大學圖書館
12 卷 1 期（民國 97 年 3 月） 頁 96-110

針對特定領域知識分類架構進行導引式分群演算法之研究

The Design of Guided Concept Clustering Algorithm for Specific Knowledge Domain

葉建華 **Jian-hua Yeh**
真理大學資訊工程學系助理教授
Assistant Professor, Department of Computer Science and Information Engineering, Aletheia University
E-mail: au4290@email.au.edu.tw

【摘　要】

本文所探討的焦點是設計出一套自動化的運作模型，讓各種機構能夠快速藉由資訊系統學習網路上的相關本體認知架構資訊，並且依照自身的需求與分類，建立適用於自身發展的概念性架構（concept hierarchy，或稱知識分類架構）。這樣的運作模型，能夠將多個既存的分類架構，如美國政府入口網站（US Government Portal）、LookSmart、開放目錄專案 ODP（Open Directory Project）和 Yahoo 等等，以自動化的網頁擷取程式（Web spider），收集各個本體認知架構下相關的網頁以及附加資訊。然後依照網頁所存在的階層架構以及網頁內容，進行特徵學習。然而，本文提出的方法將採用不同於傳統的分類學習方式，希望可以藉由使用者表達的預設需求，進行使用者所提供內容的導引式分群。這樣的運作方式，是將大量已經分類過的資訊，藉由特徵資訊的萃取，作為特定領域的專業註解，使得分類結果更能夠貼近使用者的期望。

【Abstract】

This paper discusses the automatic concept hierarchy generation process for specific knowledge network. Traditional concept hierarchy generation uses hierarchical clustering to group similar terms, but the result hierarchy is usually not satisfactory for human recognition. Human-provided knowledge network presents strong semantic features, but the generation process is both labor-intensive and inconsistent in a large scale hierarchy. The method proposed in this paper combines the results of specific knowledge network generation and automatic concept hierarchy generation to produce a human-readable, semantic-oriented hierarchy. This

臺灣大學學術期刊資料庫

generation process can efficiently reduce the efforts of manual classification, an exhausting task for human beings. An evaluation method is also proposed in this paper to verify the quality of the result hierarchy.

關 鍵 詞：知識分類架構；階層式分群；樹狀結構相似度
Keywords：Concept hierarchy; Hierarchical clustering; Tree similarity

壹、前言

長久以來，人類產生知識分類架構的方式，一直是以大量的人力手動產生為主，其中主要的原因是在於人類具有有效產生語意架構的能力。但是隨著近代真實世界中大量新概念的產生，以及伴隨著日益複雜的概念關聯特性，使得人類越來越不容易產生大型的知識分類架構。這是因為隨著架構逐漸增大，人類不再有能力持續保持產生語意架構的一致性。無法保持一致性的問題會導致知識分類架構失去其應有的品質與公正性，因此在近年來，特別是網際網路興起之後，人類手動產生大型知識架構便不再具有其適用性。以網頁分類架構（Web directory）為例，目前已經無法單純以大量的人力來進行維護（maintain）與擴充（span），而必須輔以適當的資訊技術做為前置處理，以降低後續人工的處理負擔與問題。筆者之前所參與的知識分類架構建置計畫（Chen, Yeh, & Sie, 2005），相對地，試圖運用傳統的資訊技術，以資料聚合演算法（data clustering algorithm）來解決知識分類架構的建構議題。但這樣的方法同樣地無法適當的產生合適的語意架構。本文便以解決上述問題為焦點，提出了一個稱之為 GCC（Guided Concept Clustering）的導引式演算法，以降低人類進行架構分類的負擔，同時又可以自動地建構出具合理與一定品質的語意結構出來。

貳、背景與相關議題討論

雖然目前已經有部分建構完成的大型知識分類架構，但針對特殊的知識領域，知識分類架構仍有建置的必要性。同時由於維護知識分類架構的成本高昂，且人工建置容易引起不一致或是錯誤的情形，這些問題都影響了特定領域知識分類架構的品質，因此越來越多相關研究致力於發展出客觀且正確的自動演算方式，來建構知識分類架構。

知識分類架構可以有很多種建構方式，它可以從無到有的方式產生，或是由既有的其他知識分類產生，或是由特定資訊來源的語料或辭典（Chuang & Chien, 2002, 2003a, 2003b;

