網路論壇之知識搜尋

Knowledge Searching over BBS

楊錦潭[1,*] 陳玟志[2] 廖慶榮[3] 簡世宇[4]

[1]南台科技大學
數位設計學院
71005 台南縣永康市南台街一號
yangdav@mail.stut.edu.tw

[2]國立高雄師範大學
資訊教育研究所
80243 高雄市和平一路 116 號
koach@ccsh.kh.edu.tw

[3]中原大學
資訊管理學系
32023 桃園縣中壢市中北路二００號
cjliao@mail.cycu.edu.tw

[4]南台科技大學
電子工程系博士班
71005 台南縣永康市南台街一號
stn3388@yahoo.com.tw

Jin Tan David Yang[1,*] Wen Chi Chen[2] Ching-Jung Liao[3] Shih-Yu Chien[4]

[1]Graduate Institute of Information Communication, Southern Taiwan University of Technology
Yungkung, Tainan County, 710, Taiwan
yangdav@mail.stut.edu.tw

[2]Graduate Institute of Information & Education, National Kaohsiung Normal University
Yungkung, Tainan County, 710, Taiwan
koach@ccsh.kh.edu.tw

[3]Department of Management Information Systems, Chung Yuan Christian University
200, Chung Pei Rd., Chung Li, 320 Taiwan
cjliao@cycu.edu.tw

[4]Graduate Institute of Electronics Engineering, Southern Taiwan University of Technology
Yungkung, Tainan County, 710, Taiwan
stn3388@yahoo.com.tw

Received 17 June 2006; Revised 29 July 2006 ; Accepted 18 September 2006

* 通訊作者

摘　要

本研究旨在透過「社群行爲特性」(traits of community behavior)的考量，改善目前「非同步網路論壇」(Bulletin Board System, BBS)的搜尋效益，以增進學習者對研究資源的複用性(reusability)程度。社群行爲特性包括三項，即：時間趨勢、相關的活躍會員、與參考資源，再提供使用者從這三個特性進一步的資料過濾。網路論壇提供了個人將其隱性知識表現出來的機會，進而成爲可分享的顯性知識，達到知識共享的功能。這些被記錄下來的討論過程一旦數量龐大，若要具有效的被利用性(reusability)，則必須要借助特定情境而設計的搜尋機制才能辦到。本研究中的演算法爲「最長共同子序列」(Longest Common Subsequence, LCS)與最長共同連續子序列(Longest Common Consecutive Subsequence，LCCS)。本研究的個案爲高師大的「瑞士團隊」(RAS, Research Assistant System)論壇爲例，RAS主要是用來增進研究生學習社群的互動性，以提高成員對研究資源的掌握度。本研究的成果有二：一爲在內建式的搜尋結果之上，加入排序權重演算法，得以改善傳統上依照資料庫欄位設計的排序方式；二爲多元化的搜尋結果呈現。未來將對此新的功能將進一步的作實徵性的研究。最後，對未來研究的建議亦一併討論。

關鍵詞：社群行爲特性、非同步論壇、最長共同子序列、最長共同連續子序列搜尋

ABSTRACT

The purpose aims to adopt traits of community behavior for searching effectiveness in a Bulletin Board System (BBS). The improved searching algorithm is used to promote reusability on the existing research resource of BBS. There are three traits of community behavior, time trend, relative active members, and referential resources. These three parameters can be adjusted as knowledge filtering. Therefore, they can attain what learners intend to get easily. BBS provides a mediated communication in asynchronous practice while members in the BBS community transfer implicit knowledge to explicit knowledge. This explicit knowledge can be also shared among all members. As time goes by, more and more knowledge in BBS will be accumulated as pool of knowledge. The key algorithms adopted in this study are LCS (Longest Common Subsequence, LCS) and LCCS (Longest Common Consecutive Subsequence). The subject of this study is RAS was established for research project at National Kaohsiung Normal University. RAS is originally designed for promoting reusability of existing research resource among RAS's members. The results of this study consist of (a). A weighted ranking algorithm by LCS and LCCS is developed to improve ranking by traditional database approach. (b). A multiple presentation on searching results is designed for knowledge retrieval with easier interfaces. The next experiment study for verify the effectiveness of this improved search will be conducted soon. Finally, the implications for future research are also discussed.

