尋易(Csmart-II)：智慧型網路中文資訊檢索系統

(An Intelligent Chinese Information Retrieval System for the Internet)

簡立峰，李明哲，陳明權，陳宏明，陳丁旗，
董惟鳳，李宏俊，黃敦怡，張元貞

中央研究院資訊科學研究所

E-mail: lfchien@iis.sinica.edu.tw

摘要

考慮網路中文電子資源的有效利用，我們將原本針對中文全文檢索設計的尋易(Csmart)系統發展成新一代 Csmart-II 智慧型網路中文資訊檢索系統，包括整合資源發掘與過濾、資訊檢索與語音介面技術。使得 Csmart 系統可以以語音或鍵盤輸入近似自然語言查詢檢索網路即時新聞、BBS 論壇、中文 Web Pages 等網路資源。本文是有關 Csmart 系統的整體設計以及部份新發展技術的簡介。

一. 前言

隨著網際網路(Internet)的快速成長，網路上的電子資源，舉凡電子郵件、網路新聞、Web Pages、電子期刊、電子書等成長的相當迅速。為了使這些資源充分利用，網路資訊檢索系統 (Network Information Retrieval System)需求大為增加，包括Lycos, Infoseek, Alta Vista, Excite等在短短兩三年間陸續發展出來[1,2]，對網際網路的使用者而言這些系統無疑地是資訊瀚海的領航員，藉由這些系統的協助，網際網路資源的運用更加發揮。遺憾地，這些系統都是針對英文世界使用者設計的。

考慮中文電子資源的有效利用，適合網際網路資訊服務的高效率中文資訊檢索技術研究是相當迫切的。可惜國內已知的僅有中正大學的GAIS系統[3]、中央大學的CHARVEST系統等少數研究團體從事這方面的努力。這和英語世界動輒成千上萬研究人力物力的投入差距相當大。為此在過去一年中我們將原本針對中文全文檢索設計的尋易(Csmart)系統，由一般檢索核心程式(Searching Engine) [4]擴充具有網路資源發掘與簡易過濾能力，並且增加許多針對中文特性設計的檢索功能以及語音查詢技術，因此發展成新一代智慧型網路中文資訊檢索系統 [5]。

目前Csmart系統除了可以檢索包括電子辭典、建築文獻、佛學書目與摘要、產業技術報告等一般文件資料庫，也可以開始檢索網路即時新聞、BBS論壇、中文Web Pages等網路資源。

為了發展網路中文檢索，我們研究出許多技術包括 Fast Full-text Search [6]、Approximate Text Search、Qasi-Natural Language Query [7]、Speech Retrieval [8]、Word-based Text Search、Relevant Sentence Extraction、Relevance Feedback、Filtering and Subject Dissemination等，且在系統功能與結構設計上也仔細考量中文特性。這中間多數都是全新的嘗試。我們覺得這之中一些經驗可以提出來供大家參考，然而由於Csmart系統整合許多技術，限於篇幅本文嘗試僅就Csmart系統的整體設計及部份新發展技術作一簡介。進一步瞭解Csmart的技術內容可參考相關技術文件或利用Web Browser 連結至 http: // csmart.iis.sinica.edu.tw/。

二. 系統架構

Csmart 的系統架構如圖 1 所示包括 3 個子系統：資源發掘與過濾 (Resource Discovery and Filtering)、資訊檢索(Information Retrieval)與語音介面(Speech Interface)。為了發展網路檢索， Csmart 的研究朝兩方向發展。一個方向是網路資源的充分利用。我們初步設計了 Robot 程式可以在網路上自動發現有收藏價值的資源，藉此我們開始收錄網路中文 Web Pages ， BBS 論壇以及即時新聞並提供檢索，藉由 Robot 技術的發展我們得以開啓網路資源的眞正利用。以計算語言學的角度，我們可以取得源源不斷語料庫，可以統計不同領域的語言差異；以資訊服務的角度，我們有機會整合各個圖書館書目資料庫發展虛擬圖書書目檢索，整合網路新聞，發展虛擬新聞檢索；以語音辨認角度，我們可以建立豐富語言模型，發展語言模型調適技術， Client/Server 方式的語音辨認；最後以資訊檢索的角度，傳統文件分類、資訊抽取、資訊摘要、使用者行為分析等研究，由於有豐富的的資源與使用者，也因此可以較深入發展。

另一方面，我們開始研究語音與自然語言人機介面。由於我們發現許多智慧型檢索功能，如自然語言檢索，由於中文輸入的困難而無法發揮；另外我們發現智慧型的檢索技術多數必須藉由良好的人機互動才能展現，因而我們嘗試發展語音檢索技術與設計人機互動式查詢功能，目前我們以金聲三號為基礎已完成允許使用者以說話的方式詢問 Csmart 系統的語音檢索技術(Speech Retrieval)[8]。我們發現中文單音節特性使語音檢索技術相當接近實用程度。這對發展中文口語交談系統將是很好的開始。

為配合網路資源利用與語音自然語言人機介面的需求，在資訊檢索方面我們發展很多新的技術。舉例，由於網路傳輸不便，檢索結果必須有更佳的精確率以

減少不必要的網路傳輸。為此，我們除了持續加強近似自然語言檢索的檢索精確率與 Ranking 能力，也發展相關文句擷取功能，可使得檢索出的文件有更好的提示以提高檢索結果的可讀性，還有新增 Relevance Feedback 技術，允許從選取的查詢結果自動產生更精確查詢，以及新增主題自動選粹技術，讓使用者自訂新聞主題，而系統隨時主動提供相關新聞。這些技術的開發使得 Csmart 檢索功能的豐富如表 1 的比較說明與國際著名系統相較並不遜色[1,2]。