Kashyap, Ramakrishnan, & Rindflesch, 2003）中產出，甚或是合併後兩種方式產出。有許許多多不同程度的自動化演算方式被發展出來，部分也引用自機器學習和自動文件分類相關演算方式（Ghani, Slattery, & Yang, 2001; Li & Yang, 2003; Yang, 1999, 2001; Yang & Liu, 1999; Yang & Pedersen, 1997;Yang, Zhang, & Kisiel, 2003; Zhang & Yang, 2003），或是近年來 Web 環境中所發展出的相關技術（Chakrabarti, Dom, & Indyk, 1998; Yang, Slattery, & Ghani, 2002）。以目前的研究進度來看，此類技術大都可應用在範圍受限且較小規模的認知架構上，以階層式的建構方式進行之（Palmer, et al., 2001）。

一般而言，產生知識分類架構的方法，可以大致歸納成以下幾種：1. 由上而下的方式，由最一般性的概念漸次細分為特定的概念（如KACTUS知識分類）；2. 由下而上的方式，由基本的概念逐漸匯整成為一般化的概念（如企業知識分類架構）；3. 由內而外，也就是由最重要的概念為出發點，分別發展一般化以及特殊化的概念層次。以下我們列舉一些在知識分類架構領域中具重要性的研究：

一、InfoSleuth（MCC）

InfoSleuth 是一個 MCC（Microelectronics and Computer Technology Corporation）的研究計畫，其目的是在發展出一個發掘公司與外部網路各項資訊的技術，焦點在於解決資訊定位、評估、擷取、合併等等問題，最主要的建構目標則是在於產生一個以知識分類為基礎的軟體代理人架構。InfoSleuth 的知識分類建構程序（Hwang, 1999）如下：1. 首先由領域專家提供一個集合的種子詞彙集合，用以表達較高層次的概念（也就是較一般性的概念），且相關於此詞彙集合的Web文件也會被擷取進來；2. 系統接著將含有種子詞彙的文件取出，並產生相關聯的概念詞彙片語，將其放置於知識分類中正確的位置上，同時為下一輪程序產出新的種子詞彙集合，這項程序會不斷反覆至特定的滿足狀況；3. 針對之式分類中的相關連概念定義關聯屬性，如 “is-a”、“part-of”、“manufactured-by”、“owned-by”等等；4. 在每一個反覆的流程中，領域專家都會介入確保產出概念與關聯的正確性，並進行適度必要的修改。

目前這個系統已經有許多應用，但也有一些尚待改進的問題，如語意結構上的模稜兩可問題、同義不同形的片語辨識問題、詞義問題、以及異質性文件的處理問題等等。

二、SKC（史丹福大學）

史丹福大學的 SKC（Scalable Knowledge Composition）計畫的目標在於發展出一個可以解決資訊系統中語意差異性的演算方法。此計畫試圖推導出一個可以應用於各個專門領域

臺灣大學圖書館

的知識分類整合方法。這個研究計畫是由美國軍方所支持（United States Air Force Office of Scientific Research/ AFOSR 與 United States Defense Advanced Research Project Agency/ DARPA）。本計畫中部分研究（Jannink, 1999; Jannink & Wiederhold, 1999）甚至將韋氏詞典的資料轉為圖狀結構（graph structure），用以支援產出知識分類架構。

本計畫中主要的進行方式為：1. 計算圖狀結構中節點間的關聯（arc）；2. 透過一個 ArcRank 演算方式過濾關聯值；3. 將關聯延伸為更一般性的關係；4. 反覆整個程序至滿足為止。這個演算方式將會輸出一個根據關聯程度強弱產生的一個節點集合。此演算法的問題也大都在於文字描述的錯誤與語法結構錯誤對產出的影響上。

除了以上的研究計畫之外，尚有許多相關的知識分類工程研究，大多與知識分類的自動學習相關，如 Karlsruhe大學AIFB的知識分類學習研究（Maedche & Staab, 2000, 2002）、ECAI2000 workshop（Wagner, 2000）、德州大學奧斯丁分校的歸納邏輯等等。這些相關計畫都是以半自動或是全自動產生知識分類架構為主。綜合以上相關議題，本文所介紹的導引式分群演算法，就是以如何應用部份文件自動分類架構的處理技術，達到本體認知架構的產生，而所藉由的方式，將放在少量人工介入建立導引性資訊，配合上統計以及機器學習機制，來進行有效語意的認知架構產生工作。

