應用乏晰概念網路於中文文件擷取之研究

The Study of Applying Fuzzy-Concept Network to Chinese Document Retrieval

計畫編號: NSC87-2213-E-009-087
執行期限: 1997/8/1- 1998/7/3

主持人:梁婷 執行機關:交通大學資訊科學系 職稱:副教授
e-mail: tliang@cis.nctu.edu.tw

一、摘要 (中英文)

關鍵詞: 乏晰概念網路、模糊語意、中文、文件擷取。

資訊擷取系統主要的目的在有效的滿足使用者的查詢所需。雖然目前多數商業資訊擷取系統仍然多以佈林邏輯模組為基礎，然而這類系統卻無法有效地表達語言中未臻明確的語意資訊，以致系統無法適切地解決這類的查詢。

一般解決這些語帶模糊的查詢，多要求先在查詢詞彙上明確定義其模糊和歸屬函數，再以分群法則依模糊門檻值分群處理以減少搜尋時間。然而這些方法在效率上尚未臻令人滿意。原因在於語言中一些語意概念間的訊息並未充分利用。

因此在本計劃裡我們一方面將研究中文詞類中可資利用的詞彙，建構模糊語意變數的語法結構和函數，以免除佈林式查詢的複雜邏輯運算，並可將資訊需求中的不明確性表達出來。另一方面在處理巨量文件分類問題上，我們希望能從乏晰概念網路的理論來研究其應用到中文文件查詢程序的可行性。基本上乏晰概念網路可解釋為繼承階層的集合。由於繼承階層結構早已應用在知識訊息表達上，對物件提供良好的分類法則，同時也易於執行。因此乏晰概念網路將可有效地表示在不明確條件下的概念間的關聯性值。利用概念間的相似性和衍生性關係，將有助於搜尋過程在繼承階層結構下做橫向或縱向推演。

關鍵詞:Fuzzy concept network, fuzzy linguistic, Chinese, text retrieval.

The primary goal of retrieval system is to effectively satisfy user's information request. Though most commercial systems are still based on the Boolean logic model, such model cannot efficiently solve the uncertain information in a user's query.

One general solution for uncertain information requests is to require explicit definition of the fuzzy terms and membership functions or do clustering with respect to different fuzzy threshold value to reduce the search space. However, these approaches still cannot solve the problem efficiently since the relationship among the concepts existed in the natural languages does not fully employed and expressed.

In this project, we will investigate those Chinese words which can be used in fuzzy linguistic variables and construct linguistic structure and associated functions so as to avoid the complexity in Boolean operations and handle the fuzziness in users' queries

appropriately. As to the text classification, we will investigate the feasibility of applying fuzzy concept network to Chinese text retrieval.

Basically fuzzy concept network can be interpreted as the collection of inheritance hierarchy. Since the inheritance hierarchy has been used as good knowledge representation of object classification and supports easy implementation, fuzzy concept network can efficiently express those relations under uncertain condition. By using the similarity and generalization relations among concepts, searching can be conducted as inferring among the inheritance hierarchy and be speeded up.

二、緣由及目的

隨著電子文件的蓬勃增長，快速有效的文件檢索擷取功能，將有助於發揮電子文件潛在的資訊價值、方便龐大資訊的管理，因此這方面的研究也就日益的重要。在英文文件檢索擷取的探討上目前已有相當丰富的研究〔1，2，3，4，11〕。而在中文文件擷取的處理研究上，學術上的文獻近年包括文件分類研究〔5，6，9〕以及中文全文檢索查詢系統製作〔7，8，10〕。

基本上資訊擷取系統可分成傳統的向量空間模式、機率模式、佈林式模式和非傳統的乏晰集合模式〔13〕。 在向量模式中，文件和類別分別以向量空間表示，由內積運算值決定文件的歸屬類別。此種分類基本上有賴於文件和類別的向量空間表示中關鍵詞的選取和夾角的計算法則。而機率模式的擷取方式由於是以統計假設為基礎，來建構所使用的數學模組，因而存在一些實驗結果與數學模組所推演出的結果不一致的現象〔4〕。是故這兩種的分類法目前仍多應用在實驗性的擷取系統上。

