之探勘。在實驗的過程中，我們發現所蒐集到的文件內容僅能進行到第五個回合，亦即最多可找到有 5 個特徵詞同時出現在兩篇文件中。

表 7 為 DHP 關聯分析法進行特徵詞搜尋之結果，表中記錄特徵詞_1、特徵詞_2、特徵詞_3、特徵詞_4、特徵詞_5、出現次數以及出現之文件檔名，由表中可以看出最多有 5 個特徵詞同時出現在兩篇文章檔名之中。

表 7：DHP 關聯分析法進行特徵詞搜尋之部分結果

編號	特徵詞_1	特徵詞_2	特徵詞_3	特徵詞_4	特徵詞_5	次數	文件檔名
1	皮膚	搔癢				2	098. 皮膚搔癢症的中醫治療.txt
2	皮膚	搔癢				2	175. 膚質過乾　少泡溫泉.txt
3	心肌梗塞	動脈硬化				4	001. 青豆腰果飯.txt
4	心肌梗塞	動脈硬化				4	002. 鮮蔬腰果炒.txt
5	心肌梗塞	動脈硬化				4	003. 腰果蝦仁.txt
6	心肌梗塞	動脈硬化				4	004. 通心粉沙拉.txt
1	乾燥	皮膚	乾性皮膚			2	105. 冬季乾燥皮膚的保健.txt
2	乾燥	皮膚	乾性皮膚			2	105. 冬季乾燥皮膚的保健.txt
3	腦中風	心肌梗塞	動脈硬化			4	001. 青豆腰果飯.txt
4	腦中風	心肌梗塞	動脈硬化			4	002. 鮮蔬腰果炒.txt
5	腦中風	心肌梗塞	動脈硬化			4	003. 腰果蝦仁.txt
6	腦中風	心肌梗塞	動脈硬化			4	004. 通心粉沙拉.txt
1	大蒜	芹菜	胡蘿蔔	洋蔥	開水	2	076. 七彩飲年後體內大掃除.txt
2	大蒜	芹菜	胡蘿蔔	洋蔥	開水	2	095. 年後減肥營養師推薦七彩飲.txt

表 8：特徵詞對對稱關聯性之部分計算結果

編號	特徵詞_1	特徵詞_2	特徵詞_1 對 特徵詞_2 的重要性	特徵詞_2 對 特徵詞_1 的重要性
1	化妝品	力高	0.50	1.00
2	卸妝徹底	力高	0.50	1.00
3	粉底	力高	0.50	1.00
4	三角眼熟	三角	1.00	1.00
5	春暉醫星球	下半身肥胖	0.01	1.00
6	範圍	下半身肥胖	1.00	1.00
7	纖細	下半身肥胖	1.00	1.00
8	升期	下降期	1.00	1.00
9	春暉醫星球	下疳	0.01	1.00
10	梅毒	下疳	0.50	1.00

2. 分析特徵詞對之對稱關聯性

我們分別針對前置處理過後所擷取之特徵詞，分析其兩兩特徵詞之對稱關聯性。此實驗部分主要找出特徵詞_1 與特徵詞_2 在文件分布之關聯重要性。從表 9 中可以看出特徵詞_2“力高”對特徵詞_1“化妝品”、“卸妝徹底”以及“粉底”極具重要性，亦即只要有出現特徵詞“化妝品”、“卸妝徹底”、“粉底”，就一定會出現“力高”，反過來出現的機率則為 0.5。

3. 關聯矩陣之呈現

為了呈現特徵詞之關聯矩陣，我們必須找出相關之特徵詞，以便建立特徵詞之間之關聯式矩陣關係。以特徵詞『大蒜』為例，與之最為相關的為特徵詞『果汁機』、『芹菜』、『胡蘿蔔』、『開水』以及『鳳梨』最為相關，從圖 8 中可以清楚的看出，特徵詞與特徵詞間共同出現之篇數。例如：特徵詞『大蒜』在資料庫中出現了 33 篇，特徵詞『大蒜』與『果汁機』共同出現的篇數為 2 篇，特徵詞『大蒜』與『芹菜』共同出現的篇數為 4 篇。