參、導引式分群演算法的設計

因此本計畫的目標是發展一個導引式的分群方法（Guided Concept Clustering, GCC）。其中包含用以參考的文件集合，抽取其中我們所認定的特徵（features），然後進行一種特別改良的階層式聚合演算法（Hierarchical Agglomerative Clustering, HAC）來達到分群的目的。在設計這樣的演算法時，我們將整合利用兩種不同的資訊作為我們認定的特徵。第一是傳統研究中所使用的內容相似度（Similarity）的資訊，第二是由既存架構所學習得來的分類觀點資訊。在內容相似度資訊的基礎上，長時間以來各種研究已經證明，由下而上的階層式聚合演算法仍然是最容易在只有資料的情況下，用以建立分類架構的分群演算法。因此我們在設計導引式的分群方法時，也將以此方式為出發點。

階層式聚合演算法（hierarchical agglomerative algorithm）由樹狀結構的底部開始層層聚合。一開始我們將每一筆資料視為一個群聚（cluster），假設我們現在擁有 n 筆資料，則將這 n 筆資料視為 n 個群聚，亦即每個群聚包含一筆資料：

1. 將每筆資料視為一個群聚Ci

臺灣大學圖書館期刊資料庫

2. 找出所有群聚間，距離最接近的兩個群聚Ci、Cj
3. 合併Ci、Cj成為一個新的群聚
4. 假如目前的群聚數目多於我們預期的群聚數目，則反覆重複步驟 2 至 4，直到群聚數目已將降到我們所要求的數目。

階層式聚合演算法的計算結果，我們以圖1為例。

在步驟 2 中，所謂的「距離最接近的兩個群聚 Ci、Cj」的定義，事實上有各種不同的定義方式，每一種方式所得到的結果都不太相同。這些常用的群聚距離的定義可以說明如下。

1. 單一連結聚合演算法（single-linkage agglomerative algorithm）：群聚與群聚間的距離可以定義為不同群聚中最接近兩點間的距離：

$$d(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b)$$

2. 完整連結聚合演算法（complete-linkage agglomerative algorithm）：群聚間的距離定義為不同群聚中最遠兩點間的距離：

$$d(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a, b)$$

3. 平均連結聚合演算法（average-linkage agglomerative algorithm）：群聚間的距離則定義為不同群聚間各點與各點間距離總和的平均（其中，n_{Ci} 表示Ci的資料個數。）：

$$d(C_i, C_j) = \frac{1}{n_{C_i} n_{C_j}} \sum_{a \in C_i, b \in C_j} d(a, b)$$

4. 沃德法（Ward's method）：群聚間的距離定義為在將兩群合併後，各點到合併後的群中心的距離平方和（其中，μ 表示Ci∪Cj的平均值）

$$d(C_i, C_j) = \sum_{a \in C_i \cup C_j} \| a - \mu \|^2$$

由於群聚距離直接影響了群聚的結果，因此要如何將更具語意的資訊成分，整合進階層式聚合演算法，便成了本研究中導引式分群演算法的重要焦點議題。基於上述觀察，我們希望將更具語意的分類資訊，整合進分

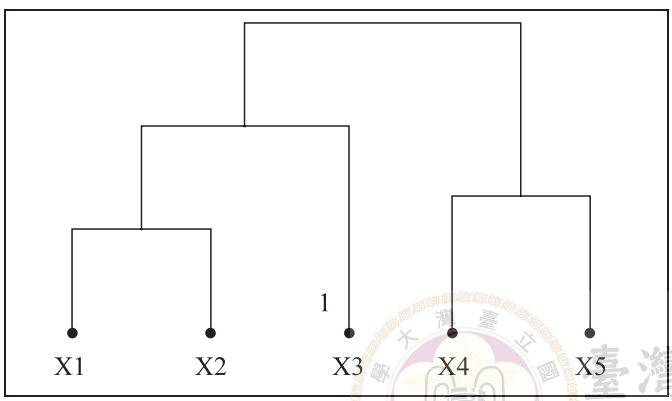


圖 1　階層式聚合演算法的結果

臺灣大學學術期刊資料庫

群演算法的設計中。網路上可獲取更具語意的分類資訊可以分成三類形態。第一種是網路上既存的分類網站（Web Directories），如 LookSmart、Yahoo、ODP 等。第二種是其他網頁參考的連結文字（anchor text）。第三種是在其他網站上的相關文字說明（如meta tags等等）。若依分類資訊的強烈程度來區分，則以第一種資訊，也就是既存的分類網站架構所學習得來的分類觀點資訊，最具有分類語意性質。