相對於向量空間和機率模式的擷取模組，以佈林式模組為基礎的設計目前仍為多數商業性的擷取系統所採用。這些系統設計多著重在提供便利的自然語句、關鍵詞、同音詞、同義詞和容錯查詢等一般查詢或個人化資訊服務等功能。所使用的擷取法則多建構在快速字串比對或以反轉法、特徵法為基礎的索引檔等傳統搜尋機制上〔6，8，9〕。然而對於查詢語言中未臻明確的語意資訊，以及語意概念間的訊息處理及其在文件分類與搜尋推理過程的運用，則尚需進一步的探討。

在解決語意模糊的查詢問題上，一般乏晰集合模式的擷取設計多要求先在查詢詞彙上明確定義其模糊和歸屬函數，再以分群法則依模糊門檻值分群處理以減少搜尋時間〔11〕。或者運用模糊集合理論中模糊語意變數的模糊分佈觀念（以一種連續分佈的重要程度值）來取代單一的權重數值（常用在加權式佈林模組中），以提高系統的使用度〔7〕。 然而這些方法在效率上尚未臻令人滿意，原因在於語言中一些語意概念間的訊息並未充分利用。

因此在本計劃裡，我們一方面將研究中文詞類中可資利用的詞彙，建構模糊語意變數的語法結構和函數，以免除佈林式查詢的邏輯運算，並可將資訊需求中的不明確性表達出來。另一方面在處理文件分類問題上，我們希望能從乏晰概念網路的理論來研究其應用到中文文件查詢程序的可行性。

乏晰概念網路基本上可解釋為繼承階層的集合[13〕。由於繼承階層結構早已應用在知識訊息表達上，對物件提供良好的分類法則，同時也易於執行。因此乏晰概念網路將可有效地表示在不明確條件下的

概念間的關聯性值。在本計劃裡我們將利用概念間的相似性和衍生性關係〔12，13〕，使搜尋過程在概念繼承階層結構下做橫向或縱向推演，以加速查詢過程。同時在概念網路架構下執行相關文件查詢將可導引查詢程序的實現，進而提高擷取輸出品質。我們希望經由這方面的研究對日益蓬勃的中文電子文件能有實際的應用處理。

三、結果與討論

本計劃主要包括語料分析處理設計、詞彙萃取系統設計、詞彙加權計算設計，乏晰概念網路建立、查詢模組設計、檢索模組設計、及文件相關查詢模組設計。測試用的語料是以資訊相關系所的碩、博士論文電子文件中的論文封面及中文摘要為主要實驗語料，並在工作站上執行本計劃。

在語料處理上我們首先將所蒐集到的論文資料予以格式化並標記基本的文件屬性作屬性查詢。再就摘要文件的全文處理部份，本計劃依據中文字結合能量、字頻、及字序訊息，以統計的方法迅速地萃取出摘要與標題中有意義的二音詞和三音詞彙。對於相對少數的四音詞和五音詞詞彙，若無法經由二音詞和三音詞彙組成，將以辭庫輔助萃取。辭庫的建構是以集合論文所附的關鍵詞和中研院的詞庫為主，未來再加入所萃取出來的新詞彙，並依所建的概念網路將詞庫中的詞彙做分類。

實際上，文件內涵表達和概念網路的建構均有賴於詞彙的萃取、詞彙數量的選取、以及詞彙加權函數的計算。文件內涵表達中的詞彙的萃取與數量的選取將從空間大小和詞頻為設計參數。而詞彙加權函數計算模組設計則考量詞彙出現的位置、詞彙長度和其與論文題目長度與摘要長度之比例。

在乏晰概念網路的建立上，我們利用在語料分析步驟中所得的詞彙和文件內涵間的關連訊息來建立乏晰概念網路，並利用概念間的相似性關係和衍生性關係作為搜尋推理的架構。基本上所設計的概念網路架構，最底層為碩、博士論文摘要，往上的幾層為包含這些論文的概念類別。概念間的相似性值或衍生性值，藉由詞彙對類別的關連性值和各類別間關連性值的計算得到。為了避免在概念網路中做檢索推演時，對一些路徑做重複不必要的推演，我們分別利用概念對概念矩陣及概念對文件矩陣來描述乏晰概念網路，將網路中的檢索推演轉換成矩陣間單純的算術運算，而矩陣內的元素即為概念和概念間的關連性值或概念和文件間的關連性值。

