行政院國家科學委員會專題研究計畫　成果報告

概念瀏覽介面的數位圖書館個人化文獻管理系統之研究

計畫類別：個別型計畫
計畫編號：NSC94-2416-H-009-029-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學圖書館

計畫主持人：黃明居
共同主持人：柯皓仁

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 9 月 27 日

行政院國家科學委員會專題研究計畫成果報告

概念瀏覽介面的數位圖書館個人化文獻管理系統之研究[1]

A Concept Browsing Interfaces of Personal Bibliographic Management System for Digital Library

計畫編號：NSC-94-2416-H-009-029.

執行期限：94 年 8 月 1 日至 95 年 7 月 31 日

主持人：黃明居，國立交通大學圖書館

E-mail: mjhwang@lib.nctu.edu.tw

一、中英文摘要

資訊科技的進步與數位資源數量快速地成長，數位圖書館的建置與個人化服務的發展亦日趨成熟。本研究提出以概念空間（Concept Space）為基礎的文件視覺與管理方式，讓使用者更能整體性的閱讀自己所蒐集的數位資訊。核心技術是以關鍵字分群的方式達到概念萃取的目的，且將文件以多種概念描述後，進行文件概念分群，並透過文件內引用文章(Citation Article)的相似度，建立文件間的引用關係，進而建立群與群之間的引用關係，達到建立概念之間的相關性。最後，取代傳統條列式的顯示方式，以視覺化的方式展現分群結果並呈現出概念之間的相關性。根據實驗結果分析，本研究所萃取出的概念可以適合地表達出文件的整體概念，在文件分群的準確率上亦有一定水準。

關鍵詞：數位圖書館、概念式瀏覽

This project have been developed a concept browsing of personal bibliographic management system, which labeling each group clearly concepts by an automatic concept extraction method. Concept extraction is accomplished through a keyword clustering algorithm. Then a document clustering algorithm is made on documents described with the extracted concepts. In addition, citations existing between documents are used to construct documents connections. Such connections are further used on discovering group relations and concept relations. Instead of representing search results by traditional lining style, a visualization system is developed to represent both search results and the relations of concepts. Several experiments done in this work show that not only concepts extracted by our method can clearly describe documents in the same group, but also achieve good clustering accuracy.

Keywords: Digital Library, Concept Browsing.

二、緣由與目的

隨著資訊科技的進步與網際網路的蓬勃發展，加上資訊數位化的技術逐漸成

[1]本研究原計畫書所規劃之部分人力被刪除，因此僅著重於「概念式之瀏覽介面」之相關課題研究，本執行結果報告書亦以此核心爲主軸作說明。

熟，數位內容與資源快速地成長，「數位圖書館(Digital Library)」資源日益豐盈，人們可以透過網際網路方便地獲取相關資訊與知識。然而，由於數位資訊的數量過於龐大(2002 年，全球人類產生 5 exabytes 的數位資訊，相當於 3 萬 7 千個美國國會圖書館（2002 年具有 1 千 7 百萬冊館藏)）[1]，在此資訊暴漲環境中，雖然已有許多效能優異之檢索引擎（如 Google 等），但是檢索結果之數量往往過於龐大而需要花費許多時間閱讀才能真正讀到自己所要的資料。因此，圖書館如何提供一個很有效率，且具備知識架構式的機制，以「概念式」顯示方式（Concept Browsing）呈獻給讀者，讓讀者很快的瞭解找到的資料中，整體的架構為何，以致於很快地找到自己所需的資料，是未來圖書館相當重要的工作之一。

另一方面，經過查詢所得的資料，與其他資料的關係為何？此資料在這領域的「知識地圖」(Knowledge Map)，是在那個位置，可以延伸的主題或資料有哪些？更重要的是提供一個友善（Friendly）閱覽介面？這些問題的解決方式，均為本研究主要之動機與目的。

三、研究方法與範圍

有鑑於「概念式」顯示方式（Concept Browsing）需將資料文件先行分群後，轉換成圖像的「概念式」顯示方式，因此本研究核心的研究方法為文件分群演算法，過去研究者均將文件內容轉換成向量空間 (Vector Space)，並透過距離表示文件間的相似度，將相似度高的資料群聚在一起。