我們將利用搜尋引擎或是網頁收集程式（Web spider）從網路上取得上述的資訊。以此種分類網站的標記資訊為例，我們將分類網站的標記資訊（分類目錄）與分類目錄下的網頁建立關聯，這個步驟我們稱之為標籤資訊擷取程序，過程如圖2所示。

筆者所提出的方法是在傳統的資料聚合演算過程中，設計引入一個投票程序（voting procedure）。假設我們要針對某個概念 C 產生一個知識分類架構，而我們所擁有預備分類的資料集為 D，這個資料集取自其他既有分類架構的文件資料。如果我們的資料取自 k 個既有的分類架構 $\{N_1, N_2, \cdots, N_k\}$，其各自擁有的文件資料集稱為 $\{DS_1, DS_2, \cdots, DS_k\}$，我們可以將每個文件集進行分類資訊標籤資訊擷取，如圖 2 所示。而在所有文件資料集的標籤資訊擷取程序完成之後，我們便擁有了大量帶有架構標籤資訊的文件資料。接下來，我們便

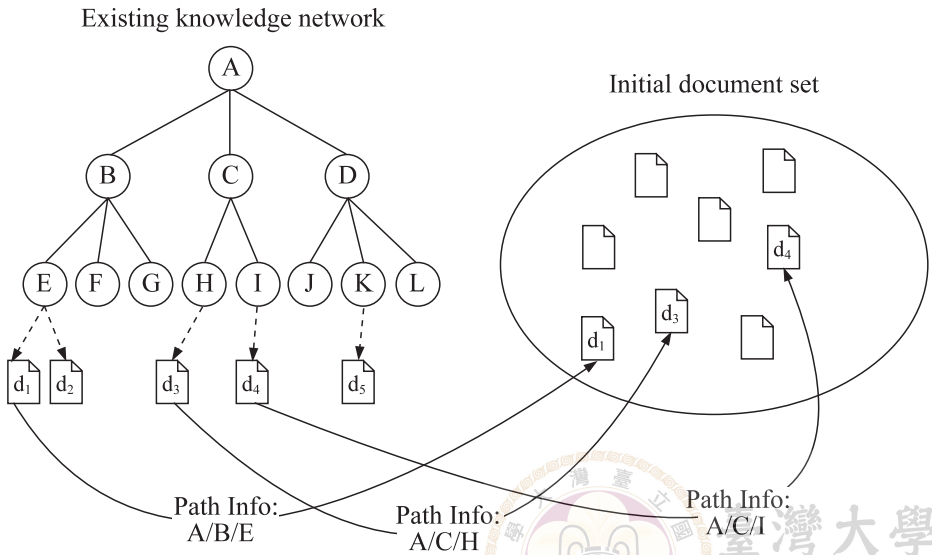


圖 2　標籤資訊擷取程序

臺灣大學學術期刊資料庫

著手進行文件的比對工作。我們將欲分類的文件資料與既有的文件資料集進行比對，如果比對成功，則我們便會把既有的標籤資訊轉嫁在我們欲分類的文件上，這種「貼標籤」的動作提供了我們分類架構建構過程中重要的資訊，而這些標籤我們稱之為外來文件的投票（Document Foreign Vote, DFV）。

這些資訊將用特定的比例與文件的內容一起進行階層式聚合演算法計算（請見以下演算法決策函數）。但是不同於過去的階層式聚合演算過程，我們的目標並不是要自動產生一個分類階層，而是當本體認知架構的設計者，已經有初步的設計構想時，將網路上的資訊進行合適的分群，並且能夠將分群的結果歸納至原先的初步架構之下，甚至能夠對原有的架構提出調整的建議。我們將傳統的階層式聚合（Hierarchical Agglomerative Clustering，HAC）演算法進行改良，將我們所提出的外來文件投票（DFV）資訊引進HAC演算法中。因此我們提出將HAC原有的聚合決策函數修改如下：