在查詢模組建構上，本計劃以兩種方法來設計。第一種方法是建構模糊語意變數的語法結構和其所需函數。選取中文語言中部分修飾詞和數量詞，以取代數字上的權重，使查詢方式更趨便利自然，並可將資訊需求中的不明確性表達出來。 另一種方法則是將查詢者的需求轉換為一個概念，並利用這個概念在乏晰概念網路中進行搜尋推演。查詢者可根據需求，選擇要看的某種概念論文，並根據需求程度對各個概念下重要性權重值或修飾詞；亦可利用查詢詞彙來表達所選擇的概念類別中有包含該詞彙概念的論文。同時也可利用權重來表示各詞彙的重要性程度。我們將計算詞彙在乏晰概念網路中各概念裡所佔的關連性值，並根據查詢者對各概念所下的重要性權重值做調整。

在檢索模組設計上，有別於一般檢索系統多以關鍵詞索引檔設計為主，我們的檢索模組將依據查詢者對概念網路中各個概

念的需求做概念檢索。概念檢索是以概念對文件矩陣（從概念網路建構步驟中所得）和查詢對概念矩陣（從查詢模組步驟中得到查詢者對各概念之需求），將這兩個矩陣做矩陣運算，我們即可得到一個查詢對文件矩陣。矩陣內的元素代表各篇碩、博士論文對查詢者的相關重要性程度值。查詢者可依所需察看相關論文的篇數，調整相關重要性程度門檻值。

在文件相關查詢模組設計上，此一階段主要是根據查詢者所選的某一篇論文，搜尋其它和此篇相關的論文；一般的文件相關查詢，有利用此篇文件的特徵表示和其它文件的特徵表示做相似性比對，取出吻合程度高的文件；或利用已建立的文件對文件矩陣來找出和該篇文件相似程度大者。我們所設計的文件相關查詢，主要是根據該篇論文對乏晰概念網路中各概念的關係程度值，依據概念間的相似性和衍生性關係來搜尋。

在計畫的成果方面，我們利用所建立的乏晰概念網路，設計概念式文件相關查詢並利用實際文件資料，與「關鍵詞式文件相關查詢」模組做比較。在不同「文件查詢次數」的實驗測試中，我們將「廣義關係門檻值」和「相似關係門檻值」固定為0.7，並將「文件輸出個數」定為20。經由實驗的結果得知（圖一），當查詢次數為1至10時，「概念式」明顯的優於「詞彙式」。同樣的，在不同「文件輸出個數」的實驗測試中，我們將「廣義關係門檻值」和「相似關係門檻值」固定為0.7。經由實驗的結果得知（圖二），當輸出個數為2時，「詞彙式」的表現優於「概念式」。這是因越少的輸出，其文件內容均為與所選文件含最多相同詞彙。一般在輸出個數增加時，概念式模組都優於詞彙是模組。

在語言查詢模組設計與製作上，我們利用史皮曼公式作為不同模組的查詢滿意度測量。從實驗結果，我們發現所提得改良式組合查詢法在不同的查詢個數與類型均較傳統組合查詢法有較高的滿意度(見圖三與圖四) 。在巨集條件式實驗中，我們發現在不同的查詢個數上巨集條件式亦優於傳統的佈林模組（見圖五）。最後在結合式模組實驗中，改良式查詢法不論在資料加權與否都要比傳統組合查詢法產生較高的查詢滿意度（見圖六）。

四、計畫成果自評

本研究計畫內容均依原提計畫之內容執行，並達到預定的目標，包括建立一個中文文件乏晰網路擷取系統(見圖七)和中文語言式查詢模組系統。計畫中所提方法與理論均有實際資料驗證其可行性。從實驗數據分析顯示本計畫所設計的概念式相關文件擷取確實較一般關鍵詞式模組提高使用者對資訊擷取的滿意度。同時，所提的語言式模組也較佈林式查詢便利，在不同的查詢狀況下均能提供較高的滿意度。