一般而言，表達文件的方式通常是選取文件中較具有代表性的特徵，將這些特徵給予較多的權重，顯示它在文件當中的重要性。然而，單一個關鍵字並不足以表示整篇文章的概念，概念應由多個關鍵字組合而成；除此之外，有時文章會包含二個以上的概念，當兩篇文章的在多個概念上都有某種程度的相似時，即可以認定這兩篇文章是相似的。

本研究透過特徵選擇 (Feature Selection)方法，取出文件中重要的關鍵字，利用關鍵字分群的方式達到概念萃取的目的，且將文件以多種概念描述後，基於這些概念進行文件分群。最後，取代傳統條列式的顯示方式，以視覺化的方式展現分群結果並呈現出概念之間的相關性。

另外，本研究的研究範圍主要是專注於英文論文文件之概念萃取，而不考慮一般性的非學術文章，如新聞文件與一般網頁等；而在論文概念萃取方面，由於作者對於概念萃取沒有太大的幫助，為了簡化系統的複雜度以及處理人名的問題，因此不加入作者相關資訊，僅利用文件的標題與摘要進行概念萃取。

四、概念萃取之文件分群與視覺化

本研究提出的概念萃取之文件分群演算法，以及如何將分群結果進行視覺化的動作。概念萃取的方法與視覺化主要分為四個程序，分別為資料前置處理 (Data Pre-processing)、文件分群 (Document Clustering)以群聚後置處理 (Cluster Post-processing)及視覺化 (Visualization)，說明如下：

4.1 資料前置處理

本研究以 CiteSeer[25]內的論文做為語料庫，CiteSeer 是一個收集科學領域論文的數位圖書館，其收錄的主題相當多，包含 Architecture, Artificial Intelligence, Information Retrieval, Networking, Operating Systems, Theory 等，並且提供論文的相關欄位，其中包含標題(title)、作者 (author)、摘要 (abstract)、出版社 (publisher)及引用相關的資訊等。除了論文本身的欄位之外，CiteSeer 亦建立了引用索引 (Citation Index)，可以完整地記錄文章間的引用關係。

本研究選取標題、摘要及引用做為資料

來源，摘要部分所收錄的文字大約只有 1000 個字元，這個數量相當於以搜尋引擎透過關鍵字查找所得到的結果資料。在這種資訊不足的情況下，我們將研究如何準確並有效率的萃取出合適的概念。

在 CiteSeer 的眾多主題中，本研究選定較為熟悉的 Information Retrieval 相關論文做為語料庫，並針對文件標題與摘要進行前置處理，主要的前置處理方法包含斷詞切字 (Tokenization)、小寫化 (Lowercase)、刪除停用字 (Stop Words)、詞性判斷 (Part of Speech, POS) 、詞幹轉換 (Word Stemming)、片語化 (Chunk)等等[24]。

4.2 文件分群

經過前置處理將大部分文件內的雜訊或不必要的資訊刪除後，在這些經過資料清除 (Data Cleaning)後的字中，挑選出較能代表文件概念的字彙及片語，計算這些字彙與片語之間的相關度，並據以建立網路語意關係圖，再依據圖形理論 (Graph Theory)原理進行過濾及分群，以達到概念萃取的目的，也做為我們文件分群的基礎。其中包括：1.特徵選擇；2.概念萃取與特徵分群；3.語意相似度向量之文件分群等。

其中特徵選擇 (Feature Selection)挑選較具代表性的概念字詞，除了傳統方式依詞性及文件在語料庫內出現次數進行過濾之外，本研究再利用字詞間的相關度來進行特徵選擇。概念萃取的方法是將的 Topic Keyword Clustering 演算法[4]加以修改，以符合本研究的需要。圖一為概念萃取的步驟，包括：1. 建立特徵語意網路圖；2. 概念萃取之方法，為 *k*-Nearest Neighbor Graph Approach[6]；3.由特徵候選組產生候選特徵子群；4. 合併候選特徵子群；5. 修正並產生概念子群。

4.3 群聚後置處理

後置處理主要分為兩個方向，第一是進行群聚標記(Labeling)的動作配合文件分群的結果視覺化，第二則是以論文內的引用文章 (Citation Article)之相似度建立群與群之間的關係。