	Let $d_u = \{w_{u,1}, \ldots, w_{u,n}\}$ and $d_v = \{w_{v,1}, \ldots, w_{v,n}\}$ be two vectores of correlation values for the document d_u and d_v, $w_{u,i} = tf_{u,i} \times idf_i = \frac{freq_{u,i}}{\max freq_{u,i}} \times \log \frac{N}{n_i}$
EQ1	the document similarity estimation function is $sim_1 = sim(u,v) = \frac{\vec{d_u} \bullet \vec{d_v}}{\lvert\vec{d_u}\rvert \times \lvert\vec{d_v}\rvert} = \frac{\sum_{i=1}^{n} w_{u,i} \times w_{v,i}}{\sqrt{\sum_{i=1}^{n} w_{u,i}^2} \times \sqrt{\sum_{i=1}^{n} w_{v,i}^2}}$
	Let $p_u = \{l_{u,1}, \ldots, l_{u,n}\}$ and $p_v = \{l_{v,1}, \ldots, l_{v,n}\}$ be two path information for the document d_u and d_v, the path similarity estimation function is
EQ2	$sim_2 = DFV_{u,i} = \sum_{i=1}^{n} h(l_{u,i}, l_{u,i})$ where $h(l_{u,i}, l_{u,i}) = \frac{1}{n-i}$ *if* $l_{u,i} = l_{u,i}$, 0 otherwise.
EQ3	The grand similarity estimation function is shown below : $sim(u,v) = w_1 sim_1 + w_2 sim_2, 0 \le w_1 < 1, 0 \le w_2 < 1, w_1 + w_2 = 1$

本研究中階層式聚合演算法決策函數（similarity estimation function）

臺灣大學學術期刊資料庫

本研究提出的分群演算法如下：

$$
\begin{aligned}
&C_1 = \{d_i\}, \forall i \\
&M_1 = sim(C_1, \phi) \\
&k = 1 \\
&\text{while} |C_k| > cluster_threshold \text{ do} \\
&\quad C_{k+1} = cluster(C_k, M_k); \\
&\quad M_{k+1} = sim(C_{k+1}, M_k); \\
&\quad k = k + 1; \\
&\text{end;} \\
&answer = C_k;
\end{aligned}
$$

分群演算法虛擬流程

上述流程中，M_i 表示由各分群之間的相似程度所構成的相似度矩陣，ϕ 表示空集合，*cluster* 函數則為根據上述決策函數進行分群合併動作，依照群聚距離進行分群合併，遞迴反覆至分群門檻條件滿足為止。由於HAC演算法所產生的架構為二元樹（binary tree）結構，因此後續必須再使用群聚分割演算法（Chuang & Chien, 2003; Mirkin, 1996）來產生最終的知識分類架構。我們所設計的導引式分群的運作方式，並不是以合併到剩下一個群組為目標，而是希望藉由分類標記資訊的輔助，將傳統的分群作一個導引式的修正，進而達到融合既有分類資訊到預設的少量架構資訊的目的。

整個方法的流程如圖3所示：

肆、實驗結果

以下我們使用一個網際網路上的真實例子做為 GCC 演算法的實驗對象。我們選取 LookSmart 網站的網頁目錄（looksmart.com）做為我們的實驗資料，並假設我們要在這樣的資料集上產生太空類（Astronomy）的知識分類架構。這個太空類別同時也出現在 Yahoo（yahoo.com）與 ODP（dmoz.org）的網頁目錄中，表 1 說

