行政院國家科學委員會專題研究計畫成果報告
中文文本超鏈結自動建構方法之研究
The Study of Automatic Construction of Hyperlinks in Chinese Texts

計畫編號：NSC 89-2213-E-002-067
執行期限：88 年 08 月 01 日至 89 年 07 月 31 日
主持人：陳信希　　國立台灣大學資訊工程學系
計畫參與人員：　　丁永偉

一、中文摘要

近幾年來，文本數位化的現象日漸普遍。大量數位化的資料透過網際網路，傳播各處。數位圖書館，電子書，...，也提供不同型態、素材的資料，大量機讀資料捶手可得。如何輔助使用者在浩瀚的資料堆中，快速準確的找到所要的資料，是研究的重點。資訊檢索和瀏覽是目前使用者擷取資訊的兩大模式，由於瀏覽模式對使用者來說具有簡單方便，一目瞭然的特性，很適合用在創作電子書、建立數位化圖書館、以及網頁製作等應用上。在這種情況下，我們有必要針對如何建立一個好的 ”導覽架構”，作進一步的探討。本計畫報告提出文本間超鏈結的自動建立的方式。

關鍵詞：中文語言處理，自然語言處理，超文本建構，資訊檢索，語意模型，瀏覽系統。

Abstract

Digitizing is very common recently. Much digitized information is disseminated elsewhere through Internet. Digital libraries, electronic books, and so on provide multi collections. How to help users extract information from large-scale digitized database quickly and accurately becomes an important issue. Information retrieval and browsing are two basic models for acquiring information. Browsing is simple and convenient for users, thus the model is useful for many applications such as electronic books, digital libraries, web pages, and so on. How to develop an elegant browsing model is worthy of studying. This project report presents the automatic construction of hyperlinks in Chinese texts.

Keywords: Browsing System, Chinese Language Processing, Hypertext Construction, Information Retrieval, Natural Language Processing, Semantics Model.

二、緣由與目的

由於近幾年來網際網路的蓬勃發展，愈來愈多的文件以 HTML 的格式呈現在網路上，而文件與文件之間也可以藉由超鏈結（Hyper-link）互相串連，形成一個文件網路，因此一份文件可以藉由超鏈結連結到其它相關網頁，使用者可以藉由一份文件的超鏈結無限延伸視野。諸如此類的情形如數位圖書館(Digital Library) 、電子百科全書、新聞鏈結 …等，愈來愈多的應用使得我們必需正視文件之間鏈結該如何建立。

由於使用者由一份文件瀏覽至下一篇文件時，是一種 ”導覽模式”，也就是文件的產生者建立一些超鏈結，使用者藉由這些超鏈結的導引，閱讀相關的其它文件。由於這種、”導覽行為”的日漸普遍，我們覺得有必要對這種導覽行為，以及如何建立一個好的 ”導覽架構”，作進一步的探討。

由於現在流通於網路上的文件數以萬計，並且隨時在不停地更動，要如何能夠對大量的資料進行快速有效的分析，建立合宜的導覽架構，並且能反映文件動態的增刪，也是一個刻不容緩的議題。

在文件鏈結系統的建構上，有兩個最主要的問題，第一個是如何找出文件中可能的鏈結點，第二個是這些鏈結點該連結到那裡; 整體來說, 這牽涉到使用者所感興

趣的主題以及如何建議使用者去尋找相關的資料，在後面的部分，我們會以 “實體” 這個詞彙來表示一個能的鏈結點。

自動鏈結技術所必須直接面對的問題就是實體的辨識，包括偵測與標定邊界，以及在不同文件的實體間建立鏈結，而如何偵測一篇文件中的重要實體，實體的重要性與文件主題有沒有關係，都是我們要分析的問題。另外我們也會分析文件類別與實體型態，並且找出常常出現的實體型態，以決定這些重要的實體型態對於文件鏈結的影響程度。

由於我們不能預測所要處理的文件數量與型態，因此，我們大概也無法期待能由所要處理的文件中訓練得出什麼資訊，所以一個強健又能保持一定精確度的實體辨識流程是我們所要致力完成的。而文件鏈結的精確率與回收率也是影響系統效能的關鍵，我們也會去分析它，並嘗試找出影響它的幾項原因。