	大蒜	果汁機	芹菜	胡蘿蔔	開水	鳳梨
大蒜	33	2	4	8	3	3
果汁機	2	38	4	3	14	6
芹菜	4	4	23	7	5	2
胡蘿蔔	8	3	7	50	4	4
開水	3	14	5	4	70	7
鳳梨	3	6	2	4	7	15

圖 8：以特徵詞間『大蒜』為例之關聯式矩陣

接下來，將圖八之結果進行第三章所介紹之公式轉換，即可得到圖 9 的結果，從圖九中可以清楚的看出特徵詞與特徵詞之間的關聯程度。例如：特徵詞『果汁機』對特徵詞『大蒜』之重要性為 6.1%。

	大蒜	果汁機	芹菜	胡蘿蔔	開水	鳳梨
大蒜	100.0%	6.1%	12.1%	24.2%	9.1%	9.1%
果汁機	5.3%	100.0%	10.5%	7.9%	36.8%	15.8%
芹菜	17.4%	17.4%	100.0%	30.4%	21.7%	8.7%
胡蘿蔔	16.0%	6.0%	14.0%	100.0%	8.0%	8.0%
開水	4.3%	20.0%	7.1%	5.7%	100.0%	10.0%
鳳梨	20.0%	40.0%	13.3%	26.7%	46.7%	100.0%