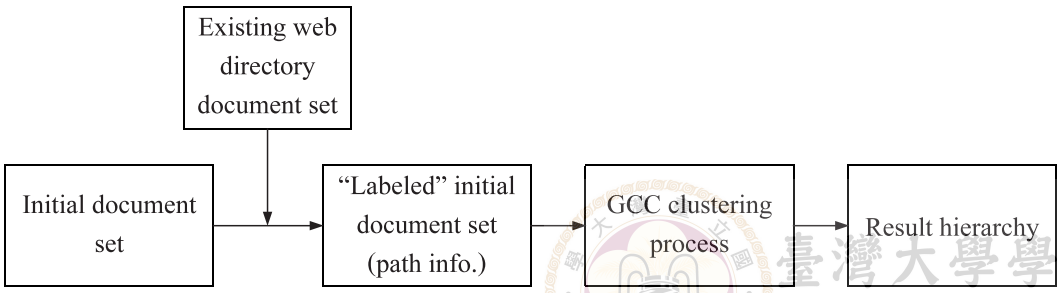


圖 3　導引式分群方法處理流程

臺灣大學學術期刊資料庫

表 1　網頁目錄中的分類資訊（a）ODP 與（b）Yahoo

（a）ODP 中太空類部分資料（標示底線的類別為實驗結果所出現的標籤）

Level1 （Astronomy）	Level2 （Astronomy>Astrophysics, for example）
… Archaeoastronomy@ （95） **Astrophysics**@ （63） Calendars and Timekeeping （47） **Cosmology** （42） **Dark Sky**@ （144） Eclipses，Occultations and Transits （84） Extrasolar Planets （43） Extraterrestrial Life （147） **Galaxies** （54） Interstellar Medium （16） **Personal Pages** （145） Solar System （716） **Star Clusters** （17） Stars （127） …	Computational Software@ （7） History@ （184） Institutions **Black Holes**@ （63） **Cosmology**@ （42） Neutron Stars@ （24） Stellar Evolution@ （12） …

明了這兩個網頁目錄中的太空類資料部分架構。由於我們是使用 LookSmart 網頁目錄下的太空類文件做為基本資料集，因此文件在 LookSmart 中的分類資訊將被捨棄不用。整個實驗過程我們用以下的步驟進行說明：首先，每個實驗用的文件都會使用我們所提出的標籤資訊擷取程序進行處理，也就是說，如果該文件正好出現在Yahoo或是ODP分類中的話，其Yahoo/ODP 分類資訊（也就是分類路徑）將會以標籤的形式附加在該文件上。接著，我們便使用本研究所提出的聚合演算法決策函數進行文件分群。當分群動作結束後，我們便可以得到如表 2 所示的最終分類架構，同時相關的數據呈現在表3中。觀察表2我們可以得到最終的架構有點類似 Yahoo 與 ODP 分類架構的綜合體，同時在表 1（a）與（b）中以底線標示的類別都出現在表 2 的架構之中。由此可以知道，這樣的程序不但可以將 Yahoo 與 ODP 的分類概念以新的方式結合後帶入新的分類架構中，同時也達到了藉由這兩種分類架構中所擷取資訊的建議，適當地融入成為一個新的分類架構，達到了藉由既存架構導引建構出新的分類架構的目的。

國立臺灣大學圖書館 期刊資料庫

表 1　網頁目錄中的分類資訊（a）ODP 與（b）Yahoo（續）

（b）Yahoo 中太空類部分資料（標示底線的類別為實驗結果所出現的標籤）

Level1 **（Astronomy）**	**Level2** **（Astronomy>Astrophysics, for example）**
… Astronomers （102） Astronomical Calendars （20） **Astrophysics** （127） Auroras@ Comets，Meteors，and Asteroids@ Companies@ Conferences and Events@ Cosmology （113） **Deep Sky** （13） Eclipses@ **Education** （174） Extrasolar Planets （31） FITS （Flexible Image Transport System） （1） **Galaxies** （49） History （16） Humor@ Infrared （11） Interstellar Medium （72） Light Pollution@ News and Media （29） Organizations （23） Pictures （43） Planetaria （62） **Radio Astronomy** （78） Research （248） Scientific Notation@ **SETI （Search for Extraterrestrial Intelligence）** @ Software （14） **Solar System** （987） **Space**@ Stars （222） Telescopes （26） Web Directories （12） **X-Ray** （20） …	Ask an Expert （3） Conferences （1） **Cosmology**@ **Galaxies** （7） Gamma-Ray Observation （4） Institutes （48） Neutrino Detection@ Observatories@ Relativity@ **Research** （22） **Space Physics**@ Stellar Phenomena@ Surveys （6） Universal Origins@ …

