網頁資訊檢索與分群技術於數位典藏知識庫建構之應用分析

林致祿, 夏敏翔, 王怡聖, 林高弘, 高宏宇, 盧文祥

Dept. of Computer Science and Information Engineering

National Cheng Kung University

摘要

近年來各典藏機構已成功地扮演數位知識內容的提供者之重要角色，然而一般民眾廣泛的知識需求仍然快速地增加；同時專業數位知識內容建構仍然需要內容專家協助。整體而言，各典藏機構在數位知識內容的增補與更新仍相當緩慢，對於使用者的需求仍不甚明瞭或無法迅速提供適合的知識服務。本論文將分析如何整合一些有效的網路探勘技術和網路搜尋技術，嘗試利用網路蘊藏豐富的多語言資訊來協助內容專家擴充知識庫的建構，以發展一個可行的半自動數位知識內容建構技術。不同於傳統以人工為主的知識內容建構技術，我們將分析如何透過網路資源的萃取、過濾以及分群技術，產生一個可即時擴充的實用知識本體以提供使用者最彈性的跨領域的知識服務或數位學習。

1. 簡介

「數位典藏國家型科技計畫」自民國 91 年 1 月正式展開，目的在於妥善保存國家的文化資產，推廣精緻藝術的流傳與品賞，強健文化的傳承與發展，並鼓勵資訊與知識的分享。近年來國內許多學術研究機構與團體，已協力進行相關技術的研發，並共同配合支援各典藏機構的數位典藏工作，例如多媒體數位典藏資料庫管理、建置、搜尋技術等。各典藏機構現階段已成功地扮演數位知識內容的提供者及生產者之重要角色，然而符合一般民眾更廣泛的知識需求仍然等待快速地增加；更深入的專業數位知識內容建構仍然需要內容專家協助。整體而言，各典藏機構在數位知識內容的增補與更新仍相當緩慢，對於使用者的需求仍不甚明瞭或無法迅速提供適合的知識服務，因此各典藏機構仍期待學術研究機構提供前瞻性之創新技術，朝向實際可行的半自動數位知識內容建構。本論文將以如何將 Web 上結構化及非結構化資料匯整於典藏知識庫為主要探討問題，進行 Web 資源與數位典藏關連的研究。

本文朝著這個方向提出一個利用網路探勘技術來輔助數位典藏單位進行相關的知識庫（Knowledge Base）的建構和擴充的一個基本雛形架構。不同於以往典藏單位利用不易獲得的專家人力來建構相關的知識庫和知識本體，本論文將利用全球資訊網(WWW)作為我們的知識寶庫，發展出一套半自動化甚至全自動化的知識建構系統，以改進以人工方式在知識建構的效率上與知識建構的完整性上之不足。下面我們將就本論文核心問題網路內容探勘加以探討。

隨著網際網路的發展和網頁動態技術的成熟，全球資訊網網頁的數量呈現爆炸性的成長，其內涵的資訊可說是相當廣泛且多樣化，在本計畫中我們利用搜尋引擎，例如：Google, Yahoo 等來當作是我們網路資訊搜尋的入口。對於某特定搜尋字詞，我們都可以得到許多與它相關的網頁，不過有些網頁中雖然都跟我

們的搜尋字詞有關，但是可能在語意上卻不見得是我們要的。例如：如果我們要找尋關於 Rembrandt(一位荷蘭畫家)的相關資料，我們可以在搜尋引擎中的確可以找到許多描述這位畫家的資料，不過我們也會找到很多不相關但卻含有這個特定字詞的網頁，包括用這個名字來命名的旅館，餐廳網頁。所以，如何在這相關的網頁中找到與我們需求的知識相關的網頁是我們想要解決的課題之一。