在未來研究上，針對概念網路的設計，我們希望利用經驗法則以避免因為文件個數的增加，而造成檢索效率的低落。對於儲存空間因文件增加而不斷地增大的問題，可考慮利用「壓縮」方式處理。

在語意查詢模組的研究上未來將加入斷詞程序。另外也將研究自然語言中其他可用的修飾詞以增進查詢模組的便利性。

本研究的成果已為兩位碩士畢業論文主要部份（參考文獻十四，十五）。同時部份成果已改寫成學術期刊論文（參考文獻十六）正在外審中。本計畫的執行實有助於圖書館自動化之實現。經費的補助亦提供實際執行所需。

五、參考文獻

[1] J. H. Lee, Myoung Ho Kim and Lee Yoon Joon, "Ranking documents In thesaurus-based Boolean retrieval Systems," Information Processing & Management, Vol.30, No.1, 1994, 79-91.

[2] W. B. Croft, " National center for intelligent retrieval," Communications of ACM, Vol. 38, No. 4, 1995, 42-43.

[3] C. H. Lin and H. Chen, "An automatic indexing and neural network approach to concept retrieval and classification for multilingual (Chinese-English) documents," IEEE Transaction on System, Man, and Cybernetics, part. B-cybernetics, Vol. 26, No.1, 1996, 75-88.

[4] W. S. Cooper, "Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval, " ACM Transactions on Information Systems, Vol. 13, No. 1, 1995, 100-111.

[5] 楊允言,"文件自動分類," 清大資科碩士論文, 1992.

[6] 林淑美,"財經新聞文件自動分類," 台大碩士論文, 1995.

[7] 趙建宏等,"模糊語意在中文全文檢索中之應用," 第六屆國際資訊管理學術研討會, 1995.

[8] 梁婷, "The Study of Character-based Signature Methods in Chinese Text Retrieval," 交通大學博士論文, 1995.

[9] 吳匡時，黃森原，林志清 ，"Automatic classification of Chinese documents," 全國影像處理研討會 1996.

[10] 簡立峰, "尋易(CSMART)的智慧型中文檢索系統," 中文資訊檢索技術及應用研討會(1996), 中央研究院資訊所。

[11] D. Lucarella and R. Morara, "FIRST: Fuzzy Information Retrieval SysTem," Journal of information Science, 17, 1991, 81-91.

[12] Lee-Kwang Hyuaan, Yoon-Seon Song and Keon-Myung Lee, "Similarity measure between fuzzy sets and between elements, " Fuzzy Sets and Systems, Vol. 62, 1994, 291-293.

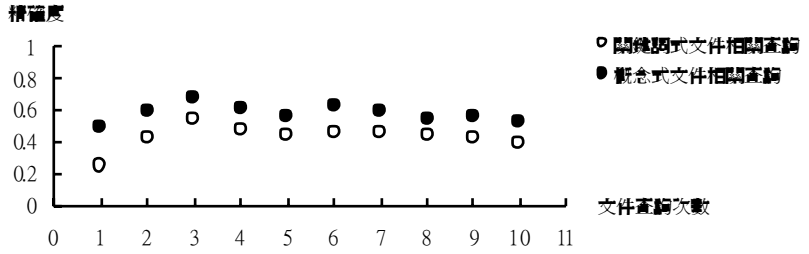
[13] S. M. Chen, and J. Y. Wang, "Document retrieval using knowledge-based fuzzy information retrieval techniques," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 25, No. 5, 1995, 793-803.

[14] 張慶權,"以乏晰概念網路為基礎的中文文件擷取研究," 交通大學碩士論文, 1997.

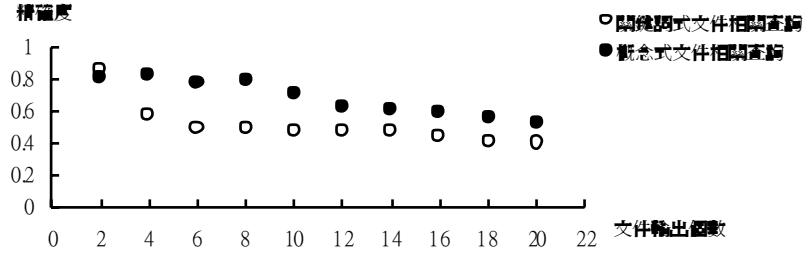
[15] 吳子強,"應用於中文文件擷取中的語言查詢模式設計與製作," 交通大學碩士論文, 1997.