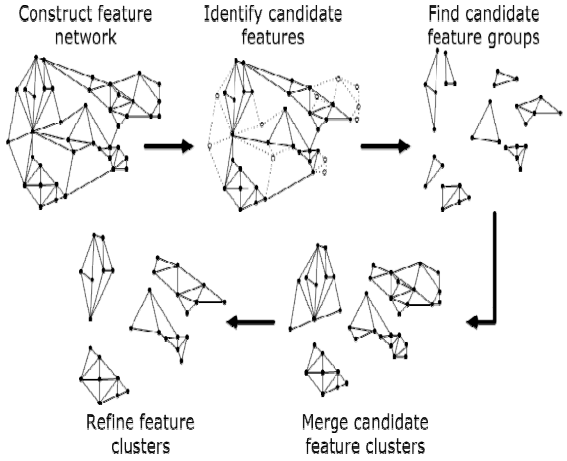

圖 一 概念萃取之步驟

本研究中，標記的方法是依據群中心點的語意相似度向量來選擇相似度最高的概念子群，並取概念子群內的特徵為候選標記特徵。在概念子群的特徵當中，由觀察得知名詞及名詞片語較能表達群的概念，且名詞片語的重要性又大於名詞。經過多次實驗及分析，本研究依照下列幾項規則來挑選特徵：

1. 選出權重前十名的特徵

2. 在這十項特徵中，優先選取名詞及名詞片語為候選標記特徵

3. 利用人力輔助過濾出有意義的特徵取權重最高的二項特徵做為最後群的標記

除了將分群結果視覺化之外，還有什麼資訊可以協助使用者了解分群的意義及概念，並且可以透過視覺化以供參考？由於本研究是以 CiteSeer 做為語料庫，而 CiteSeer 的特色在於其建立了完整的引用索引，利用此項特性，本研究先透過文件內引用文章的相似度，建立文件間的引用關係，進而建立群與群之間的引用關係，並將結果用於輔助視覺化。

4.4 視覺化

視覺化主要分為兩個方面，一為文件

的分群結果，本系統將每一個群都以一個圓表示之，其優點在於可以用半徑大小來表該群內包含的文章個數，並且用顏色的深淺表示該群內文章的同質性是否一致，顏色愈深代表同質性愈高，群聚內部緊密，如圖二所示。另一方面，為群與群之間的關係，本系統以雷達圖表現之，雷達圖的特性在於可以明確的描述群與群之間的關係強度，如圖三所示。

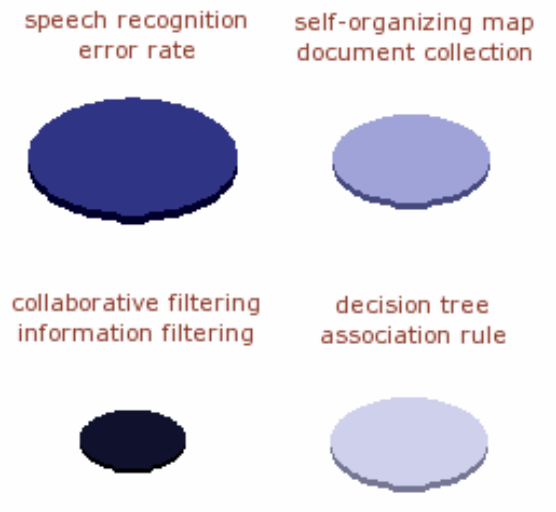

圖 二 分群結果之視覺化

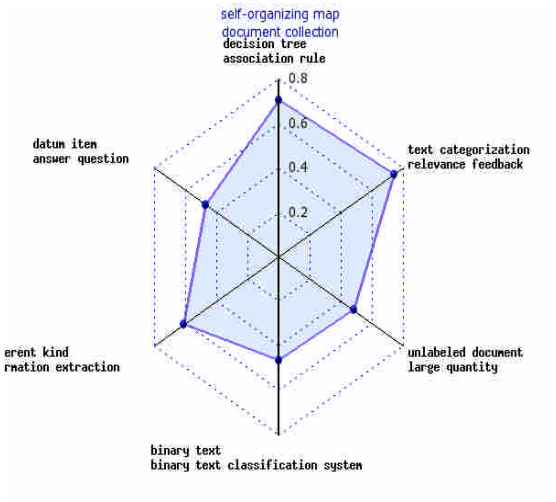

圖 三群與群關係之視覺化

五、研究結果與建議

本研究依據上述之方法與成果，建立管理系統，並且於最後分析影響文件分群結果的因素：主要分為兩方面，一為文件向量，二為語意相關度。文件向量係指以不同的方式產生供文件分群的文件向量。語意相關度則討論本研究使用的 Log Likelihood Ratio[6][10]及 Topic Keyword Clustering [4]使用的語意相關度計算方法。