而這個問題一般可以使用範例搜尋(Searching by example)來解決[Alani 2003]，這個方法中必須有人工介入來選擇一些合適的網頁，透過這些網頁內容的比對，再去找到相關的網頁。雖然這個方法對於結果有一定程度的改善，不過對於一個自動化知識擷取系統而言，這樣的人工需求並不實用。在這裡我們將改良一種自動化群聚搜尋結果的方法(Clustering on Search Results)先將搜尋結果進行分群動作，再依群組重要度取出我們需要的網頁。如Figure 1 [Zeng 2004] 所示，當我們想要搜尋 Jaguar 這個關鍵字的時候，我們得到的結果可分群為 Jaguar 汽車、大貓(動物)、Mac 作業系統和俱樂部名稱等，內容雖然都與 Jaguar 有關，但是描述的東西卻是南轅北轍。透過分群的分析，我們可以更清楚瞭解目前我們搜尋到的資料中知識的種類和差異性。這也對於我們之後的語意分析有所幫助。

Figure 1: 搜尋結果群組範例

在本論文中我們將以一些初步的實驗分析來探討如何整合一些有效的網路探勘技術和網路搜尋技術，以發展一個可行的數位知識內容建構技術。不同於傳統以人工為主的知識內容建構技術，本論文將特別專注在利用豐富的網路資源，整合一些有效的知識探勘技術。

2. 相關研究

目前英國有一個六年八百萬英鎊的跨校合作計畫 Advanced Knowledge Techonology (AKT, http://www.aktors.org/akt/) 正在執行。該計畫主要著重在下列六個議題，包括 Knowledge Acquisition, Knowledge Modeling, Knowledge Reuse, Knowledge Retrieval and Extraction, Knowledge Publishing, Knowledge Maintaince，目前已有不少研究論文出版，另外實際的系統和相關工具也陸續開發。根據我們仔細的探究，和本計畫最相關的部分是他們也積極地朝向Web-based的知識庫及知識本體建構研究 [Alani et al. 2003]，例如Figure 2展示Artequakt系統架構，該系統可以完全自動地從Web收集某些藝術家的資訊，利用這些相關訊息自動建立知識庫及知識本體，然後根據知識庫及知識本體配合使用者不同的喜好，呈現不同格式的藝術家簡歷。該系統對於知識擷取處理充分整合一些語法和語意分析工具，主要包括Syntax Parser, Name Entity Recognition和WordNet等，因此可以有效地分析人事時地物的關聯或階層關係，然後自動地建構或不斷地增補到知識庫及知識本體。

網路探勘和文件探勘是資料探勘（Data Mining）領域新的研究議題[Cooley et al. 1997; Hearst 1999]，主要是利用半結構化的網頁內容和非結構化的文件內容來擷取有用的資訊或知識，最近有許多研究提出各種方法從大量的文件語料庫擷取關鍵詞或 template，有些

則開始利用網路自動建構知識本體[Ahonen 1999; Bar-Yossef and Rajagopalan 2002; Feldman 1995, 1997; Soderland 1997]。另外有許多研究利用網路鏈結結構和鏈結文字（Anchor Text）來擷取相關的或 authoritative 網頁[Brin and Page 1998; Chakrabarti et al. 1998; Dean and Henzinger 1999; Kleinberg 1998]。過去幾年我們著重在開發蘊藏於網路內的豐富的多語資訊作為術語翻譯分析的動態語料，主要是想解決網路未知多語術語自動翻譯問題 [Lu et al. 2002, 2003, 2004; Wang et al. 2004; Cheng et al. 2004]。