[16] Tyne Liang and Ching-chyuan Chang, "Chinese Textual Retrieval Based on Fuzzy Concept Network (submitted)."

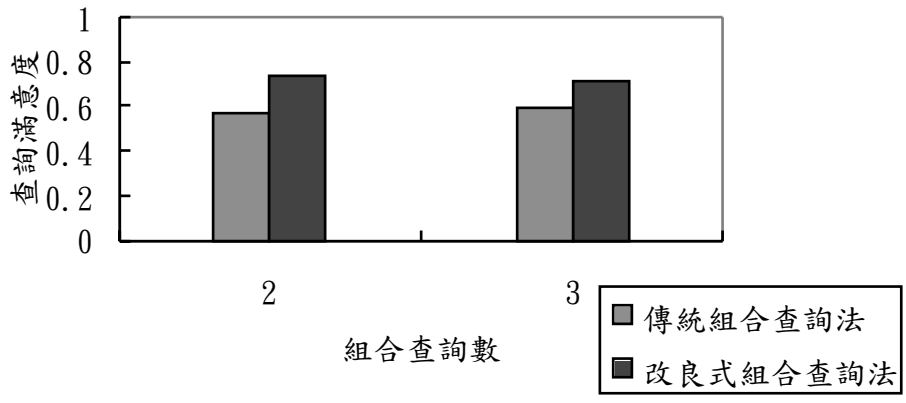
. . 圖表



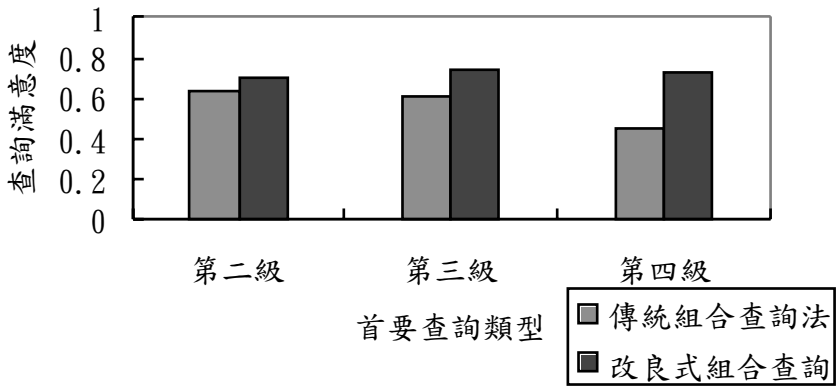
圖一： 不同文件查詢次數的精確度表現



