行政院國家科學委員會專題研究計畫　成果報告

利用現存之分類架構針對特殊知識網路自動產生分類典之研究

計畫類別：個別型計畫
計畫編號：NSC93-2218-E-002-138-
執行期間：93 年 11 月 01 日至 94 年 07 月 31 日
執行單位：國立臺灣大學工程科學及海洋工程學系暨研究所

計畫主持人：黃乾綱

計畫參與人員：徐天威，許智均

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 10 月 31 日

一、 中文摘要

本研究計畫的目的在設計一套自動化的方式讓各色組織能夠快速收集網路上的相關資訊，並且依照組織的需求及指定的分類架構，建立有利於企業或組職發展的知識庫。這一套自動化的設計中，將利用網路上已經存在的多個網頁分類架構，如 Yahoo 和 DMOZ，以及搜尋引擎，收集企業及組織知識架構相關的網頁。依照網頁本身的內容及互連的結構，進行網頁分群。不同於過去的做法，我們希望分群的結果並不是只反映了內容的相似性，而是能夠依照企業的預設架構，並參考網頁上已經標記過的人工分類資訊，如 Yahoo 的分類架構，或是特定專業人士的註解，將網頁的分類更貼近企業知識管理的需求。本研究計畫突破過去必須利用訓練資料(Corpus)將文件做分類的方式，企圖在僅有不完整的預設知識架構下，整理網路上的資訊以建立資料庫。

關鍵詞： 自動抽辭、文件分類、機器學習

二、 英文摘要

The object of research project is to study on the automatic approaches that can help government, enterprise or any organization collect the information on the web, and establish a knowledge database. This study will try to design such an automatic mechanism that will utilize the existing classification hierarchical – like Yahoo, DMOZ, and search engines to collect information interested by the organization. With measuring the similarity and checking the link structures between page contents, the collected pages will be clustered with proposed Guided Clustering Method. Compare with previous approaches, the proposed approach will exploit the existing class label from web directory site like Yahoo, or other metadata labeled by experts in our clustering method. This study will try to break the traditional limitation, which is building the classifier by training from lots of existing corpus, and organized the web information according to the special requirements of organizations.

Keywords: Term Extraction, Document Classification, Machine Learning

三、 前言

利用此研究計畫，我們整合過去資訊檢索、機器學習、與資料探勘等技術，並將網路上所能夠取得的，以及上述領域中所需要使用到的工具，進行整理。目前整理的工具包含了文字索引(Indexing)，中文搜尋(Searching)，網頁集取(Spidering)，分群法(Clustering)及機器學習(Machine learning)等工具。並將整理的工具用於本研究計畫的開發。未來期望在長時間的努力之後，能夠匯整出一套有用的開放源碼 (Open Source) 的知識管理工具。以使得在研究新的知識管理議題上，能夠有更穩固的基礎。

四、 研究目的

知識管理已經是現代政府和企業組織中重要的課題，目的就在提供政府及企業組織的競爭力。且知識管理也是學習性組織中，加快學習過程，提升學習成效的重要元素之一。但是各行各業對於知識管理所想要建立的知識架構並不相同，而且在知識架構建構的過程中，重度的依賴專業知識工作者的腦力付出，不僅無法達成知識管理提升組織競爭力的目標，其所耗費的成本也不斷的增加。

以人力來建立知識架構的過程是個複雜的程序。首先專家必須先從大量的相關資訊中萃取出具有代表性的概念(concept)。概念之間的關係，如同義、廣義、狹義等包含或是被包含的關係，也必須靠專家找出以建立階層式的架構。這是一個複雜而且長時間的過程。人工建立的知識架構，最大的好處就是在語意上比較容易被其他的人理解，而且能夠表現知識的抽象關係。

但是人工建立出來的知識架構並不容易維護，原因有二。第一，不同的人對於同一個知識架構的認知並不完全相同，因此如果想要增加人力來加快知識網的建立或是用以建立大範圍的知識架構，維持架構中各分類內容的一致性，就會變成很大的困擾。其次，網路資訊、組織內部資訊、以及組織對資訊的需求，都會隨著時間的流逝而快速演變。原有的分類，以及爲了分類所建立的辭典、索引典(thesaurus)等，都可能不再符合新的需求。如前所述，企業建立知識管理系統的目的是為了提升競爭力而非如公眾圖書館般