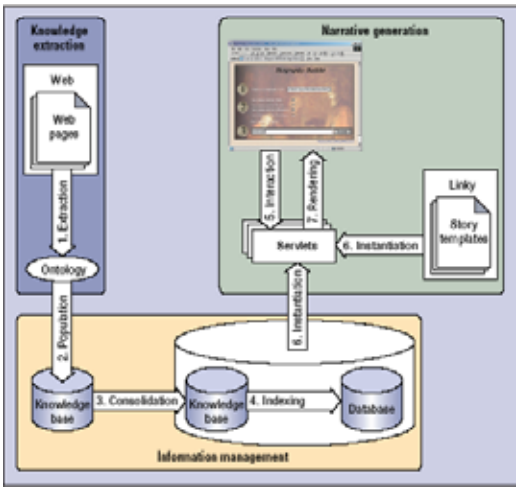

Figure 2 : Artequart 系統架構

一般在處理文件內容擷取的研究中，多半是採用 Wrapper 的作法[Cohen 1999; Kushmerick 1997; Lin 2000]。Wrapper是透過一些訓練的機制，來得到一些演繹式的語言敘述表達通式，透過這些通式使用者可以用來擷取一些相同結構的文件資料。而有另一部份的研究[Adelberg 1998; Hsu 1998; Laender 2002]則是提供一個半自動化介面讓使用者可以透過輸入一些學習資料讓系統可以學習到一些擷取規則，而透過這些規則，系統便可以去擷取固定格式的資料。不過，如果當我們考量的環境是整個全球資訊網的時候，由於文件的多樣化和複雜化，這些需要透過人工設定學習資料的作法，需要花費相當大的人力，從實用面與效率等層面來看，都並不適宜。

在先前的研究中[Kao 2004][Kao 2002]，我們利用網頁的連結和資訊熵(entropy)的分析我們可以擷取出一個網頁具有資訊的部分，稱為資訊結構。這部分的研究把資訊理論與文件架構的分析相結合，利用一個由上而下的搜尋方式，找到資訊結構的骨幹架構，後再利用一些提出的群聚演算法來找到最後的資訊結構。

全球資訊網蘊藏豐富的多語言資源，而且不斷地快速成長，另外這些資源非常容易由網路取得，可以實際有效地而且大規模地被開發來建構或擴充數位典藏知識庫。因此網路探勘和文件探勘等新議題和相關技術之研究是值得注意。

3. 網頁檢索與分群技術之應用

包含世界各國各種語言資訊的全球資訊網不斷地快速成長，因此當系統從典藏知識探勘或使用者需求探勘結果發現典藏知識仍有不足而且急需增補，我們可以利用網路知識擷取技術透過搜尋引擎找尋全球相關資訊（包含豐富的文字和圖片等），然後利用資訊萃取技術來分析資訊蘊藏的人事時地物等語意關係，並建構這些語意訊息彼此的關聯結構。

在本論文中，我們利用網路蘊藏豐富的網路資源，用以開發一個網路資訊擷取系統來分析文件物件模型的探勘系統以建立每個網頁的樹狀物件模型，以進行網頁重要區塊與資訊的擷取。從網路的資料搜尋資訊，我們利用知識萃取過濾與結果分群分析的技術，讓網路資源可以更有效地被利用。如Figure 3所示，兩個知識擷取流程能夠將網路資訊有效率地萃取與過濾，並能透過分群的呈現使得不同知識內涵的呈

現更能區分。以下我們將分別針對此兩個知識擷取流程做更進一步描述。

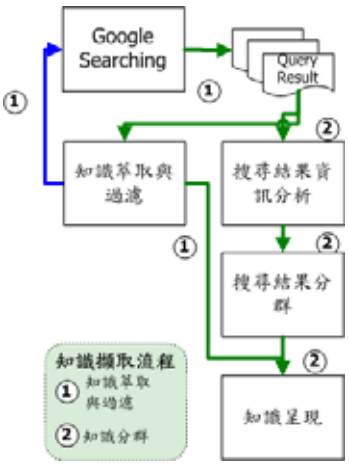

Figure 3: 系統流程圖

3.1 搜尋結果的知識萃取過濾

由於網路的快速成長，各種領域的資源非常的豐富，目前可以從網路上取得大量的資訊相對較為容易，因此在網路的知識擷取上，現階段我們主要利用搜尋引擎，根據搜尋結果來進行重要知識區塊的擷取。這部分主要利用資訊萃取技術來分析資訊蘊藏的知識概念關係，並建構這些知識概念彼此的關聯結構，我們將整合資訊檢索和自然語言處理已開發成熟的一些技術，例如中央研究院詞庫小組開發的具備未知詞偵測與句法詞類預測能力的中文分詞技術 [Ma & Chen, 2003]，嘗試開發一個實用的知識概念擷取工具，本節將介紹兩個主要模組：資訊收集及知識萃取過濾。