若要建立文件之間的鏈結，首先必須決定鏈結所必須扮演的角色，一般來說，若是以文件相似程度為鏈結存在與否的標準，則鏈結的存在意味著兩篇文件有相同或相似主題，若是以詞與文件的相似程度為鏈結存在的依據，則鏈結的存在便代表著文件的主題就是該詞代表的意義。M. Agosti 等人 [Agosti, Colotti & Gradenigo, 91] 曾提出過文件之間可以建立 4 種鏈結，即文件與文件鏈的鏈結(D-D links)、索引點與所引點的鏈結(T-T links)、文件與索引點的鏈結(D-T links)以及索引點與概念間的鏈結(T-C links)，我們在此提出一個以建立實體(Entity)與實體之間的鏈結方法，而鏈結的存在表示有兩個相同實體存在，或者兩個實體的表現行為相似。

三、結果與討論

3.1 文件分析部份

一篇文件的實體通常是一篇文件的基本元素。使用者可由這些實體拼湊出整份文件所要表達的內容，因此使用者也很有可能要對這些實體做更深一層的認識，因此，選擇實體，就變成幫助使用者選擇相關資訊的一個重要步驟。那到底什麼樣的字串該被當作實體呢？在我們的考量之下，只要是使用者有可能感興趣的地方，都應該被視為實體。在這樣的考量之下，我們發現一些事實，人名、組織名及事件名幾乎都會是使用者有興趣的地方。又如地名在大部份的文件中，會是使用者感興趣的資訊。其它的一般名詞，在不同類型的文件中，有著不相同的情況，例如：”石頭” 在旅遊類的文件中似乎不會引起使用者興趣；但是若在礦物類，就很有可能引起使用者興趣。再如： ”天王星” 在物理或天文類的文件中幾乎是很顯眼的字串，不管 ”天王星” 是文件主要內容還是隨便提提。

有了以上的瞭解，我們覺得如果要選擇實體必需先對文件類別做分類，再根據文件類別選擇某些類別的鏈結點當作鏈結處。在我們的模型中，我們根據中研院 Sinica Corpus，得到了 67 類的文件類別，以及由梅家駒等人所編之“同義詞詞林”，得到了 1364 類的實體類別，而每一類的文件，都有對應的實體類別串列，對於任一類別的文件只有對應於該類別的實體串列的字串才有可能成為可能的鍊結點。

就如前面所提，由於人名、組織名與地名在大部份的文件都是重要的實體，並且根據實驗分析，人名、組織名與地名在文件中也站有相當高的比例，因此，以專門的方法來辨識這些實體是絕對有必要的，因此我們使用了中文人名辨識模組、英譯人名辨識模組、地名辨識模組與組織名辨識模組與 N-gram 模組來辯識這些常出現且重要的實體。

除此之外，我們也提出了一個辨識其它實體的方法，我們期待這樣能使得在實體的辨識上能夠更為完備。

3.2 導覽部份

由於目前資料的流通非常快速，因此，系統應該要能對資料的變動作出適當的反應且及時的反應，才可以讓使用者的檢索發揮最大效率。

為了能夠即時且正確的建立鏈結，系統的作法是將所有實體概念資料存放在實體資料庫內，一開始時，所有的實體概念的狀態(status)都是 0 (SLEEP)，也就是沒有任何的鏈結，此時，系統會逐一檢查每一

個實體概念看看是否有其它的實體概念擁有相同的實體字串，如果有，就把擁有相同實體字串的實體概念設定為活動 1 (ACTIVE) 狀態, 這些處於活動狀態的實體就可以互相鏈結，在系統將實體概念設定為活動狀態的同時，也意味著系統必須修改實體所在的檔案，使得該實體在檔案中呈現高亮度，並可接受使用者點選。

對於之後新加入的文件，系統會先去辨識文件中的實體, 形成實體概念, 之後，系統採取相同的作法，也就是去撿查新的實體概念是否有在實體資料庫中有沒有對應的其它實體概念，如果有，變去更動相關實體概念的狀態與修改文件。

對於文件的刪除，系統的作法是，在刪除一篇文件時，同時也會刪除處於這篇文件中的所有實體，此時系統必須對該文件中的所有實體作一次撿查，如果一個實體處於沉睡狀態，則系統可以忽略，如果處於活動狀態，意味著有其它的也處於活動狀態的實體也會鏈結到本篇文件，這時必須在去檢查會鏈結到本篇文件的實體數目，如果不只一個，系統還是可以忽略，如果只有一個，系統必須將該實體狀態設定為 0 (SLEEP) ，以免該實體鏈結到一個不存在的文件。