作知識的整理，因此如何讓企業能夠快速的吸收外界的變化，並調整其知識管理的架構，也是重要的課題。

過去幾年網際網路的發展雖然促進的文件分類或資訊整理自動化技術的進步，不過這些技術的發展多是走在兩個極端，一個是以企業內部文件的自動分類為研究對象。一個是以整個 World Wide Web 上資訊分類的研究為主要課題。反而對於如何將網際網路資訊，自動化的快速整合進企業知識管理系統，並用以改善分類架構的研究，反而著墨不多。

本研究計畫的研究目的，就在研究如何利用導引分群的方法(Guided Clustering Method)來自動產生分類典，協助政府組織或企業在建立知識管理系統的過程中，大幅的減少人力的負擔。並研究可行的自動改善知識管理架構的方法。

五、 文獻探討

目前國內外進行的研究以自動產生網頁及文件分類架構為主的相關研究議題，主要有下列二類：

1. 將網頁或文件自動分類至現存的分類架構。例如將網頁自動分類到 Yahoo 的分類架構之下。這一類的研究在處理典型的文件分類(Document Classification)的議題。研究人員利用各種統計或機器學習的方法來建立各種分類器。這些文件分類方法的比較，包括分類方法的效能，分類方法的延展性等，都在 Yiming Yang 一系列的論文中有嚴謹的討論[5,8,10,12,13,14,15,15]。針對網頁分類的研究中則進一部加入網頁的額外資訊，如超鏈結(Hyperlink)，連結文字(Anchor Text)或是 Meta-tag 的資訊來輔助網頁的分類[1,11]。

   目前文件分類的研究已經很成熟，只是所有的文件分類研究，都假設各分類項目之下，已經有可供訓練的文件內容(corpus)，不然就無法進行分類的工作。這對於常常僅有分類架構而缺乏訓練資料的知識管理系統來說，並無法提供幫助。

2. 將未分類網頁或文件自動產生出類似分類網站的分類架構。例如將搜尋的網頁，透過階層式分群法建立 Yahoo。這一類的研究在處理階層式分群法的(Hierarchical Clustering)的議題。階層式分群法在各種相關研究中並沒有改變基本的步驟，此類研究的方向都集中在特徵選取的問題上，藉由特徵選取方式的改變來建立不同取向的分類架構。

   這種方式可以將收集好的資料自動建立出分類的架構。此類方法的缺點在於建立分類架構的過程並無法加入人的參與。此外，建立出來的階層常常不容易讓人理解，做為資訊瀏覽的架構並非十分合適。因此這種方式並未真正用於建立目錄網站，主要是用於搜尋結果的整理。[9]

由於上述兩種方式的皆無法滿足企業或是政府組織想要建立彈性知識管理架構的特性，因此近年來部分相關的研究，便從文件分類的議題先轉而考慮如何建立分類典(Taxonomy)，也就是下段所述的第三類相關的研究。

從文件中抽取自動建立分類典(Taxonomy)或是分類用辭典(Dictionary)。這類研究不先處理網頁分類的議題，而是先建立分類典以利後續的網頁分類。分類典的建立可以透過從文件中抽取關鍵詞及資訊[6]。也可以由分析使用者所下的查詢辭(Search Query)來建立分類典[2,3,4]。分類典的建立初期仍然是需要人工建立的，只是建立之後後續的分類典擴充的工作可以利用統計或是機器學習(Machine Learning)等演算法達到自動擴充的目的。分類典建立的方法對本研究計畫來說是重要的工具，我們需要利用分類典將自動分群結果與使用者預先定義的知識管理結構做接合的工作。並且我們希望能夠改善目前分類典趨於靜態的特性。讓分類典能夠依特定領域的需求自動做調整。