Keywords: traits of community behavior, asynchronous BBS, LCS, LCCS, searching

一、緒論

「網路論壇」(Bulletin Board System, BBS)具有訊息交換、線上交談、問題解答、經驗交流等多項功能。將BBS用於在教學之中，它提供了學生/教師個人將其隱性知識表現出來的機會，進而成爲可分享的顯性知識，達到群體知識分享的功能。因而，全世界日益增加的教師，利用它來進行非同步教學。顯示出它已成爲師生社群非同步互動上最爲普遍性的方式。眾所同知，「網路論壇」的後端平台係採用「關聯式資料庫」(relational database)的規劃方式在存放資料，因而在做文章檢索時，可以很輕易地根據欄位做爲查詢條件的依據，查詢的結果也可直接以欄位做爲「排序」(rank)輸出的依據，然而這種排序方式雖然在速度上可以很有效率，但是在意義上卻很難達到人性化的考量。

目前的「網路論壇」查詢結果的排序方式，僅是根據系統資料庫的欄位做排序（如日期、主題等），而且無法更具智慧化的將文章排序(例如：類似同位語的自動轉換機制、或依作者回應量等重要因素考量於其中)，則勢必造成當搜尋結果的資料數很多時，使用者同樣需求花費大量時間精力去逐一過濾，無法快速找到有價值的相關文章。

因而，本研究提出改善目前網路論壇的搜尋效益，以增進成員對研究資源的複用性程度。成員可以透過簡易的操作界面而完成資訊擷取的參數調整，以快速取得想要的資料。要讓「討論版」的內容能夠有效的被利用，則應該需要有良好的搜尋介面、搜尋機制、與資料呈現的介面，才能讓使用者輕易找到他們所需的內容。例如：讓他們能以「自然語言」式搜尋、運用社群的「情境屬性特性」搜尋等爲考量來改善搜尋結果的呈現。

本研究在網路論壇內容的搜尋上，除了單純的依照資料庫欄位去排序外，應該要深入了解在論壇內情境屬性特性，提供符合論壇特性的排序，改良搜尋結果的呈現方式，以期降低使用者在搜尋資料時的困擾，並減少使用者找到符合資料的時間。爲了印証本研究的成效，我們實作改良式搜尋功能於「研究助理系統(Research Assistant System, RAS; http://ia.nknu.edu.tw/ras)，上以提高該會員使用。

二、文獻探討

基於本研究的目的，本節的文獻探討並分爲二個：即最長共同子序列(Longest Common Subsequence, LCS)與最長共同連續子序列(Longest Common Consecutive Subsequence , LCCS)的演算法技術。前者是用來比對兩串序列內容的共同部份，在不同的研究領域上，常利用找出共同子序列的部份來代表不同序列間的相似度；後者則是加入連續的子序列。

1.最長共同子序列(Longest Common Subsequence, LCS)

最長共同子序列(Longest Common Subsequence, LCS)演算法的目的在找出兩個以上

的序列之間所包含的共同序列部份，因此常被用來代表兩個序列的相似度，或是間接取得之間的差異部份。該研究最早是由[1]所提出。

因爲 LCS 可達到機器智慧中模糊比對的特性，找出不同序列之間的共同子序列，這種特性可運用於廣泛研究上。常見的應用有例如文字辨識[2] [3]、聲音辨識[4]、資料分類[5][6]、生物基因[7]等研究。