綜合上述整個Csmart系統運作如圖 1 說明。首先以左下使用者為中心，使用者可以透過網路連接Csmart，以線上(On-line)檢索方式選擇資料庫並輸入查詢以檢索Csmart所收錄的資源，另外也可以預先輸入查詢以建立個人資料檔 (User Profile)並以離線 (Off-line)方式檢索，當Csmart在網路上發現相關資源會以E-Mail方式主動通知使用者讀取。使用者在檢索時可以選擇以打字或者語音輸入。使用者可選擇的檢索功能如表1舉例說明包括邏輯查詢(Boolean Query),近似字串查詢(Approximate Query)，自然語言聯想查詢(Qasi-Natural Language Query)等，另外對檢索出的結果也可要求標示相關程度、顯示相關文句、標記相關字串直接檢索、以Relevance Feedback方式進一步檢索、以語音合成方式唸出檢索文件，以及建立個人資料檔案 (User Profile)。

另外，以資源為中心，如果是授權的特定資源如電子辭典、電子書等可以將全文儲存在 Csmart 主機，加以收錄整理提供線上檢索；若是網路資源如 BBS 、網路新聞、 Web Pages 則須透過 Robot 技術加以收錄、過濾、抽取提供資訊檢索模組建立索引。不論特定資源或網路資源如果須利用語音檢索，則須建立領域語言模型。所以資訊檢索子系統一方面須收錄資源發掘與過濾系統傳送的資源，另一方面又須提供使用者包括線上與離線以及打字與語音輸入等不同檢索。上述功能在目前 Csmart 系統都已提供。圖 2, 3, 4, 5 是有關使用者使用各種檢索方式舉例。其中打字輸入與語音輸入分屬不同介面。

三. 資源發掘與過濾

資源發掘與過濾系統是 Csmart 系統新的努力方向。如圖 6 所示資源發掘與過濾系統事實上包括發掘 (Discovery)、過濾 (Filtering)、抽取 (Extraction)技術。資源發掘主要是利用所謂 Information Spider 或 Robot 技術遊走網路發現值得收藏的資源。基本上網路資源種類很多有 Web Pages 、FTP 文件、News Groups 、BBS 等等，不同類型資源其收錄方式不一樣。以 Web Pages 收錄為例，必須利用 Hyper-link 有效遊走網路，避免收錄重複或品質差的資源、另外對有興趣的資源還必須有效加註(Annotation)。目前 Csmart 在 Web Pages 檢索方面還在實驗階段，因為中文資源相對英文還很少，因此所發展的 Robot 並不須時常上網收集資料以免造成網路擁塞。我們對有興趣的資源的摘要內容如圖 7 所示。

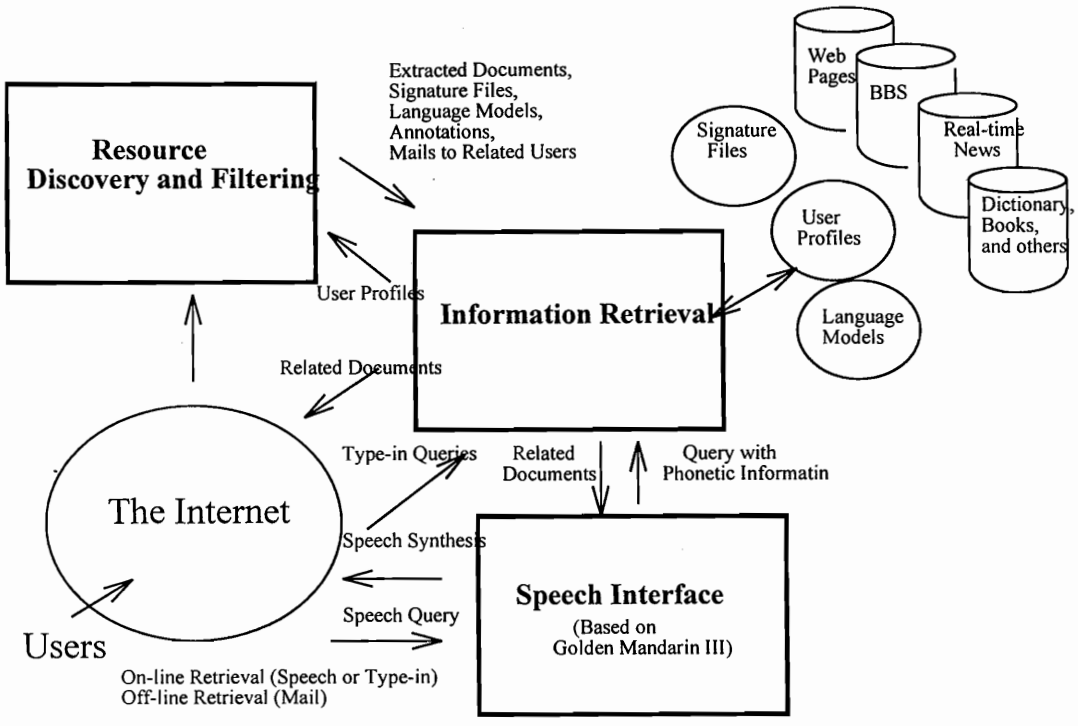

圖 1 Csmart 系統架構圖

至於資訊抽取部份，Csmart 會對收錄文件建立文件特徵(Signature File) [7]與語言模型[8]。前者提供資訊檢索子系統使用，後者是一種改良式馬可夫語言模型(Markov Language Model)提供語音介面提高辨認率以及資資訊系統作進一步特徵抽取分析之用。還有資訊過濾部份，和多數網路檢索系統一樣我們的研究還在起步階段[9]。我們以建立 User Profiles 方式進行資訊過濾，User Profiles 內主要是 User Query ，E-mail 帳號，以及從使用者提供或選定的文件(進一步說明可參見第六節) 抽取文件特徵。以即時新聞為例，使用者可輸入"國泰人壽"的查詢，並選取若干篇文章如有關國泰人壽除權、國泰人壽人事異動等，Csmart 系統嘗試從中抽取特徵，當有相關新聞即可提供使用者讀取。我們很關心資訊過濾技術的研究，包括資源自動分類(Classification) [10]、關鍵詞抽取 (Keyword Extraction)、個人化資訊服務(Personalized Service) [11]等。因為網路資源過多，如無有效分類使用者檢索負擔大；對收錄文件未能抽取出關鍵詞，則以全文檢索查詢成千上萬文件，檢索精確率會很低；還有每個使用者所需資源不同，長遠看檢索系統必須對不同使用者有不同檢索策略。目前 Csmart 在這方面的研究還須不斷加強。