六、 研究方法

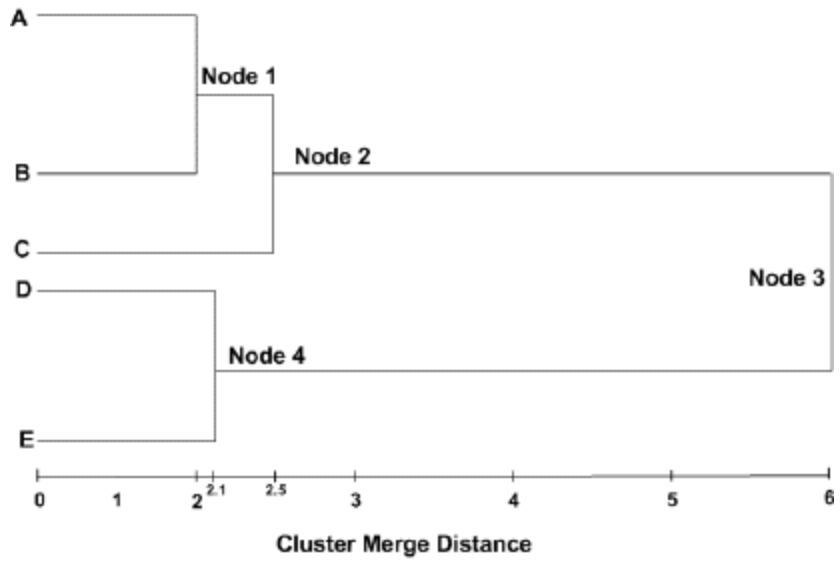
本計劃發展一個導引分群方法(Guided Clustering Method)。在設計導引分群演算法時，我們將整合利用兩種不同的資訊。第一是過去已充分研究並使用的內容相似度(Similarity)的資訊，第二是人工標記的 Metadata 資訊。

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，由下而上的階層式分群演算法(Hierarchical Agglomerative Clustering)仍然是最容易在僅有資料的情況下，用以建立分類架構的分群演算法。因此我們在設計導引分群方法(Guided Clustering Method)時，也將從階層式分群演算法的理論出發。

階層式分群演算法(Hierarchical Agglomerative Clustering)的概念如下：

1. 起始條件：將每個物件(Object)指定給不同的群組(Cluster)
2. 計算每兩個群組之間的距離(Distance)或是相似度(Similarity)
3. 將步驟 2 所計算的數值，建成一個矩陣(Matrix)
4. 找出距離最近(或是相似度最高)的兩個群組(A, B)
5. 從矩陣中移除並且合併這兩個群組(A, B)為一個新群組 C
6. 計算其他群組到這個新群組 C 的距離(或相似度)，並更新矩陣中的數值。
7. 重複步驟 1~6 直到矩陣中只剩下一個群組為止。

階層式分群演算法的計算結果，可以圖表 1 為例。



圖表 1 - 階層式分群演算法

階層式分群演算法的優點有下列幾項：

1. 可以顯示物件的距離或相似程度。在圖表 1 中，明確的看出 A 和 B 的距離比 D 和 E 的距離要近。
2. 可以產生比較小的群組，如圖表 1 中的 Node 1、Node 2、Node 4 等，對

於發現群組內和群組間的資訊很有幫助。

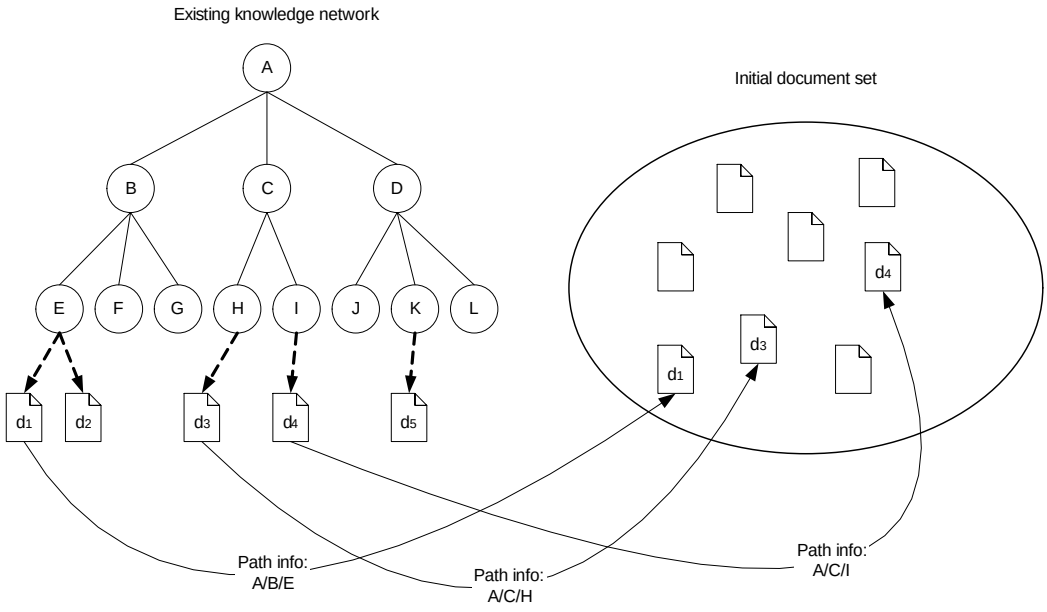
3. 使用不同的閥值(threshold)，可以產生從 1 到 N(物件總數)任意數量的群組。例如 threshold = 3 時可以產生 2 個群組，而 threshold = 2.3 時可以產生 3 個群組。