在「網路教學」的實徵性的研究中，序列的資料流代表網路學習者使用的路徑，因此也可將 LCS 演算法運用在兩個人以上的網路使用路徑上；用 LCS 來分析學生於網路上的學習路徑，分析不同學習者的學習路徑的相似度，以進一步可以自動分析學生可能所屬的群組，達到所謂因材施教的可能[6]；另外，透過將使用者於網站中所經過的網頁路徑記錄下來，再利用 LCS 演算法引入權重(Weight)的概念，再依照網頁的意義予以轉化成概念化的網頁路徑，以解決實體網頁路徑相似度過小的問題，最後利用 LCS 演算法來判斷網頁路徑的相似度，將相似的網頁路徑做自動分群[5]。

LCS 演算法處理常需耗較大的運算時間和空間，因而較適合應用在比對的目標長度較短的情況，它常被用在關鍵字或主題文字的搜尋。然而，採用 LCS 模糊比對的特性，實作可容錯的文字比對功能，讓使用者的查詢字串即使未完全符合資料庫系統的部份字串時，亦可有機會抓出相關的資料[2]。例如:查詢字串若爲”聯華電子”，傳統的資料庫比對方式將無法取出內容爲”聯電”的相關資料，此時利用 LCS 的特性即可抓出。然而這種方式較適合應用在比對的目標長度較短的情況。一旦 LCS 演算法運用在資料庫內容的全文比對上，則使用者必須等待更長的時間。尤其甚者，若同時有多人使用時，LCS 演算法對電腦主機的資源也將是一大挑戰。

後來，LCS 演算法的改良之後的演算法被稱爲最長最多共同片段(Longest Most Common Segments, LMCS)演算法[8]。LMCS 應用於搜尋的目的在於協助使用者找到真正想要的產品。因而，主要是利用產品的名稱、序號來辨識產品是否相同。其研究中指出，因爲大多數同類型的產品，其名稱和序號常會有大量的相同文字序列(LCS 值較高)。若單純以 LCS 值來判斷產品的異同，則可能會產生產品的誤判，因爲在研究中發現有些產品有相同的 LCS 值，然而產品卻是相異的情形。因此作者提出以改良的 LMCS 來做爲產品比對的演算法，找出序列之間的共同片段，用以解決當序列有較長的相同序列而局部卻有微量相異的比對。

2.最長共同連續子序列(Longest Common Consecutive Subsequence , LCCS)

最長共同連續子序列(LCCS)，和上述最長共同子序列(亦簡稱爲 LCS)的差異在於 LCCS 強調子序列必須是連續的。這種連續的特性可用來降低 LCS 可能因爲允許文字間不連續而產生誤判的情形，例如：” cream cheese ” 和” CRMS” 的 LCS 是” crms” 四個字，但事實上兩者是完全不相關的文字；而若是以 LCCS 判斷，則會發現只有” cr” 兩個字而已，在比例上只有 LCS 的 1/2，較不會誤判可能是代表相同的意思。

在「網路教學」的實徵性的研究中，將 LCS 和 LCCS 結合，運用在音樂資料的判斷研究中[4]。先將使用者輸入的音訊轉爲中介資料，再與資料庫裡的音訊資料比對，用來判斷不同的聲音訊號資料，是否其實是同一段音樂。

例如使用者所讀入的聲音序列爲 α，要比對的資料庫聲音序列爲 β。

$$\alpha = [1，1，2，0，-1，0，1，2，0] \tag{1}$$

$$\beta = [-3，1，1，2，4，-1，1，2，5] \tag{2}$$

則 LCS(α ，β)爲[1，1，2，-1，1，2]，長度爲 6。而 LCCS(α ，β)爲[1，1，2]，長度爲 3。最後則是利用 LCS 和 LCCS 各自乘以不同權重值(λ ， μ)來取得兩串音樂序列的相似值。

$$sim(\alpha,\beta) = \lambda \times \frac{Lcs(\alpha,\beta)}{N} + \mu \times \frac{Lccs(\alpha,\beta)}{N} \tag{3}$$