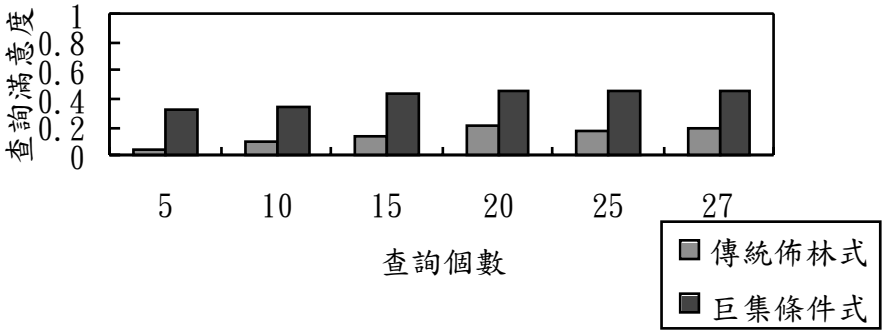
圖二： 不同文件輸出個數的精確度表現



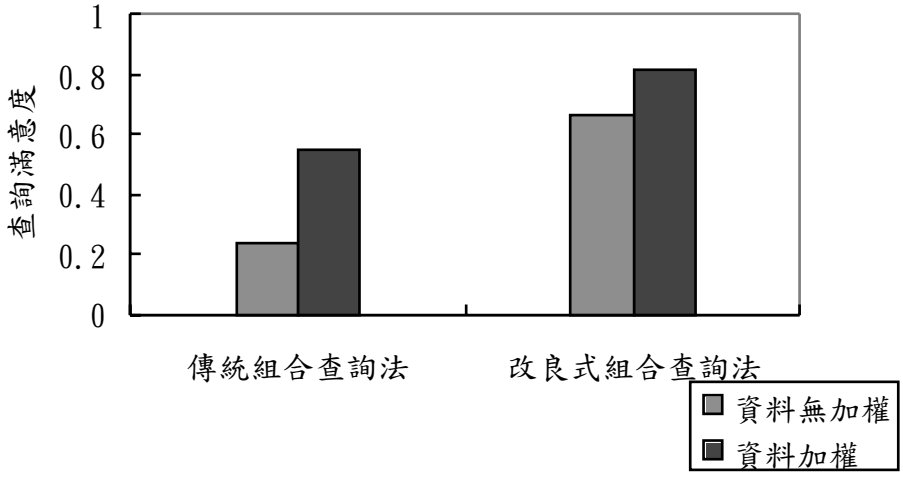
圖三: 語意權重式實驗分析(I)



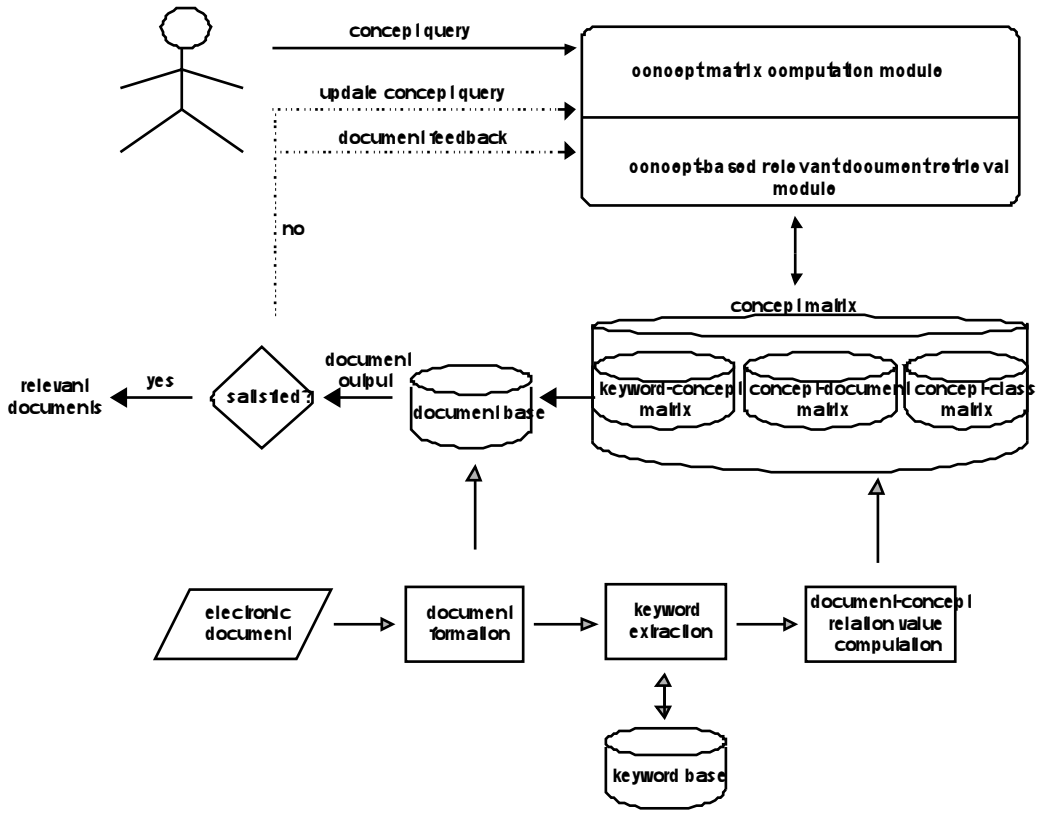
圖四: 語意權重式實驗分析(II)



圖五: 巨集條件式實驗分析圖



圖六: 結合式實驗分析圖



圖七：中文文件乏晰網路系統