階層式分群演算法的缺點則如下：

1. 如果{A, B, C, D, E}為文件時，並不容易找出可以表達群組的辭彙。
2. 自動產生的結構，常常與資訊分類以及資訊檢索的習慣不相同。而且無法加入人工標記做結構的調整。

因此我們希望將人工標記的分類資訊或是 metadata，整合進分群演算法的設計中。網路上的人工標記資訊可以分成三類形態。第一種是別的網頁參考到該物件的連結文字(anchor text)。第二種是網路上現存的分類網站(Directory Search Engine)，如 Yahoo、DMOZ 等等。第三種是在其他網站上，特別是網誌(Blog)型態的網站，提到該物件的文字說明。

我們利用搜尋引擎或是網頁收集程式(Spider, Crawler)從網路上取得這三類的資訊。以第二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，如圖表 2 所示。

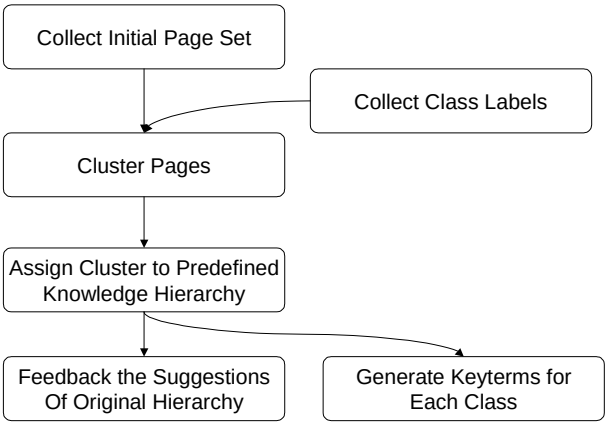


圖表 2– 將分類網站的標記資訊(分類目錄)與分類目錄下的網頁建立關聯

這些資訊將用特定的比例與文件的內容一起做階層式的分群計算。但是不同於過去的階層式分群計算過程。我們的目標並不是要自動產生一個分類階層，而是當知識管理架構的設計者，已經有初步的架構時，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，甚至能夠自動對原有的架構做調整的建議。

因此，分群的過程並不是以合併到剩下最後一個群組停止，而是根據既有的分類架構，來決定分群的中止條件。最後再將分群的結果歸納至既有的知識管理架構之下。如果有未能歸入的群組，則根據此群組與其他群組的詳細度，以及該群組內網頁的標記資訊，建議知識管理架構的設計者，是否要調整原有的架構。

整個方法的流程如圖表 3 所示：



圖表 3– 導引分群方法(Guided Clustering Method)的流程

七、 結果與討論

因為在研究計畫的發展過程中，有機會與法鼓山中華佛研所合作，以中華佛研所的佛學資料庫作為研究的對象。我們將網上佛學相關的資料以 spider 收集，查詢 DMOZ 及 Yahoo 上該網頁的分類位置，並利用抽辭工具及佛學大辭典將網頁中的佛學詞彙抽出。對照佛研所圖書分類的方式，建立網頁的自動分類機制。

目前最大的困難是，分類正確性的判斷尚難評估，因而正在與佛研所的專業研究者討論以人工的方式，檢查分類的成果。期望能夠有量化的數據出現。但由於佛學詞彙本身似乎有獨特性，因而本方法目前所呈現的效果尚屬可用的情況。