	Functions/Systems	Lycos	Alta Vista	Excite	Csmart	Note
Indexing	Inverted File/Signature	Inverted File	Inverted File	Inverted File	Signature	
Language	English/Chinese	English	English&2- byte	English	Chinese&English*	
Searching Functions	Boolean Query	Min/Max	AND/OR/NOT	AND/OR	AND/OR	
	Approximate Query	Y	Y	Y	Y	工業技術研究院/工研院
	NLQ and Ranking	Y	Y	Y	Y	最新內閣名單/禁止英國牛肉國家
	Field Searching	Y	Y	Y	Y	標題/作者/關鍵詞/全文
	Relevance Feedback			Y	Y	
	Speech-Input Query				Y	
	Word String Match				Y	腦科/電腦科學
	User Profile and Mailing				Y	
Searching Results	Scoring	Y	Y	Y	Y	
	Relevant Sen. Extraction			Y	Y	張德培的排名/張德培擊敗..排名世界..
	Speech Synthesis				Y	
Information Extraction	Spider	Y	Y	Y	Y	
	Annotation of Web Page	Y	Y	Y	Y	
	Auto Keyword Ext.			Y	Under Developing	
	Information Filtering			Y	Under Developing	

表 1 Csmart 技術與功能和國際著名系統比較

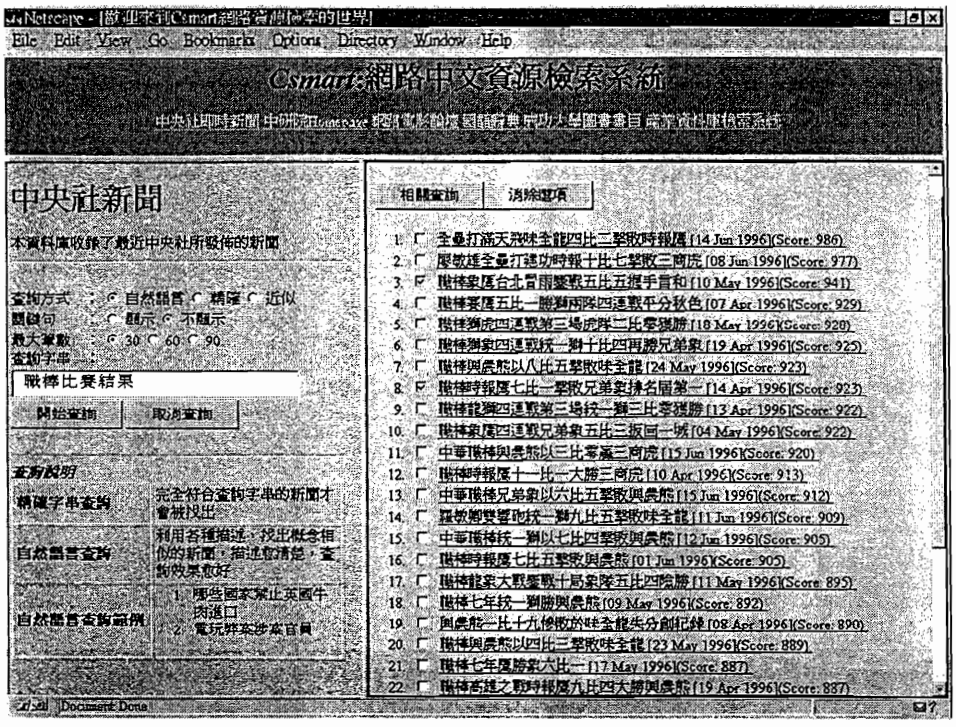

圖 2 Csmart 系統畫面 （以近似自然語言與 Relevance Feedback 檢索即時新聞）

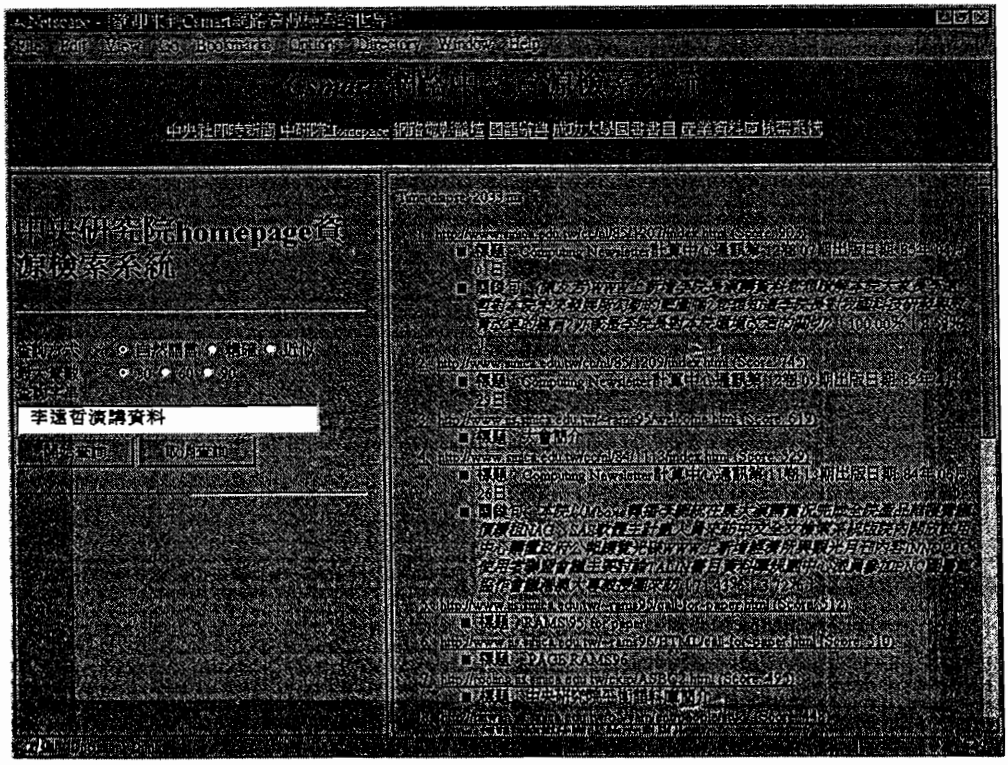

圖 3 Csmart 系統畫面（以近似自然語言檢索 Web Pages 資料庫）

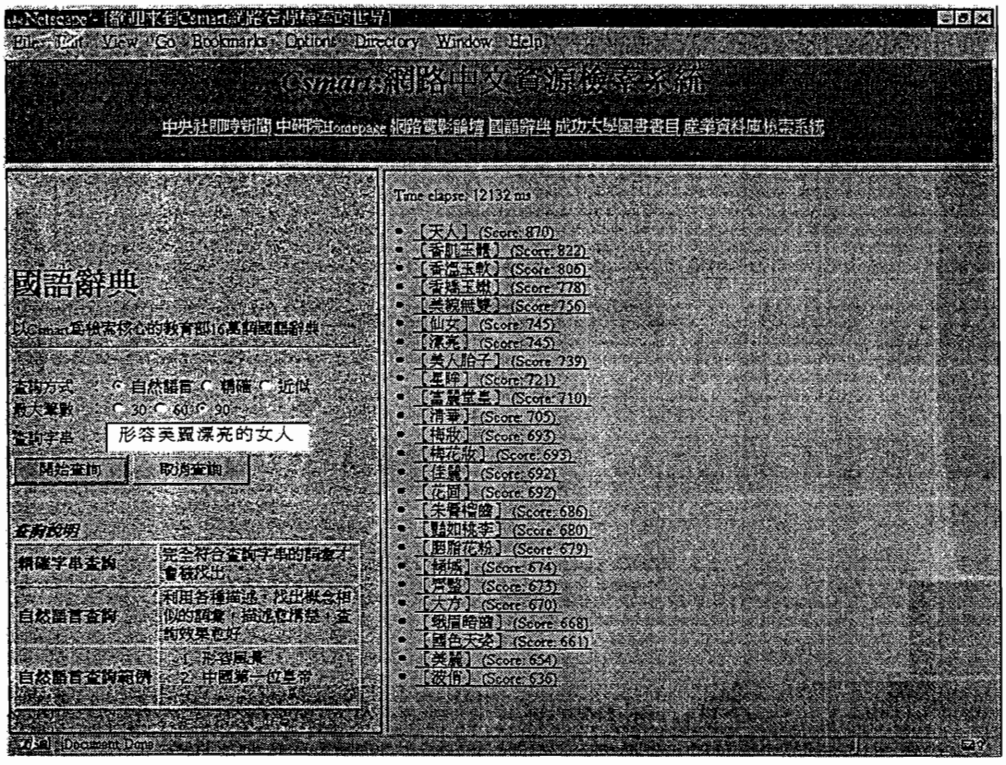

圖 4 Csmart 系統畫面 (以近似自然語言檢索國語辭典)

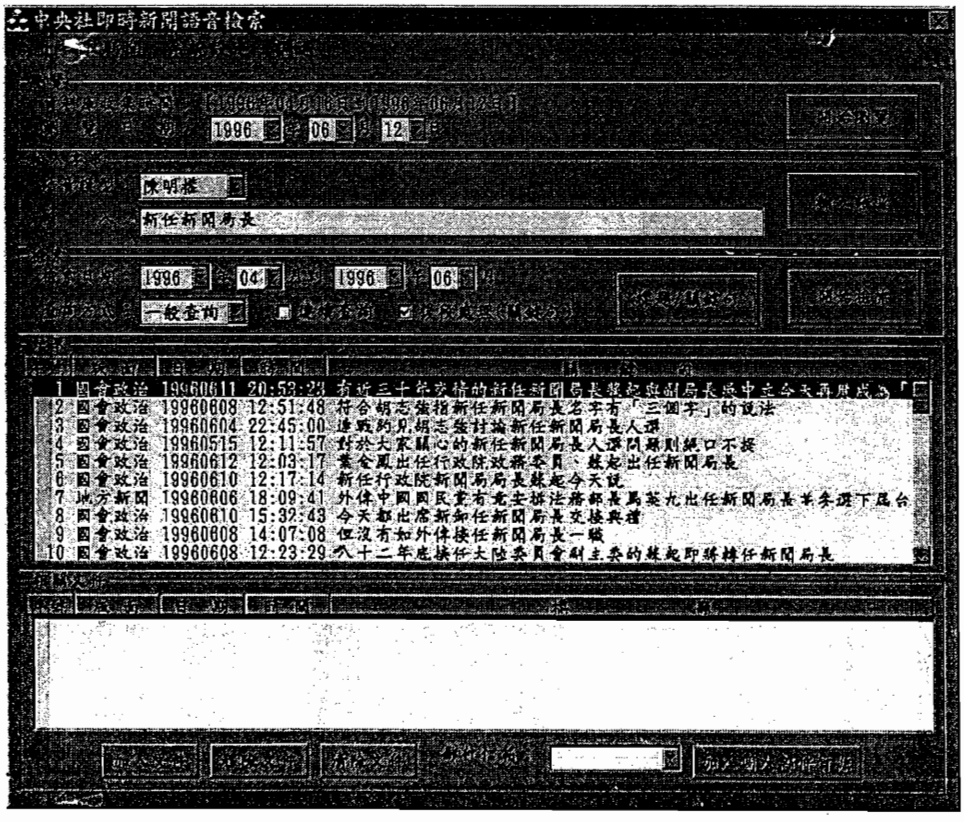

圖 5 Csmart 系統畫面 (以語音輸入查詢檢索即時新聞)