3.1.1 資訊收集

我們嘗試使用搜尋引擎來收集知識萃取時所需要的大量相關資訊，例如 Google 目前已收集超過 80 億的網頁，可以方便我們即時取得充足的資訊。對於使用者想要了解或檢索的知識概念，我們提出的想法就是根據這些知識概念和它的細部特徵(Refined Feature)來擷取更詳細的知識概念，譬如使用者輸入醫學疾病概念"心臟病"，我們就將它的細部特徵，如症狀、原因、治療等分別和"心臟病"，產生新的查詢字串 Refined Query = (Query/Concept，Refinement/Feature)，例如 Refined Query = (心臟病，症狀)，然後利用搜尋引擎來收集更詳細的知識，知識概念詞(Query/Concept)用來決定知識概念的相關資訊，細部特徵詞(Refinement/Feature)更進一步的取得更精確的資訊。對於不同領域的知識概念擷取，我們可以採用一致的方式來處理，譬如使用者輸入動物概念"台灣黑熊"，我們就將它的細部特徵，如分布、特徵、習性等分別和"台灣黑熊"，產生新的查詢字串 Refined Query = (台灣黑熊，分布)等來搜尋更詳細的知識。我們將新的查詢字串輸入搜尋引擎，取回 100 筆的搜尋結果(Search Result)，而每一筆都會有從相關網頁擷取的片段文字來描述相關資訊，此片段描述文字(Snippet)就是我們想要進步擷取的詳細知識概念來源。初期我們使用簡單的啟發式方法(Heuristics)，依據片段描述文字中的特別符號"。"和"…"將這 100 筆片段分成許多段落，而這些段落就是我們要進一步萃取的知識片段。

3.1.2 知識萃取過濾

利用上面的資訊收集方法，一般而言對於常見的知識概念都可以收集到充足的知識片段，接下來我們要決定那些段落含有正確的知識，目前仍然是很大的挑戰，初步的想法是參考資訊檢索的關鍵詞擷取技術，我們假設句子裡最重要的就是名詞跟動詞，這部份採用中央研究院詞庫小組開發的中文斷詞技術來找出段落中的名詞跟動詞，然後計算他們的出現機率，將機率前n名的名詞和動詞作為關鍵詞(目前設n = 10)，而機率值就當成它們的重要性。根據知識概念詞、細

部特徵詞、和關鍵詞，利用下面的公式來計算各知識片段S_i的重要性：

$$Score(S_i) = \alpha \times W_c + \beta \times W_f + (1 - \alpha - \beta) \times \sum_{k \in K} W_k$$

W_c是知識概念詞的匹配值，W_f是細部特徵詞的匹配值，假如該片段有出現這些詞，目前匹配值皆設為1，若未出現則設為 0。W_k是關鍵詞的匹配值，假如該片段有出現這些詞，目前匹配值設為關鍵詞的機率值，若未出現則設為 0。另外α和β是權重值，我們使用α = 0.5，β = 0.3。經過計算每一知識片段都有一個重要性分數，經排序取前m名為相關性最高的知識片段(目前設m = 5)。

3.1.3 知識概念擷取工具

Figure 4是我們開發的知識概念擷取工具的介面，我們將使用這套工具來協助數位典藏單位可以較快速來增補某些常見知識概念內容的不足。Query 輸入使用者要查詢的知識概念，像是動植物名稱，例如台灣獼猴、台灣黑熊、烏蕨、...等等，refinement 輸入使用者想了解知識概念的一些細部特徵，例如動物的分布狀況、生活習性等等，舉例來說，我們想知道台灣彌猴有哪一些特徵、生活在什麼地方，我們就可以輸入Refined Query = (Query，refinement) = (台灣彌猴，特徵、分布)來搜尋知識概念較精確的知識片段。下面將舉兩個例子來說明，如何使用這些知識片段來增補國立自然科學博物館自然與人文數位博物館(http://digimuse.nmns.edu.tw/index.jsp) 某些數位知識內容的不足：