圖 9：以特徵詞間『大蒜』為例之相對關聯式矩陣

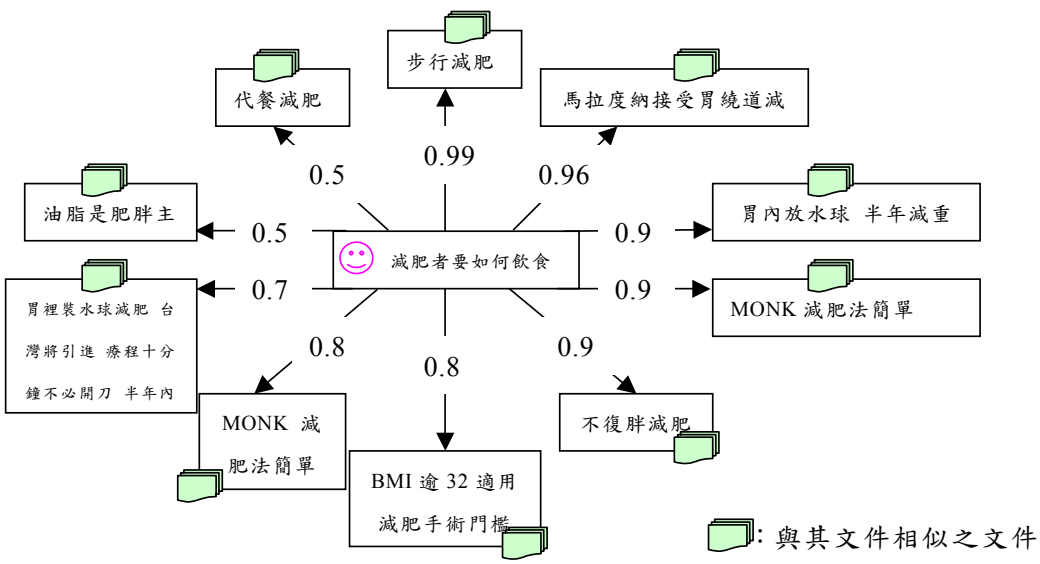

圖 10：文件相似度之知識地圖

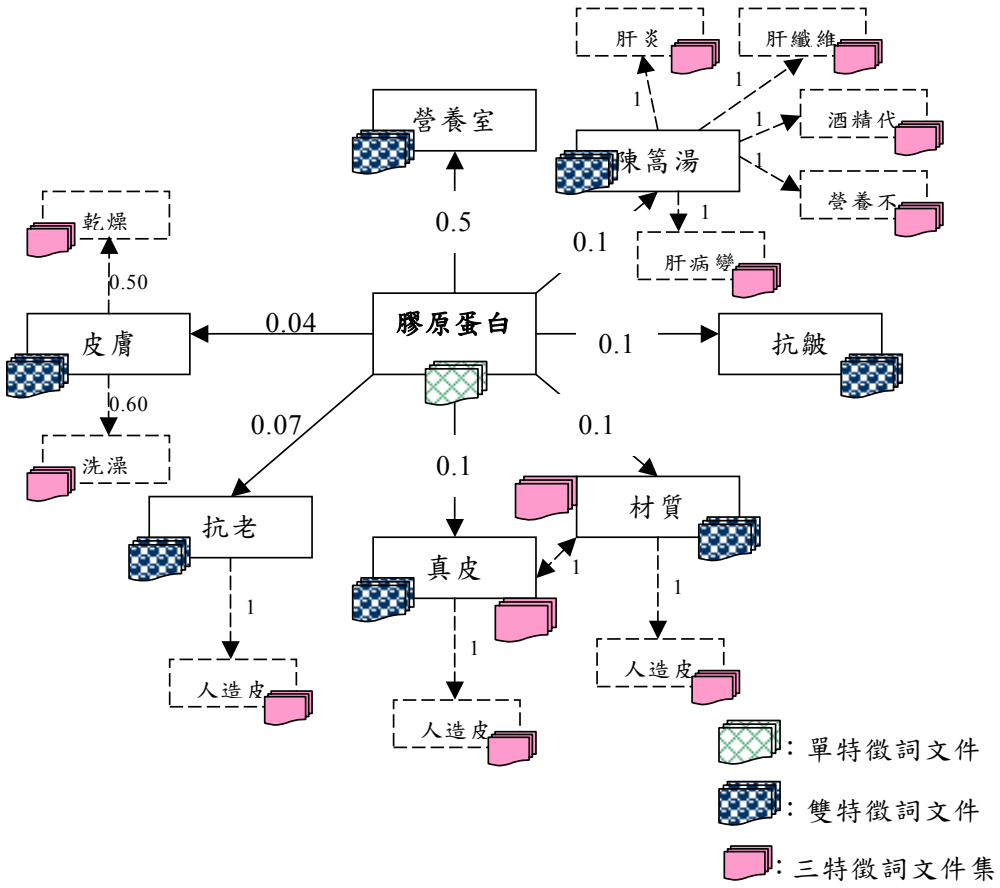

圖 11：特徵詞相似度之知識地圖

二、實驗結果

本研究之知識地圖繪製主要可分為三種形式，分別為階層式文件分類之知識地圖呈現、文件相似度之知識地圖呈現以及特徵詞相似度之知識地圖呈現。

（一）文件相似度之知識地圖

此部份所想要呈現的知識地圖，是利用計算文件相似度所描繪出來的。如圖 10 所示，首先讓使用者輸入想要尋找之目標文件，以“目標文件”為起始點，延伸出去之分支為具有高相似度文件，此知識地圖可以進一步慢慢找到其相關的文件。舉例來說，目標文件為“減肥者要如何飲食控制？”，系統便會自動產生與“減肥者要如何飲食控制？”相似之文件，並呈現文件相似度之值，以讓使用者可以清楚的知道該選哪篇文章，假設選取了“步行減肥法”，還可找尋與“步行減肥法”之相關文章。

（二）特徵詞相似度之知識地圖

此部份所呈現的知識地圖，是利用計算特徵詞相似度所描繪出來的。如圖 11 所示，首先讓使用者輸入想要尋找之相關特徵詞，以相關之“特徵詞”為起始點，延伸出去之分支為具有高相似度之相關特徵詞，此知識地圖可以進一步慢慢找到與特徵詞相關的文件。舉例來說，相關特徵詞為“膠原蛋白”，系統便會自動產生與“膠原蛋白”相關之文件以及相關特徵詞之文件，並呈現特徵詞相似度之值，以讓使用者可以清楚的知道該往知識地圖的哪個方向選取。