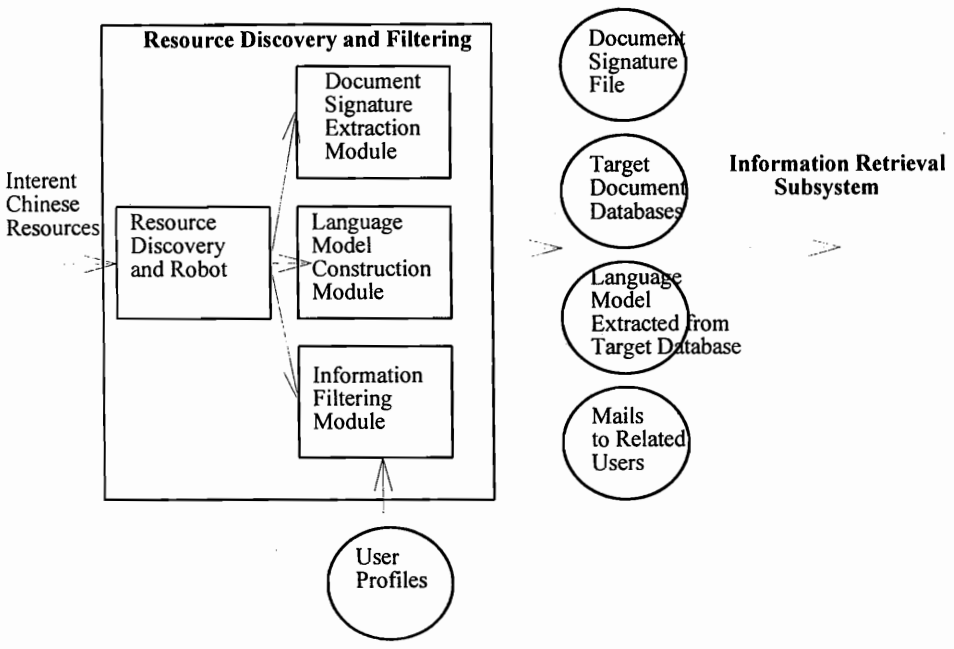

圖 6 資源發掘與過濾子系統

1. the URL
2. the first 200 Characters in the header fields
3. 20 lines or 20 percent of the document
4. number of other pages with links to this document
5. at most 20 lines of hyperlink text from other pages with links to this document
6. extractable keywords (under research)
7. the date the last downloaded
8. the date the last modified
9. document size in bytes

圖 7 Csmart 對收錄的 Web Pages 抽取的資料

四. 資訊檢索 -- 二段式搜尋

資訊檢索子系統是 Csmart 中比較有基礎的部份。Csmart 所使用的檢索技術是以特徵檔(Signature File)為核心的二段式搜尋。根據我們的經驗中文資訊檢索所須克服至少包括檢索詞的收錄與認定，專有名詞與新詞的抽取與斷詞歧異的解決，以及在近似檢索功能的提供 [12]。為此我們發展出以特徵檔(Signature File)為核心的二段式搜尋[7]。

這兩段式搜尋機制主要是將索引比對(Index Matching)與文件比對(Text Matching)分開以克服中文不易使用詞索引以及前述困難。這個方法如圖 8 所示包括第一階段以特徵檔為主的 Fast Search 以及第二階段以文件比對為主的

Detailed Search。由於中文斷詞困難，詞索引建立不易。因此我們覺得索引比對主要作用只是加速過濾多數無關文件，只要索引在比對時有很高的召回率且比對效率高，索引記錄的訊息可以較為模糊。因此我們發展出字層次的特徵檔技術。我們所發展的特徵檔搜尋方法與英文方法接近，不過在設計特徵擷取方式(Signature Extraction Method)時[13]，除了考慮資訊過濾程度，也要考慮給每個Signature Bit有較高語意蘊含以便施行近似檢索。至於文件比對部份，我們發展具備斷詞能力的比對技術，由於第一階段已把多數無關文件過濾掉，因此可利用豐富辭典內容與語言知識，施行精確斷詞以及近似分析，最後將正確文件選出。

在這種搜尋架構下，不論使用者選擇邏輯查詢、近似字串查詢或近似自然語言查詢都必需先產生查詢特徵(Query Signature)。若是邏輯或近似查詢即交由精確或近似比對搜尋程式處理，這包括第一階段先將查詢特徵和所有文件特徵比對，未滿足該查詢特徵的文件將被濾掉。未被濾掉的文件內容在第二階段將會被讀出及與查詢仔細比對，眞正滿足該查詢文件才會檢索出。若使用者是以近似自然語言方式查詢文件，則將交由最佳比對搜尋程序處理。在第一階段該程序會將查詢特徵與所有文件特徵一一比對估算出其查詢與文件之初步相似度(Relevance Value)，相似度足夠高的文件才會在第二階段繼續處理。第二階段基本上是將這些文件內文讀出，並對查詢句子進行關鍵語抽取，以及仔細比對這些關鍵語在相關文件中出現的頻率、位置與加權，以進一步判斷其相似度，最後相當相似的文件才會檢索出。事實上目前愈來愈多的東方語言檢索機制採用特徵檔技術，我們相信兩段式搜尋技術對東方語言檢索非常合適[14-16]。