LCCS 對於同根詞的比對有相當的幫助[8]。例如{absent}和{absence}的兩個字雖然相異，但 LCCS 所得到的值{absen}很高，即可判斷爲可能是相關的字。在文章中主要是利用 LCCSR(LCCS ratio)來計算相似度。

$$LCCSR(w_1, w_2) = \frac{|LCCS|}{\max\{|w_1|, |w_2|\}} \tag{4}$$

並且定義下列條件來計算字面的相似度 LS(Literal Similarity)

$$LS = \begin{cases} LCCSR & LCCSR \geq 0.5 \\ 0 & LCCSR < 0.5 \end{cases} \tag{5}$$

由以上的文獻探討，在研究中發現使用 LCCS 技術結合 RAS 情境下的重要屬性搜尋策略，即可在不對系統的運算負苛造成延遲的情況下，對文章主題關聯性的判斷有不錯的效果。在文章中，關鍵字出現的位置應該列爲重要的搜尋考量依據[9]。因此本研究中，即是要針對某些關鍵欄位做處理。例如：利用文章主題、附件檔案、描述、版面標題，結合使用 LCCS 演算法來計算使用者關鍵字和論壇中各主題的關聯性，以便對不同的主題給予不同的排序依據。

三、研究方法

在本研究中所設計的改良式搜尋功能，是採用在既有的 PHPBB 搜尋功能上，以模組化嵌入的方式而設計的，目的是希望將來 PHPBB 系統升級後，本研究的成果能以同樣的方式融合到新的 PHPBB 系統中，以降低因系統升級而帶來的維護成本。

本研究的系統架構圖如圖一。系統架構內的改良式搜尋功能，包括了五個模組，1.權重排序處理；2.時間趨勢分析；3.推薦活躍會員；4.主題特徵擷取；5.搜尋結果呈現。依序五個模組的功能描述如下。

壹、系統模組介紹

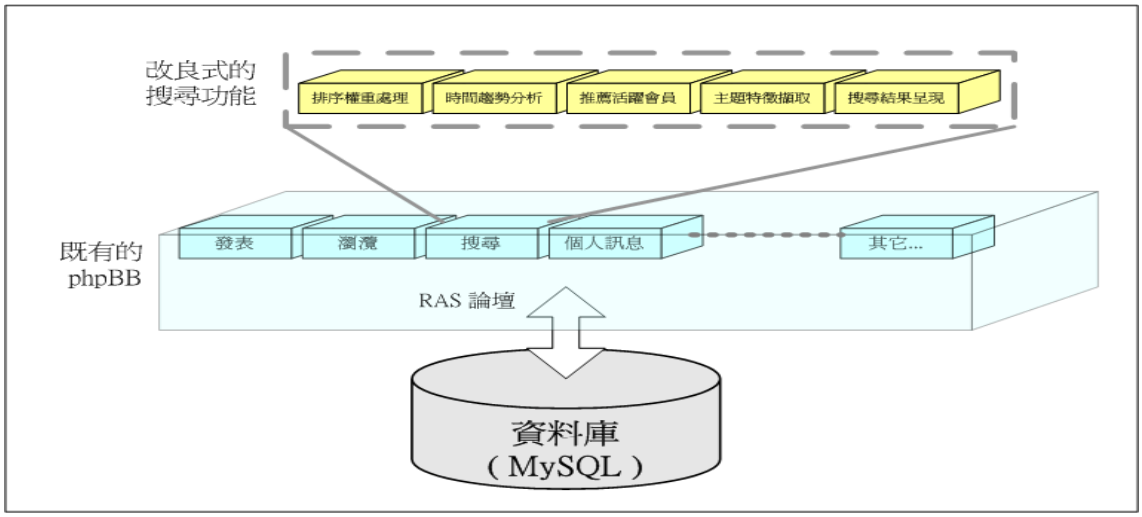
如系統架構圖(圖一)所示，本研究的實作包含了五個模組，其主要功能及設計方法說明如下：