至於評估分群結果的優劣，通常依據數學理論，計算群的內聚力及分離度判別分群結果的優劣；但是有時數據與人類的語意並不一定相符；因此，需要再以人工判別的評估方法當做參考。本研究中，人工判別的方式為隨機挑選文章組合做為樣本，並且請專家群對樣本標示相似度，最後和分群結果做比較。

實驗結果分析，本研究提出的以概念萃取為基礎之文件分群演算法，經多方面評估後都得到不錯的結果；故使用 log likelihood ratio 能夠建立正確的字詞相關度，經由概念萃取得出的概念也能正確的包含語料庫中的主題概念，並且以概念相似度表示文件，亦能準確的表達文件內容，自然可以提升文件分群的正確率。

六、計畫成果自評

本研究著重於二個方面：第一為以特徵分群萃取文件中所包含的概念，再依概念的相似度描述文章，最後利用文件分群演算法將文件分群；第二則是將分群的結果以視覺化的方式呈現給使用者。本研究中，視覺化系統除了以圖形化的方式展現分群結果，同時亦可顯示出群與群之間的關係。此研究過程所開發的分群與視覺化技術與成果，將整理撰寫成論文，正投稿中。

七、參考文獻

[1] Information on this study, "How Much Information?", along with the study's findings, are available on the project Web site: <http://www.sims.berkeley.edu/research/projects/how-much-info/>.

[2] P. A., Probability, Random Variables and

Stochastic Processes. ,Second Edition ed.New York: McGraw-Hill, 1984,

[3] R. K. BLASHFIELD, "Finding Groups in Data - an Introduction to Cluster-Analysis - Kaufman,l, Rousseeuw,pj," J. Classif., vol. 8, pp. 277-279, 1991.

[4] H. C. Chang and C. C. Hsu, "Using topic keyword clusters for automatic document clustering," IEICE Trans. Inf. Syst., vol. E88D, pp. 1852-1860, AUG. 2005.

[5] K. Chidananda Gowda and G. Krishna, "Agglomerative clustering using the concept of mutual nearest neighbourhood," Pattern Recognit, vol. 10, pp. 105-112, 1978.

[6] L.E.L.,Testing Statistical Hypotheses. Wiley, 1986,

[7] B. S. Everitt, Statistical Methods for Medical Investigations. ,2nd Edition ed.Edward Arnold, 1994,

[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000,

[9] J. He, A. Tan, C. L. Tan and S. Y. Sung, "On quantitative evaluation of clustering systems." in Clustering and Information Retrieval Anonymous 2003, pp. 105-134.

[10] J. Neyman, E.S. Pearson, "Joint statistical papers," 1967.

[11] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: A review," ACM Comput. Surv., vol. 31, pp. 264-323, SEP. 1999.

[12] G. Karypis, E. H. Han and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," Computer, vol. 32, pp. 68-+, AUG. 1999.

[13] D. A. Keim, "Information Visualization and Visual Data Mining," IEEE Trans. Visual. Comput. Graphics, vol. 8, pp. 1-8, 2002.

[14] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," in ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 170-178.

[15] J. R. Levine, T. Mason and D. Brown, Lex & Yacc. ,2nd ed.O'Reilly & Associates, Inc, 1992,

[16] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Math. Statist, Prob., vol. 1, pp. 281-297, 1967.

[17] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. 1999.

[18] G. Minnen, J. Carroll and D. Pearce, "Applied morphological processing of English," Nat. Lang. Eng., vol. 7, pp. 207-223, 2001.

[19] D. G. Roussinov and H. C. Chen, "Information navigation on the web by clustering and summarizing query results," *Information Processing & Management*, vol. 37, pp. 789-816, NOV. 2001.

[20] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv., vol. 34, pp. 1-47, 2002.

[21] M. Steinbach, G. Karypis and V. Kumar, "A comparison of document clustering techniques," *KDD Workshop on Text Mining*, 2000.

[22] Y. Yang, "Noise reduction in a statistical approach to text categorization," in SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 256-263.

[23] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," Mach. Learning, vol. 55, pp. 311-331, JUN. 2004.

[24] 劉政璋 Liu, "以概念分群為基礎之新聞文件自動摘要系統 Concept Cluster Based News Document Summarization," pp. 67, 民 94.

[25] CiteSeer - http://citeseer.ist.psu.edu/