(1) Figure 5是自然與人文數位博物館對於台灣彌猴的簡介，使用者若是覺得簡介稍嫌不足，可以利用我們知識概念擷取的工具來增補內容。Figure 4是我們使用(Query，refinement) = (台灣彌猴，簡介，特徵型態、分布、習性) 得到的結果，可以發現Figure 4簡介欄知識萃取的結果並不理想，為了快速擷取知識，以減少分析內容的運算時間，現階段我們只採用 Search Result 的結果來萃取知識，並沒有更深一層的去搜尋網頁裡的知識，但是一般對於動物的簡介都用特徵、分布、習性來介紹，因此我們使用這三項來替代我們簡介欄裡的知識。在Figure 4分布欄、習性欄(第三、四欄)正方形方框裡的資料是數位博物館簡介內容所沒有的，而這些增補知識可以讓使用者更了解台灣彌猴的一些習性，因此我們就可以使用這樣的方法來增補數位博物館裡對於動植物知識介紹的不足。

Figure 4 : 知識概念擷取工具

Figure 5: 自然與人文數位博物館對於台灣彌猴的簡介

Figure 6: 知識概念擷取工具對於雲豹的相關知識擷取

(2) 在自然與人文數位博物館中，目前仍然有很多本土的動植物並未提供民眾檢索，例如台灣黑熊、雲豹、帝雉...等等。對於數位博物館裡沒有的動植物知識內容，以雲豹為例，數位博物館中沒有雲豹的資料，下圖是我們的工具所搜尋出有關雲豹的知識，我們可以發現裡面有很多的資訊都可以用來介紹雲豹，而且還可以經由我們提供的連結來獲取更多的知識，所以藉由此知識概念擷取工具可以快速且有效的幫助數位博物館來收集所缺少的動植物知識內容。

3.2 搜尋結果的分群

隨著搜尋引擎的發展至今，越來越多網站的網頁已經被搜尋引擎 crawl 以及 index，目前最大的搜尋引擎 Google 已經索引了超過 80 億的網頁，這也意味著，有極大量的查詢結果能夠提供給使用者了，這直接的造成了一個問題，就是使用者只能用一個接著一個查詢結果的方式，去對搜尋引擎回傳的結果做肉眼的確認，但如果回傳結果有數百萬項，這勢必會對使用者造成一種困擾及負擔。近來有越來越多的人開始對提高搜尋引擎的服務品質產生研究興趣了，本文的研究即是對其中的一種方法做相關的研究，而這種方法即是對網頁做分群（Clustering），所謂分群即是利用一種不受監督（unsupervised）的方法，去把搜尋結果分到不同的目錄，加速及便利使用者在搜尋引擎的使用，而這裡所謂的不受監督即是指不事先定義目錄名稱，而是隨著搜尋結果之間的相關興趣自動產生目錄名稱。簡單的來說，分群的好處就像去超級市場，如果沒有吊牌顯示某區主要是賣什麼商品的話，那在最差的情況下，可能要找完所有區域才會找到相要買的東西，而分群時自動產生的目錄名稱就像吊牌一樣，可以加速讓你找到想要的網頁。這部分的研究技術核心與 Searching by example 相似，同樣是去分析網頁間關鍵字的相似度來統計網頁與網頁間的群組距離，只是這個分析的動作本來是由人工完成，現在我們透過分析可以自動地達到分群功能。這部分我們首先利用向量模型(Vector Space Model)來描述搜尋結果中的網頁，並以向量間的 cosine 數值來表示網頁間的距離，透過不同的群組演算法，我們將評估其群組效果，並且會根據資料分佈來決定此搜尋結果需不需要分群。另外群組內網頁的相關程度也是一個衡量這個結果群組所表達的知識是否集中，太分散的群組所擷取出來的知識相對的也會較分散，這將影響我們之後的知識擷取。我們可以根據某個依實驗得到的評估標準將內涵資訊太雜的群組刪除，剩下來群組內的網頁便可以供區塊與資訊擷取所用。