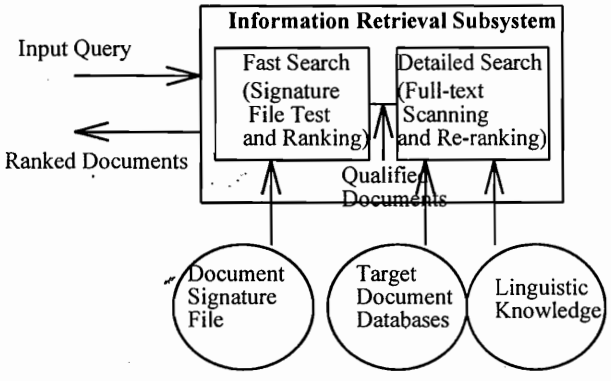

圖 8 Csmart 的二段式搜尋架構

五. 資訊檢索 -- 近似自然語言查詢與應用

Csmart 檢索技術最特殊之處應該是近似自然語言查詢技術的設計。早先我們觀察一般檢索系統成效發現多數使用者並不會使用邏輯查詢而且很多時候中

文查詢無法精確表達，如以圖書書目查詢為例當使用者不確定作者名稱，如"索忍尼新"或"索忍尼津"，書名記憶不精確，如"中國文化基本教材"或"中華文化基本教材"，出版社不能肯定，如"台視文化"或者"臺視文化"，這中間可能有簡稱，外來語言，也可能讀者記憶不清等，查詢這些資料如無高效率近似檢索功能則會形成相當不便。因此我們希望發展出的近似自然語言查詢滿足以下條件：

1. 能允許無限制(Non-constrained)的輸入查詢字串
   也就是查詢字串允許出現非控制字彙(Noncontrolled Vocabulary)
2. 檢索機制能有相當容錯能力
   這包括能檢索近似字串、容忍些許資料登錄錯誤、詞類變化等
3. 檢索出的結果能夠依據與查詢相關程度依序排列，且檢索結果相當合理

我們所發展的近似自然語言查詢方法基本概念包括以下幾個步驟：

1. 以 IDF 為主要參考依據決定查詢字串中每個可能單字與相連雙字的重要性，如"張德培網球排名"中"張"、"德"、"培"、"網"、"球"、"排"、"名"、"張德"、"德培"、"培網"、"網球"、"球排"、"排名"等每個單字與雙字。

2. 利用特徵檔快速檢測所有文件包含前述查詢單字與雙字出現的情形，據此計算出每個文件的初步相似度，相似度過低的文件則加以過濾。

3. 比對辭典，進一步抽取查詢中的可能詞彙，如"張德培"、"網球"、"排名"，將可能有關文件內容讀出，考慮相關文件包含這些詞彙的情形與可能的領域知識如詞彙出現在標題與內文會有不同加權，重新決定文件相似度。

這樣的近似自然語言查詢方法雖然簡單，卻也考慮中文斷詞困難、中文字意豐富、中文雙字鑑別率，中文檢索詞不易精確表達、大量文件檢索效率、領域知識的運用等因素。在實際觀察發現檢索精確率相當的高，目前我們還未作大規模檢測，但是在十多種資料庫超過上千次查詢，我們覺得中文近似自然語言檢索可能比英文有更好的效率，不過這只是臆測仍缺乏完整實驗證實。

自然語言查詢提供全新的查詢概念，只要使用者將可能的檢索概念盡量表達則查詢成效極高。如表 2 所示，形形色色的查詢透過以自然語言聯想查詢可以查出。

六. 資訊檢索 -- 進階檢索技術

1. 近似字串查詢

除了自然語言查詢，前述 Csmart 還發展許多特殊資訊檢索技術以克服中文檢索困難。在近似字串查詢方面，如表 3 所示很多時候中文如無近似字串檢索許多資訊無法加以檢索出來。 Csmart 的近似字串檢索主要是先比對字串頭尾含相同字(少數情形與同音字有關如巴塞隆納、巴塞隆那例外)，如中研院與中央研究院，資策會與資訊工業策進會，這些字串有一定長度關係且短字串內的很高比例的字出現在長字串中且字序一致，最重要的是短字除最後一個字外幾乎都出現在詞的左邊界上。

2. 能解決斷詞歧異的文件比對技術

中文檢索一般都是字串檢索，很少考慮斷詞歧異解決。然而一些查詢如表 4 如無精確斷詞，以詞的觀點則會出現 False Drops 。 Csmart 利用兩段式搜尋，在第二段文件比對時加上斷詞技術，所以多數斷詞歧異能夠排除。

查詢舉例	資料庫
李院長演講資料	中研院 Web Page 資料庫
資訊所圖書館	中研院 Web Page 資料庫
簡立峰電話	中研院 Web Page 資料庫
尋易 Csmart 系統	中研院 Web Page 資料庫
新內閣名單	網路即時新聞資料庫
張德培網球排名	網路即時新聞資料庫
芝加哥公牛對西雅圖超音速	網路即時新聞資料庫
禁止英國牛肉進口的國家	網路即時新聞資料庫
奧斯卡最佳影片	網路即時新聞資料庫
最好看的電影	網路電影論壇資料庫
形容美麗漂亮的女人	電子辭典
形容風景	電子辭典
諾貝爾獎得主	電子辭典
發明電燈的人	電子辭典
最新高溫超導體	產業技術報告資料庫
平行編譯器可行性	產業技術報告資料庫
中國文化	圖書書目資料庫
王姓電腦概論	圖書書目資料庫
世界最高的建築	建築文獻資料庫