臺灣大學學術期刊資料庫

表 2　GCC 分群建構結果

Level1 （Astronomy）	Level2 （Astronomy>Astrophysics）
Astrophysics（ODP/Yahoo）	**Black Holes**（ODP）
Cosmology（ODP）	**Cosmology**（ODP/Yahoo）
Dark Sky（ODP）	**Galaxies**（Yahoo）
Deep Sky（Yahoo）	**Research**（Yahoo）
Education（Yahoo）	Reference
Galaxies（ODP/Yahoo）	**Space physics**（Yahoo）
Personal Pages（ODP）	…
Reference	
Solar System（Yahoo）	
Space（Yahoo）	
Star Clusters（ODP）	
X-Ray（Yahoo）	

表 3　GCC 分群實驗相關數據

Hierarchy processed	ODP: Astronomy	Yahoo: Astronomy	Our result: Astronomy
Total node count	3,901	3,016	1,347

接著為了確認由GCC演算法所產生知識分類架構的品質，我們參考了幾種樹狀結構的相似度計算方式（Bhavsar, Boley, & Yang, 2004; Morrison, 2003），最終的相似度結果呈現於表4中。這項結果顯示了 GCC 的實驗結果與參考的幾個分類架構綜合起來擁有超過 60%的結構相似性。也就是說，GCC結果所產生的結構擁有超過原參考結構中分類項目 60%的內容，亦即藉由這兩種分類架構中所擷取資訊的建議，適當地融入成為一個新的分類架構，表現出了參考架構融合的效果。在表 4 中，所謂的結構相似性，是指最終架構中節點與 Yahoo/ODP中重複的比例。

伍、結論

在這個研究中，我們將傳統的資料分群演算法，修改成為一種導引式的概念分群演算法，並從初步的實驗中得到了有趣的結果。這樣的演算方式目前仍有幾個值得後續討論與改善的議題。首先，由於GCC的決策函數目前是取決於傳統的文件相似度與分

表 4　GCC 所產生之知識分類結構的相似性度量

Experiment: $w_1 = 0.6$，$w_2 = 0.4$ Average-link HAC	Structural similarity with union of ODP & Yahoo	Structural difference from union of ODP & Yahoo
Level1	77.27%	22.73%
Level1&Level2	61.66%	38.34%

類標籤的相似度，是否有其他的資訊可以提供GCC進行更具語意的聚合分群，仍有待討論。其次，決策函數中的比重分配，尚待發展出一個客觀且自動的程序予以設定，這需要更多的機器學習理論進行探討。最後，標籤擷取程序是否可以使用模糊比對的方式進行擷取，這將會是個值得討論的部分，這是因為模糊比對將可能會使標籤資訊增加，但同時也有可能降低聚合分群結果的正確性。

本文所探討的GCC演算法，其應用範圍不只可以用在網頁的分類上，更可以推廣到其他的專門領域上，包括生物、能源、甚或是政府施政分類上。（收稿日期：2008/3/3；接受日期：2008/5/13；修訂日期：2008/6/9）

參考文獻

Bhavsar, B., Boley, H., & Yang, L. (2004). A Weighted-Tree similarity algorithm for Multi-Agent systems in E-Business environments. Computational Intelligence Journal, 20 (4), 584-602.

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In Proceedings, ACM SIGMOD International Conference on Management of Data : SIGMOD '98 : June 1-4, 1998, Seattle, Washington, USA (pp. 307-318). New York: ACM.

Chen, C. C., Yeh, J. H., & Sie, S. H. (2005). Government ontology and thesaurus construction: A Taiwanese experience. In Digital libraries : implementing strategies and sharing experiences : 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, December 12-15, 2005: proceedings. Berlin: Springer.

Chuang, S. L., & Chien, L. F. (2002). Towards automatic generation of query taxonomy: A hierarchical query clustering approach. In 2002 IEEE International Conference on Data Mining : proceedings : 9-12 December, 2002, Maebashi City, Japan (pp. 75-82).

臺灣大學學術期刊資料庫

Los Alamitos, Calif.: IEEE Computer Society.

Chuang, S. L., & Chien, L. F. (2003a). Automatic query taxonomy generation for information retrieval applications. Online Information Review, 27(4), 243-255.

Chuang, S. L., & Chien, L. F. (2003b). Enriching web taxonomies through subject categorization of query terms from search engine logs. Decision Support Systems, 35(1), 113-127.