在資料分群前，我們先對搜尋結果分析其資料的集中度和特定度，用來決定是否需要更進一步的分群。通常在一些比較通用的查詢字時，分群的效果是很顯著的，但是如果查詢字串是很特定的專有名詞時，如 Macaca cyclopis，那搜尋的結果其實會相當集中，此時便不需要額外的去分群。

4. 實驗資料蒐集與分析

我們對資料分群分析進行了以下實驗。首先先挑選 10 個查詢，再去 Google 送出查詢，每一個查詢取回 100 則結果，這份結果使用在人工分群及分析 DF（Document Frequency）這兩部份。在人工分群部分，主要是建構一個標準答案，使用人工去對每一個查詢做分群的工作，記錄下目錄名稱及相對應的目錄下包含了幾個搜尋結果，而這裡的目錄名稱是由評估者所認定最符合的目錄名稱。在 DF 分析方面，我們採用 n-gram 的方式先擷取出 term（1 ≦ n ≦ 3），並濾掉一些標準的 stop-word。接著我們統計每個 n-gram term 總共在幾則結果中出現的數目 X（1 ≦ x ≦ 100），我們將 DF 的分佈情況與先前的人工分群結果進行比較，可發現此兩者有著一定程度的關係。分群越多的查詢，其 DF 的分佈越平均，代表字詞的分佈越有群聚化現象。

我們實驗的查詢包括(1)一般性的：keyboard、star、paper、apple 和 Jodan，以及(2)專業性的動植物專業名詞：Macaca cyclopis(台灣獼猴)、bird、mouse、Blechnum orientale L(烏毛蕨)以及 PTERIDOPHYTA。各個查詢結果的人工分群和 DF 分佈如Figure 7所示，左邊為人工分群的群組與數量分佈，右邊為 DF 的統計圖。我們可發現圖形可以大略分成兩類，一類是有關動植物專業名詞這種查詢，而另一類就是比較一般性的查詢，和一般性的查詢比較起來可以看出有關專有名詞的 DF 比較有起伏且比較不會偏重於 DF=1 的字詞。另外，某些字詞因為具備了某些特殊的意義，因此常有完全不同的內容，卻使用擁有相同 query 結果的 pages。例如：許多新聞網站都會有 star 的字詞出現，而這些網站所介紹的內容未必雷同，有時反倒是大相逕庭的。此外，我們觀察到雖然 keyboard 和 mouse 都是一種電腦的硬體設備，由於有許多有關老鼠的生物