表 9：實驗結果紀錄

類別名稱	名稱	類別	網頁	系統	共同篇數	準確度	回覆率	F-measure
醫藥館	中醫天地	Cls1_1	286	908	109	12%	38%	18%
	癌症防治	Cls1_2	13	70	13	20%	100%	33%
	診療 DIY	Cls1_3	32	73	9	12%	28%	17%
兩性館	兩性館	Cls2	51	96	25	24%	48%	32%
美容館	美容專欄	Cls3_1	169	170	69	41%	41%	41%
	精油妙用	Cls3_2	41	153	42	27%	100%	43%
飲食館	健康飲食	Cls4_1	36	161	18	11%	50%	18%
	健康好菜	Cls4_2	270	266	134	53%	54%	54%
健身館	減重專欄	Cls5_1	21	17	8	47%	38%	42%
	運動專欄	Cls5_2	22	93	5	5%	23%	9%
舒壓館	舒壓專欄	Cls6_1	72	82	23	28%	35%	31%
	舒壓療法	Cls6_2	42	38	27	71%	64%	67%
平均						29%	52%	37%

在此知識地圖中，使用者利用系統所呈現之關聯特徵詞尋找文件。使用者可藉由單一特徵詞尋找其文件集，或者進一步尋找其雙特徵詞文件集，利用其關聯特徵詞找到使用者欲尋找之文件，或者進一步尋找其三特徵詞文件集，藉由此方法可讓使用者節省大量時間。

（三）實驗結果評估

此研究利用網頁已分類之結果與系統判斷之結果作比較，以衡量本研究相似度計算以及分群之結果。本研究採用在資訊檢索領域中用來評估效能的標準衡量指標—準確度、回覆率以及 F-measure 來評估本系統結果的準確性(Van,1979)。

我們分別針對每一個類別下，由網頁已經分類好的文件進行比較。在前置處理部分，我們已經找出每一個類別代表性之特徵詞，在此我們將採用卡分值大於 7.5 之特徵詞代表系統所找到之文件。

我們將此系統之準確度、回覆率以及 F-measure，呈現於表 9 中，可以看出在準確度部分，類別 6_2 最高為 0.71，即搜尋出的文件中，每 10 篇文件有 7.1 篇是使用者要

尋找的文件，其次是類別 4_2，其值為 0.53。在回覆率部分，最高為類別 1_2、3_2，其值為 1.0，即使用者欲搜尋之文件，皆有被搜尋出來。而 F-measure 為調和平均數，它結合了以上兩種指標，其最終結果為類別 6_2 表現最好，其值為 0.67，類別 4_2 次之，其值為 0.54。

由實驗結果可看出本系統判定與人工判定之準確度及回覆率，並與其它分群演算法進行比較。如表 10 所示，C4.5 與 CN2 為資料探勘之分群演算法，其分群演算法亦採用卡方檢定之方法擷取其具有代表性之特徵詞彙來進行分群，C4.5 所採用之準確度及回覆率分別為 68.31%與 77.61%;而 CN2 之準確度及回覆率分別為 63.00%與 79.02%。

本研究在準確度與回覆率之值分別為 29%及 52%。由此可知，在人工所判定的文章很多在系統判定時有 52%被搜尋出來，在準確度的表現上有些許差距，其差距原因主要是因為本研究所採用之實驗資料為非專業領域之資料，且大部分專業領域之資料已經有專業用語或慣用語法，故在特徵詞的擷取中比較容易且特徵詞較具其代表性。本研究試圖將此方法應用在一般文件中，雖然準確度及回覆率較低，但本研究較具有普遍性，且在某些類別中也產生不錯的效果。

表 10：實驗結果比較

	準確度	回覆率
C4.5	68.31%	77.61%
CN2	63.00%	79.02%
本研究	29.00%	52.00%

本研究的實驗資料為中文文件，試圖從自然語言中擷取具有意義的特徵詞是非常不容易且費時間的。本研究之實驗所採用的中文文件為一般入口網站的文件，此中文文件隱含了非常多的雜訊，與其它相關研究相比，其它知識地圖相關之研究所採用的實驗資料本有較好的特徵詞關聯性。而本研究所建立的詞性合併規則已大大減少沒有意義之中文特徵詞，其中很多所擷取到的特徵詞還是不夠具有其代表性，由此可知中文文件的處理非常的不簡單。若以本實驗數據跟其它相關研究相比，本研究之內容是有其參考價值的。