Ghani, R., Slattery, S., & Yang, Y. (2001). Hypertext categorization using hyperlink patterns and meta data. In Machine learning : proceedings of the eighteenth International Conference (ICML 2001), Williams College, June 28-July 1, 2001 (pp. 178-185). San Francisco, Calif.: Morgan Kaufmann Publishers.

Hwang, C. H. (1999). Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), Linkoping, Sweden, July 29-30, 1999.

Jannink, J. (1999). Thesaurus entry extraction from an On-line dictionary. In FUSION '99 : proceedings of the Second International Conference on Information Fusion, July 6-8, 1999. Sunnyvale, California, USA. Mountain View, CA: International Society of Information Fusion.

Jannink, J., & Wiederhold, G.. (1999). Ontology maintenance with an algebraic methodology: A case study. In Ontology management : papers from the AAAI Workshop. Menlo Park, Calif.: AAAI Press.

Kashyap, V., Ramakrishnan, C., & Rindflesch, T. C. (2003). Toward (semi-)automatic generation of bio-medical ontologies. In Biomedical and health informatics from foundations to applications: November 8-12, 2003, Washington, DC [proceedings]. [Madison, WI]: Omnipress.

Li, F., & Yang, Y. (2003). A loss function analysis for classification methods in text categorization. In Proceedings, Twentieth International Conference on Machine Learning : August 21-24, 2003, Washington, DC, USA (pp. 472-479). Menlo Park, Calif: AAAI Press.

Maedche, A., & Staab, S. (2000). Discovering conceptual relations from text. In ECAI 2000 : 14th European Conference on Artificial Intelligence, August 20-25, 2000, Berlin, Germany: proceedings. Amsterdam; Washington, DC: IOS Press.

Maedche, A., & Staab, S. (2002). Mining ontologies from text. In Knowledge engineering and knowledge management : methods, models, and tools : 12th international confer-

臺灣大學學術期刊資料庫

ence, EKAW 2000, Juan-les-Pins, France, October 2-6, 2000 : proceedings. New York: Springer.

Mirkin, B. G. (1996). Mathematical classification and clustering. Dordrecht; Boston: Kluwer Academic Publishers.

Morrison, T. (2003). Similarity Measure Building for Website Recommendation within an Artificial Immune System. Unpublished doctoral dissertation, University of Nottingham, England.

Palmer, C. R., Pesenti, J., Valdes-Perez, R. E., Christel, M. G., Hauptmann, A. G., Ng, D., & et al. (2001). Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results. In Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries : June 24-28, 2001 : Roanoke, Virginia, USA. New York, N.Y: ACM Press.

Wagner, A. (2000). Enriching a lexical semantic net with selectional perferences by means of statistical corpus analysis. In Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000), Berlin, Germany, August 2000.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval, 1(1/2), 67-88.

Yang, Y. (2001). A study on thresholding strategies for text categorization. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval : New Orleans, Louisiana, USA, September 9-13, 2001: SIGIR 2001 (pp. 137-145). New York, N.Y.: ACM.

Yang, Y., & Liu, Xin. (1999). A re-examination of text categorization methods. In Proceedings of SIGIR '99 : 22nd International Conference on Research and Development in Information Retrieval : University of California, Berkeley, August 1999 (pp. 42-49). New York, N.Y.: ACM.

Yang ,Y., & Pedersen, J. P. (1997). A comparative study on feature selection in text categorization. In Machine learning: proceedings of the fourteenth International Conference (ICML '97), Nashville, Tennessee, July 8-12, 1997 (pp. 412-420). San Francisco, CA: Morgan Kaufmann Publishers.

Yang, Y., Slattery, S., & Ghani, R. (2002). A study of approaches to hypertext categorization. Journal of Intelligent Information Systems, 18, 219-241.

Yang, Y., Zhang, J. & Kisiel, B. (2003). A scalability analysis of classifiers in text categorization. In SIGIR 2003 : proceedings of the Twenty-Sixth Annual International ACM

臺灣大學學術期刊資料庫

SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28 to August 1, 2003 (pp. 96-103). New York, N.Y.: ACM.

Zhang, J., & Yang, Y. (2003). Robustness of regularized linear classification methods in text categorization. In SIGIR 2003 : proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28 to August 1, 2003 (pp. 190-197). New York, N.Y: ACM.

臺灣大學學術期刊資料庫