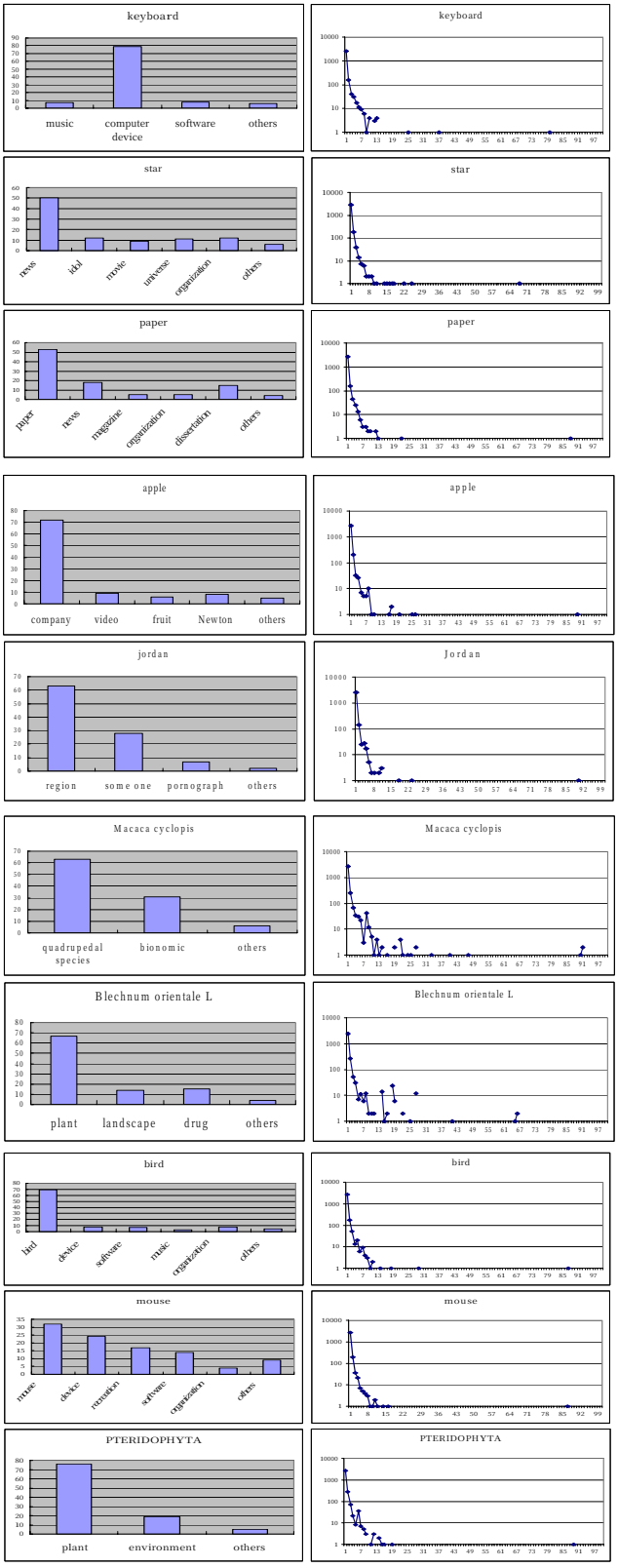

Figure 7: 人工分群與 DF 分佈比較圖

研究、卡通…等等，以至於這二者的 query 結果大不相同；前者的結果完全偏向設備，而後者卻是較為平均地分散在各個 query 中。而就 apple 的結果來看，發現有七成以上的分群結果都是 company，這是因為我們用人工的方式，只從 Google 中選取前 100 筆的 query 來看，由於資料量過少，可能前 100 筆的資料都會偏向某一方面，而不夠廣泛與全面化。另一個原因可能是只要是有關這個公司的資訊或產品消息，都直接被歸類在這個項目，因而導致結果都偏向這個項目。

另一方面，我們也嘗試對一些專有名詞來做實驗，例如：Macaca cyclopis、Blechnum orientale L…等等。因為這些都是相當專門的名詞，在查詢的結果來看，內容的方向是大致相同的，不像前面一般的名詞會出現相當多不一樣的分歧和分類。經由人工的分類結果，可得知動植物專有名詞的分類少，且大部份分在同一類，由此結果我們可得知動植物專有名詞類的搜尋結果大部分都在同一類，又因為我們取得是 Google 的前 100 筆結果，這 100 應為 PAGERANK 排名的 100 名，其被瀏覽的機率也最高。因此我們可推論對於大部份人而言，這種動植物專有名詞是不需要去做分群的，我們便可根據如此的分析去將查詢分為較需分群的以及較不需分群的。

5. 參考書目

Adelberg, B., NoDoSE—a tool for semi-automatically extracting structured and semistructured data from text documents. Proc. of the 1998 ACM SIGMOD international Conf. on Management of data (SIGMOD), 1998.

Ahonen, H., Heinonen, O., Klemettinen, M., Verkamo, A. Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery, Proceedings of IJCAI'99 Workshop on Text Mining: Foundations, Techniques and Applications, 1-9, 1999.