表2　Csmart自然語言聯想查詢舉例

種類	舉例
頭銜	李登輝、李總統登輝、李主席登輝
單位簡稱	中研院、中央研究院
單位簡稱	台大、台灣大學
人名拼字	郭李建夫、郭李健夫
相似詞	中國文化、中華文化
譯名	巴塞隆納、巴塞隆那
打字錯誤	電腦概論、電腦概論
相似片語	台北市大安分局、台北市中山分局、台北市分局、
相似片語	高溫超導技術、高溫超導體技術、高溫超導材料技術

表 3　需求近似字串查詢舉例

檢索詞	須精確斷詞的字串
語言學	組合語言學、程式語言學
腦科	電腦科學
中共	其中共有、美中共同參與
陳健康	指陳健康的重要
化學	國際化學術會議、電腦化學理、動物演化學

表 4　須精確斷詞的檢索詞舉例

3. 相關文句擷取技術

網路檢索不只檢索精確率要高，檢索結果的提示須相當可讀性，以協助判斷檢索出的文件是否相關，特別是從檢索出標題不易理解與查詢的關係，或者檢索過長的全文文件時。所以檢索結果除了顯示文件標題、可能出處外，如表 5 所示相關文句擷取與顯示可以提高檢索結果的可讀性。相關文句擷取與近似字串比對觀念上相似，可是要注意查詢中關鍵詞的抽取與分佈，如"張德培"與"排名"。高水準相關文句擷取技術須有簡單文法剖析，目前 Csmart 並未運用這類技術。

查詢	相關文句擷取
張德培的網球排名	文件標題：大滿貫杯網球賽決賽 相關文句：張德培在大滿貫杯比賽擊敗貝克排名晉升第七
公牛隊第四場	文件標題：NBA 比賽公牛初嚐敗績 相關文句：NBA 總冠軍決賽第四場西雅圖超音速隊獲勝
禁止英國牛肉進口的國家	文件標題：世界主要報紙標題 相關文句：巴基斯坦今天宣佈禁止從英國進口牛肉
語音辨認	文件標題：聲控系統技術應用結案報告 相關文句：以期在語音辨認系統實現上仍能達到實驗室水準
虎象比賽最有價值球員	文件標題：職棒虎象纏戰十一局象隊二比一獲勝 相關文句：路易士同時獲選最有價值球員

表 5 關鍵文句擷取技術提高檢索結果的可讀性

4. Relevance Feedback

當使用者輸入的查詢過於簡單或者滿足查詢條件文件過多，舉例使用者原本想查詢馬英九先生新任職務，若以"馬英九"為檢索詞可能在一個月時新聞中可檢索到超過 500 則，其中多數是關於電玩弊案、反毒、工程弊案的報導。通常使用者不知道如何再過濾出相關新聞。這時透過 Relevance Feedback ，使用者只須選擇幾篇眞正相關報導，由系統重新產生更精確查詢即可。 Csmart 的 Relevance Feedback 方法是參考查詢與相關文件特徵(Signature)，利用 Signature Bit 交集與 IDF 加權重新產生查詢特徵(Query Signature)。我們發現 Relevance Feedback 特別適合網路檢索，因為網路資源量多質差，使用者不易一次決定適當的查詢，利用 Relevance Feedback 技術，不論系統與使用者負擔都可減輕。這一點從 Excite 的系統成效也可以觀察出 [2]。

七. 語音介面與語音檢索

為了有效克服中文輸入的困難，以及嘗試設計人機互動式查詢技術以模擬人與電腦對話的效果，我們以金聲 3 號為基礎完成初步具備語音檢索效果的語音介面與語音檢索，如圖 4 所示我們允許使用者以說話的方式詢問系統(Unconstrained Speech Query)。語音檢索並不是將語音辨認與資訊檢索合併即可。語音檢索要考慮語音辨認強健性特別是專有名詞，此外語音檢索更要考慮檢索強健性及容錯能力與檢索速度。在這方面 Csmart 發展許多技術[17, 18]。首先在語音辨認強健方面，我們將語言辨認的語音解碼模組(Linguistic Decoder)獨立出。以欲查詢的資料庫為語料訓練語言模型，這個語言模型是以字為基礎特別加

重專有名詞辨認率，而且具備動態訓練能力以因應資料庫的隨時異動。另外在資訊檢索的強健方面，我們允許語音辨認系統送出最可能字串與候選音節(Syllable)。

在資訊檢索子系統方面，如須以語音檢索的系統我們會把音節訊息加入文件特徵檔中，以提高檢索容錯率。在實際使用發現，由於語音輸入便利，使用者的查詢具有較多的訊息。一般可以說出 8~10 個字。只要語音辨認率維持 70%以上，對檢索系統而言即會有很高正確率。因為 Csmart 的自然語言查詢檢索成效主要是受查詢的資訊豐富與否的影響。以資訊豐富程度來看，70%的語音正確率所輸入的查詢與打字輸入查詢往往是接近的，因為打字輸入的查詢較短。我們實際觀察語音查詢的成效發現語音檢索較為便利、也比較容易表達出眞正自然的查詢、輸入也快得多，其可行性很高。這很可能是中文檢索的一大特色，對發展口語交談系統很有助益 [19]。