對於暫時隔離文件的做法，與刪除文件的處理流程接近，但是系統必須保存一份背份，以作為將來回復之用。

根據我們實際使用導覽系統的經驗，過多的鏈結點對使用者來講除了有一種不舒服的感覺外，更糟糕的是出現密度較低的鏈結點會被密度高的鏈結點所 "遮蔽"，也就是使用者自然而然不會去注意密度較低的鏈結點，這對於使用者來說，無疑是一項資訊的損失。因此，系統在標示鏈結點的時候，也會考慮到鏈結點密度的因素。 當一個實體被系統標示為鏈結點後，下一個相同的實體至少要離該鏈結點 500 Bytes 才會被標示為鏈結點。

待文件分析部份完成後，就可以進入導覽模式瀏覽文件了部份。使用者可以由任一起始文件開始，藉由系統選擇的鏈結點往下繼續瀏覽其它文件了。 瀏覽其它文件的方式是由使用者點選一個鏈結點，之後系統便會將與此鏈結點相關的文件呈現給使用者。本系統是多重鏈結模式，所以呈現的文件數目可能不只一個，一般的情況，使用者首先看到的不是文件本身，而是文件名稱、文件類別、文件摘要等相關資訊。由於文件之間是透過相同的實體產生鏈結，也就是說，若文件之間存在著鏈結，那表示文件內都有提到同一個實體 (日期 、人 、地 、組織 、思想 、理論) 而兩份文件之間也有可能透過數個不同的實體互相鏈結。

3.3 實驗與分析

本實驗是針對新聞類的文件所作的分析。我們收集了 30 篇關於體育的新聞稿來作分析，新聞來源包括 民生報、聯合報以及中國時報，每一篇報導都先經過處理，將不必要的 HTML 標頭去除，而每一篇文件的平均大小為 2.3 K bytes，大約 1000 字。

實體擷取部份，實體回收率：所有分析文件內共有個 382 實體，只有其中 302 個被標示出。 因此實體回收率為 79.0 %。實體精確率：系統共針測到了 395 個實體，但只有其中 302 個是真正的實體，因此實體精確率為 76.4 %。鍊結點部份，鏈結點回收率：所有分析文件內應該要有個 184 個鏈結點，系統共找出其中 162 個。因此鏈結點的回收率為： 88.0 %。鏈結點精確率:系統共建議了 178 個鍊結點，但其中只有 162 個是正確鏈結點。因此，鏈結點的精確率為： 91.0 %。實體搜尋部分，此部份是根據正確的鏈結點所做的分析，也就是系統所找出的 162 個正確的鏈結點。以我們的正確答案來看，這 162 個正確的鏈結點的鏈結數總和為 437 個，也就是說平均一個鏈結點可以鏈結到 2.69 個相關實體, 而系統建議的鏈結數總和為 373 個，而鏈結點對真正相關實體的平均回收率是 85.3 %，也就是有 14.7 %的實體因為實體字串與鏈結點字串不同而無法產生鏈結。

對於鏈結到不相關實體的原因，我們做了更深入的分析：大部分的情況是實體之間擁有相同的字串但畢竟是有所不同的。我們根據無法百分之百回收相關實體的鏈結點作了進一步的分析。無法百分之

百回收相關實體的鏈結點共有 38 個，這 38 個鏈結點的鏈結數總為 125 個，也就是說平均一個鏈結點應該要鏈結到 3.29 個相關實體，而系統建議的鏈結數總共為 57 個，而鏈結點對真正相關實體的平均回收率是 46.5 %。而這 38 個鏈結點中，其中是組織名的有 28 個，地名有 6 個，人名的有 2 個，比賽或會議有 2 個。

對於實體回收率部分，雖然有 88%的實體都能被回收，但是我們仍然發現只有實體字串不同，就很難產生鏈結，僅僅是期待由文件之內找到相同實體的不同外型（實體字串）似乎會遇到許多的瓶頸，尤其文件的來源不同，對一個實體的描述方法可能差異很大，因此，由語言處理上的技術去搜尋一個實體的特徵，再去比較這些實體特徵的相似性，應該是有必要的。