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., Shadbolt, N. R. Automatic Ontology-Based Knowledge Extraction from Web Documents, IEEE Intelligent Systems, Jan-Feb, pp.14-21, 2003.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., Rajagopalan, S. Automatic Resource List Compilation by Analysing Hyperlink Structure and Associated Text, Proceedings of the 7th World Wide Web Conference, 1998.

Cheng, P. J., Teng, J. W., Chen, R. C., Wang, J. H., Lu, W. H., Chien, L. F. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval, Proceedings of the 27th ACM International Conference on Research and Development in Information Retrieval (SIGIR), July, 2004.

Cohen, W., Recognizing Structure in Web Pages using Similarity Queries. The National Conf. on Artificial Intelligence (AAAI), 1999.

Feldman, R. and Dagan, I. KDT - Knowledge Discovery in Texts, Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, 1995.

Feldman, R., Aumann, Y., Amir, A., Kloesgen, W., and Zilberstien, A. Maximal Association Rules: a New Tool for Mining for Keyword co-occurrences in Document Collections, Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997.

Hearst, M. Untangling Text Data Mining, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

Hsu C. N., and Dung, M. T., Generating Finite-state Transducers for Semi-structured Data Extraction from the Web. Information Systems, 23(8):521-538, 1998.

Kao, H.-Y., Lin, S.-H., Ho, J.-M. and Chen, M.-S., "Mining Web Information Structures and Contents based on Entropy Analysis," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 16, No. 1, January 2004.

Kao, H.-Y., Lin, S.-H., Ho, J.-M. and Chen, M.-S., "Entropy-Based Link Analysis for Mining Web Informative Structures," Proc. of the ACM 11th International Conf. on Information and Knowledge Management (CIKM-02), November 4-9, 2002.

Kim, S., Alani, H., Hall, W., Lewis, Paul H., Millard, David E., Shadbolt, Nigel R., Weal, Mark J., Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web, In Proceedings of the Workshop on the Semantic Authoring, Annotation & Knowledge Markup conjunction with the Fifteen European Conference on Artificial Intelligence, France, 2002.

Kleinberg, J. Authoritative Sources in a Hyperlinked Environment, Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 46(5), 604-632, 1998.

Kosala, R. and Blockeel, H. Web Mining Research: A Survey, ACM SIGKDD Explorations, 2(1), 1-15, 2000.

Kushmerick, N., Weld, D., and Doorenbos, R., Wrapper Induction for Information Extraction, Proc. of the 15th International Joint Conf. on Artificial Intelligence (IJCAI), 1997.

Laender, A., Ribeiro-Neto, B., Silva, A., and Teixeira, J., A Brief Survey of Web Data Extraction Tools, SIGMOD Record Vol. 31, Number 2, June 2002.

Lin, W. Y., Lam, W., Learning to Extract Hierarchical Information from Semi-structured Documents. Proc. of the ACM 9th International Conf. on Information and Knowledge Management (CIKM), 2000.

Lu, W. H., Chien, L. F., Lee, H. J. Translation of Web Queries using Anchor Text Mining, ACM Transactions on Asian Language Information Processing, 1(2), 159-172, 2002.

Lu, W. H. Term Translation Extraction Using Web Mining Techniques, PhD thesis, Department of Computer Science and Information Engineering, National Chiao Tung University, 2003.

Lu, W. H., Chien, L. F., Lee, H. J. Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach, to appear in ACM Transactions on Information Systems, 2004.

Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.

Soderland, S. Learning to Extract Text-based Information from the World Wide Web, Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997.

Wang, J. H., Teng, J. W., Cheng, P. J., Lu, W. H., Chien, L. F. Translating Unknown Crosslingual Queries in Digital Libraries Using a Web-based Approach, Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries (JCDL), 108-116, 2004.

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., Ma, J., Learning to cluster web search results, Proceedings of the 27th SIGIR, 2004.