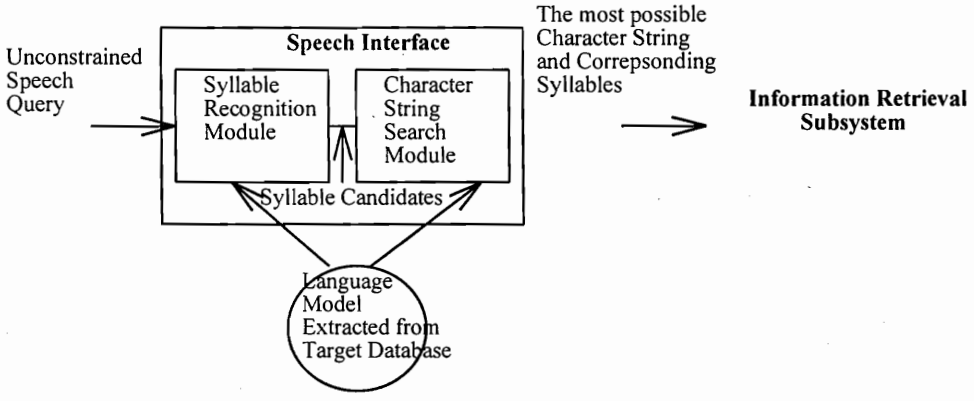

圖 9 語音介面子系統

八. 系統成效與未來研究

前述所有 Csmart 技術都已經開發完成，並且多數經過很長期測試。目前 Csmart 系統除了可以檢索包括電子辭典、建築文獻、佛學書目與摘要、產業技術報告等一般文件資料庫，也可以開始檢索網路即時新聞、BBS 論壇、中文 Web Pages 等網路資源。在搜尋速度方面，快速查詢在五千萬字(100MB)在 PC/486 環境檢索一般字串不到 1 秒。五億字(1GB)在 SPARC10 環境檢索一般字串約 2~5 秒。自然語言檢索則較花時間，在 SPARC10 檢索 2 萬則論文摘要約 1 秒，16 萬則約 4~5 秒。以台灣現有 URL 數目估計短時間不會超過 10 萬而且成長也不大因此檢索速度應無問題。而在索引成效方面，文件索引為可調式(Scalable)，索引大小視需要調整，一般文件所需之索引空間約只佔文件大小的 15~30%左右，另外索引建置時間極短，100MB 文件在 PC/486 環境約只需 4 分鐘，在 SPARC 10 只要 2 分鐘。至於檢索功能與語音檢索效率前述各節也已說明。大致上 Csmart

的資訊檢索技術已符合實用，語音介面與檢索在實驗室成效良好未來可行性高，而資源發現與擷取技術則須持續發展。為此我們已經開始研究關鍵詞抽取(Keyword Extraction)技術以因應大量網路資源所需，藉此希望發展出資訊分類與過濾技術，使我們有能力判斷出有興趣收集的資源，並且進而擷取出重要訊息加以建立索引，與發展個人化資訊服務。

參考資料

1. G. Venditto, Searching Engine Showdown, Internet World, May 1996.

2. M. Courtois, et al., Cool Tools for Searching the Web: A Performance Evaluation, Online, Nov. 1995.

3. S. Wu, Gais Home Page, http://gais.cs.ccu.edu.tw/

4. Lee-Feng Chien (95b), 尋易 (Csmart) -- A High-Performance Chinese Document Retrieval System, The 1995 International Conference of Computer Processing for Oriental Languages, *ICCPOL'95*.

5. Lee-Feng Chien , Hsiao-Tiech Pu, Ming-Chan Chen, Hung-Ming Chen and Ming-Jer Lee, Natural Language Information Retrieval with Speech Recognition Techniques for Network Chinese Resources Discovery, May 1996 International Workshop on Information Retrieval with Oriental Languages

6. Lee-Feng Chien , A Model-Based Signature File Approach for Full-text Retrieval of Chinese Document Databases. To appear on Computer Processing of Chinese and Oriental Languages, 1996.

7. Lee-Feng Chien , Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts, ACM SIGIR'95, 1995.

8. Sung-Chien lin, Lee-Feng Chien, Keh-Jiann Chen, Lin-Shan Lee, An Efficient Voice Retrieval System for Very-Large-Vocabulary Chinese Textual Databases with a Clustered Language Model, May 1996 (ICASSP'96).

9. Belkin, Nicholas J., et al, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?", Communications of the ACM, Vol. 35, No 12, Dec. 1992).

10. D. Lewis, Evaluating and Optimizing Autonumous Text Classification Systems, SIGIR'95.

11. Foltz, Peter W., et al, "Personalized Information Delivery: An Analysis of

Information Filtering Methods", Communication of the ACM, Vol. 35, No. 12, Dec., 1992.

12. Lee-Feng Chien and Hsiao-Tiech Pu, Important Issues on Chinese Full-text Information Retrieval, Invited and to be submmited for Computational Linguisitcs and Chinese Language Processing.

13. Faloutsos, C., "Access Methods for Text", ACM Computing Surveys, March 1985, 49-74.

14. Tyne Liang, Suh-yin Lee and Wei-Pang Yang, Optimal Weight Assignment for a Chinese Signature File, Information Processing and Management, Vol 32, No. 2, pp. 227-237, 1996.

15. Lee, Ahn and Shin, An Effective Indexing Method for Korean Text Retrieval, International Workshop on Information Retrieval with Oriental Languages, Korea, 1996.

16. Y. Ogawa, A New Character-based Indexing Organization Using Frequency Data for Japanese Documents, SIGIR'95.

17. Sung-Chien Lin, Lee-Feng Chien and Lin-shan Lee, A Syllable-based very-Large-Vocabulary Voice Retrieval System for Chinese Databases with Textual Attributes, Proceedings of the 4th European Conference on Speech Communication and Technology, Sept. 1995.

18. Sung-chien Lin, Lee-Feng Chien and Lin-shan Lee, Unconstrained Speech Retrieval for Chinese Docment Databases with Very Large Vocabulary, Proceedings of the 4th European Conference on Speech Communication and Technology, Sept. 1995.

19. Yen-Ju Yang, Lee-Feng Chien and Lin-Shan Lee, An Efficient linguistic Decoding System with Adaptive Learning for Mandarin Speech Recogntion, accepted by CPCOL.