由於這裡所分析的是新聞的文件，即使這些新聞的來源不同（中國時報、聯合報、民生報）但是對於各種實體的表達方式仍有相當的一致性，所以系統對於實體的搜尋仍能保持相當的正確率與回收率。不過可以發現到的是，即使對於實體的描述一致，但是實體也會有簡稱與縮寫，這裡所碰到的主要問題也是在簡稱與縮寫部分，完全不同的描述方法在這裡似乎影響性並不是很大。

四、計畫成果自評

整體而言，研究內容與原計畫所列的工作項目完全相符，並已經達成預期的目標，所提出的整合系統適合在學術期刊或會議上發表。

五、參考文獻

Agosti, M.; Colotti, R.; and Gradenigo, G. (1991) “A Two–Level Hypertext Retrieval Model for Legal Data,” *Proceedings of SIGIR*, pp. 316-325.

Agosti, M.; Crestani, F. and Melucci, M. (1996) "Design and Implementation of a Tool for the Automatic Construction of Hypertexts for Information Retrieval," *Information Processing and Management*, 32(4), pp. 459-476.

Agosti, M.; Crestani, F.; and Melucci, M. (1997) "On the Use of Information Retrieval Techniques for the Automatic Construction of Hypertext," *Information Processing and Management*, 33(2), pp. 133-144.

Allan, J. (1995) *Automatic Hypertext Construction*, Ph.D. Dissertation, Department of Computer Science, Cornell University.

Allan, J. (1996) "Automatic Hypertext Link Typing," *Proceedings of the ACM Hypertext'96 Conference*, Washington, DC, pp. 42-52.

Allan, J. (1997) "Building Hypertext Using Information Retrieval," Information Processing and Management, Vol. 33, No. 2.

Baron, L., Tague-Sutcliffe, J., and Kinnucan, M.T. (1996) "Labeled, Typed Links as Cues When Reading Hypertext Documents," *Journal of the American Society for Information Science*, 47(12), pp. 896-908.

Chen, Hsin-Hsi; Ding, Y.W. and Tsai S. C. (1998) “Named Entity Extraction for Information Retrieval,” *Computer Processing of Oriental Languages* (accepted).

Chen, Hsin-Hsi; Ding, Y.W.; Tsai C. J. and Bian, G. W. (1998) “Description of the NTU System Used for MET2. ” *Proceedings of 7th Message Understanding Conference*.

Chen, Hsin-Hsi and Lee, J.C. (1996) “Identification and Classification of Proper Names in Chinese Texts,” *Proceedings of COLING96*, pp. 222-229

Green, S.J. (1997) *Building Hypertext Links in Newspaper Articles Using Semantic Similarity*, Technical Report, Department of Computer Science, University of Toronto.

Green, Stephen J. (1997) “Automatic Link Generation: Can We Do Better Than Term Repetition?” *Proceedings of WWW Conference*.

Mahesh, Kavi (1997) “Hypertext Summary Extraction for Fast Document Browsing," *AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, pp.95-103.

Silverstein, Craig and Pedersen, Jan O. (1997) “Almost-Constant-Time Clustering of Arbitrary Corpus Subsets,” *Proceedings of ACM SIGIR*, pp.60-66.

Shin, Dongwook; Nam , Sejin and Kim, Munseok (1997) “Hypertext Construction Using Statistical and Semantic Similarity,” *Proceedings of ACM International Conference on Digital Libraries*, pp.57-63.

Tebbutt, John (1998a) "User Evaluation of Automatically Generated Semantic Hyperlinks in a Heavily Used Procedural Manual," http://www.itl.nist.gov/div894/894.02/works/papers/ user_eval.html.

Tebbutt, John (1998b) "Finding Links," *Proceedings of the ACM Hypertext'96 Conference*, http://www.itl.nist.gov/div894/894.02/works/papers/ finding_links.html.

Thistlethwaite, P. (1997) "Automatic Construction and Management of Large Open Webs," *Information Processing and Management*, 33(2), pp. 161-173.

梅家駒; **竺一鳴**; **高蘊琦**; **殷鴻翔** (1993) ***同義詞詞林***, **東華書局**, **台北**, 1993.

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　　　**中文文本超鏈結自動建構方法之研究**　　　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫

計畫編號：NSC89－2213－E－002－067－

執行期間：88 **年** 08 **月** 01 **日至** 89 **年** 07 **月** 31 **日**

計畫主持人：陳信希

執行單位：國立台灣大學資訊工程學系

中　華　民　國八十九年十月二十二日