伍、結論與展望

在本研究中的資料前置處理部份，試圖以卡方檢定之方式，先找出具有代表性之特徵詞彙；在文件關聯模組部份，先進行特徵詞之分群，再利用向量空間模組之方式計算文件與文件間之相似度，以階層式聚合分群法之平均連結聚合演算法方式進行文件之分群；在分析特徵詞關聯性部分，進行特徵詞與特徵詞間對稱與非對稱之分析與比較，並利用資料探勘之 DHP 方法找出特徵詞間之關聯程度，以在進行知識地圖繪製時，得到良好之效果。

本研究之特色在將階層式文件經過資訊檢索和資料探勘方法的處理，建構一個知識地圖。本研究主要結論與研究貢獻分別整理如下：

一、結合資訊檢索之方法與資料探勘的分群，將向量空間模組之相似度計算結果，作為階層式聚合分群法分群之依據。此方法與架構說明資訊檢索與資料探勘可被廣泛的聯合運用。

二、針對一般文件進行分析與關聯處理，改善過去以階層式呈現方式之文件集。

三、建立數條特徵詞擷取規則，以擷取出文件內容中較具有意義之特徵詞。而這些規則也可應用於企業內部文件中，考慮企業經營的特性，建立其相關規則。

四、結合學術理論與實際應用之研究，所提出的系統架構與方法，將可有效應用於企業內部部門階層文件的分類中，建構一個屬於企業本身之知識地圖，可提供企業內員工快速尋找其相關文件，協助企業內部主管在進行決策制定時，可節省大量時間與成本，提升企業競爭力。

五、現今大部分研究研究均採用專業領域之資料進行實驗，本研究試圖在非專業領域之資料進行實驗，即利用網際網路之階層式文件為實驗資料，建立其文件相關性，其結果顯示出本研究之方法論可以為一般文件建立其知識地圖，讓網際網路使用者亦可快速找出相關文件。

於未來與研究展望部分，本研究整理出下列幾點：

一、將此方法論應用於企業階層式文件中，可為企業建立良好的知識管理系統，提昇企業競爭力。

二、未來可以針對本研究所使用之文件分群演算法進一步探討，計算出分群的品質，並建立漸進式分群。

三、文件是源源不絕的，所以要如何在動態的新增文件的情況下，建立一個動態的知識地圖，是非常重要且值得持續探討的。

致謝

本研究承蒙中華大學補助，計畫編號 CHU-94-M-010，僅此誌謝。

參考文獻

1. 中央研究院詞庫小組，CKIP 中文斷詞系統，http://godel.iis.sinica.tw/CKIP/ws/.
2. 中文詞之事庫小組技術報告，1993，中文詞類分析。
3. 孫振凱、民 91，利用網頁建構知識分布圖，國立中山大學資訊管理研究所碩士論文。
4. Baker, L. D.; McCallum, A. K. “Distributional clustering of words for text categorization,” Proceeding of ACM SIGIR international conference on information retrieval, 1998, pp:96-103
5. Davenport, T.; De Long, D.; Beers, M. “Successful knowledge projects,” sloan

management review, Vol. 39, 1998, pp:43-57
6. Fu-ren, L.; Chih-ming, H. ”Knowledge map creation and maintenance for virtual communities of practice,” Information Processing & Management, 42(2), 2006, pp:551-568
7. Knapp; E. M. “Knowledge management,” Business & economic review, 1998
8. Tiwana, A. “The knowledge management toolkit : practical techniques for building a knowledge management system,” New Jersey, Prentice, 2000
9. Vail, E. F. III “Knowledge mapping: getting started with knowledge management,” Information system management, Vol. 16, No. 4, 1999, pp:16-23
10. Yang, Y.; Pedersen, J. O. “A comparative study on feature selection in text categorization,” Proceedings of 14th international conference on machine learning, 1997, pp:412-420
11. Van- Rijsbergen, C. J. ”Information Retrieval,” Butterworths, 1979