此外我們發現佛典文獻中的 quotation 和 citation 特性，應該要作為分類的重要考量因素。這一點，在一般文件中也並非不會發生。Quotation 和 citation 資訊如何在計畫初期並未被特別注意，因為一般網頁中的交互參照(reference)多以 hyperlink 的形式存在，但在文件中，更多的方式是以 Quotation 和 Citation 的方式出現，這對於知識管理的議題來說，如何判定文件中的文字是 Quotation 或是 Citation 是一個有趣的新議題。

八、 參考文獻

1. S. Chakrabarti, B. Dom and P. Indyk “Enhanced hypertext categorization using hyperlinks.” Proceedings ACM SIGMOD International Conference on Management of Data (pp.307-318), Seattle, Washington: ACM Press, 1998.
2. S.-L. Chuang and L.-F. Chien, “Automatic query taxonomy generation for information retrieval applications” Online Information Review (OIR), 27(4):243-255, 2003
3. S.-L. Chuang and L.-F. Chien, “Enriching Web taxonomies through subject categorization of query terms from search engine logs” Decision Support System, Special Issue on Web Retrieval and Mining, 30(1):113-127, April 2003.
4. S.-L. Chuang and L.-F. Chien, “Towards automatic generation of query taxonomy: A hierarchical query clustering approach” Proc. the 2002 IEEE International Conference on Data Mining (ICDM), pages 75-82, Maebashi City, Japan, Dec. 9-12, 2002.
5. R. Ghani, S. Slattery and Yiming Yang. “Hypertext categorization using hyperlink patterns and meta data” The Eighteenth International Conference on Machine Learning (ICML'01), pp 178-185, 2001.
6. V. Kashyap, C. Ramakrishnan and T. C. Rindflesch, “Towards (Semi-)automatic Generation of Bio-medical Ontologies” Poster Proceedings of the AMIA 2003 Annual Symposium, November, 2003, Washington, DC.
7. Daphne Koller, Mehran Sahami “Hierarchically classifying documents using very few words.” Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997.
8. F. Li and Y. Yang. “A loss function analysis for classification methods in text categorization” The Twentith International Conference on Machine Learning (ICML'03), pp472-479, 2003.
9. R.E. Valdes-Perez and etc, "Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results". Joint Conference on Digital Libraries (JDCL '01), Roanoke, VA (presented as a demonstration), June 24-28, 2001
10. Y. Yang, J. Zhang and B. Kisiel. “A scalability analysis of classifiers in text categorization” ACM

SIGIR'03, pp 96-103, 2003.

11. Y. Yang, S. Slattery and R. Ghani. “A study of approaches to hypertext categorization” Journal of Intelligent Information Systems, Volume 18, Number 2, March 2002.
12. Y. Yang “A study on thresholding strategies for text categorization” Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), pp 137-145, 2001.
13. Y. Yang and Xin Liu “A re-examination of text categorization methods.” Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99, pp 42--49), 1999.
14. Y. Yang “An evaluation of statistical approaches to text categorization.” Journal of Information Retrieval, Vol 1, No. 1/2, pp 67--88, 1999.
15. Y. Yang and J.P. Pedersen “A Comparative Study on Feature Selection in Text Categorization Proceedings” of the Fourteenth International Conference on Machine Learning (ICML'97), 1997, pp412-420.
16. J. Zhang and Y. Yang. “Robustness of regularized linear classification methods in text categorization” ACM SIGIR'03, pp 190-197, 2003.

九、 計畫成果自評

本計劃依照預定計畫完成的結果包含有：

1. 整理相關的資訊收集工具。包含 Spider/Crawler 程式，網頁分析工具。
2. 發展導引分群演算法(Guided Clustering Algorithms)。此演算法改進傳統的分群演算法，將單純的內容相似度以及人工標記的 Metadata 資訊整合在分群演算法的設計中。
3. 參與本計劃的研究生在資訊檢索(Information Retrieval), 資料探勘(Data Mining), 機器學習(Machine Learning)和文件處理(Document Processing)等方面獲得紮實的訓練。

但研究成果尚屬粗淺，所以還未發表於國內外的學術會議或期刊中。目前僅將初步成果與中華佛學研究所合作，作為佛學資料的整理之用。未來會先進一步把研究成果運用於計畫主持人所負責的教育部人才培育計畫的電子之事交換平台之上，並進一步探討導引分群演算法的實際使用特性。