◎權重排序處理模組

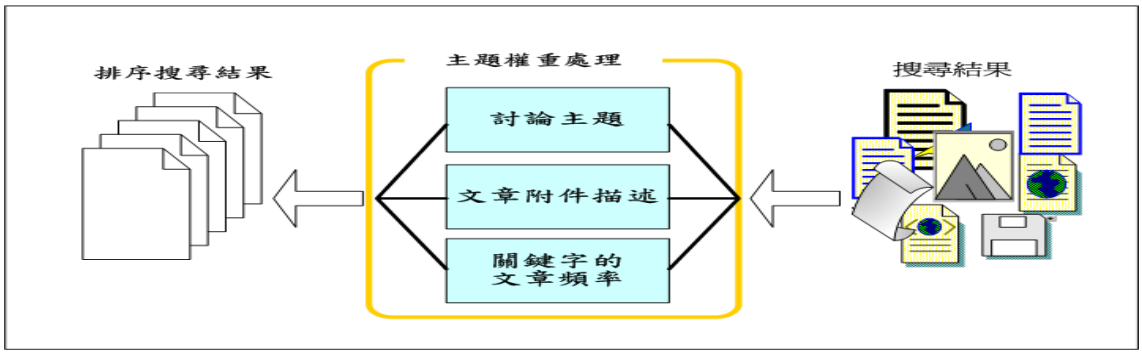
「權重排序模組」(weighted sorting module)主要是根據 RAS「社群行爲特性」的情境考量而設計的。除了依據使用者查詢字串所出現的欄位而考量外，同時結合 LCS 和 LCCS 的演算法以增加相似文字的接受度，最後計算出每個討論主題的排序權重。排序權重的設計主要考量三個屬性(圖二)：1.討論主題；2.文章附件描述；3.關鍵字的文章頻率。最後根據這三個屬性所得的值，給予各主題不同的計分排序。

◎時間趨勢分析模組

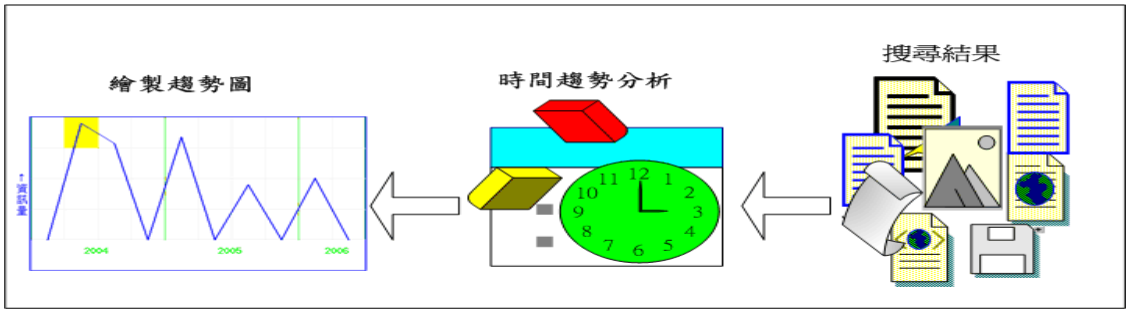
在實際搜尋的過程中，使用者通常很難明確指出欲尋找的資料時間，因此設計此一「時間趨勢分析模組」的目的，主要是希望能讓查詢者在輸入查詢關鍵字後，能快速得知該關鍵字在 RAS 論壇內的討論趨勢及時間上的統計資訊。此模組的功能是將搜尋結果的文章，以時間的觀點做統計分析，並且加以圖形化處理後，在搜尋結果的畫面中，



圖一：本研究之系統架構圖



圖二：權重排序模組

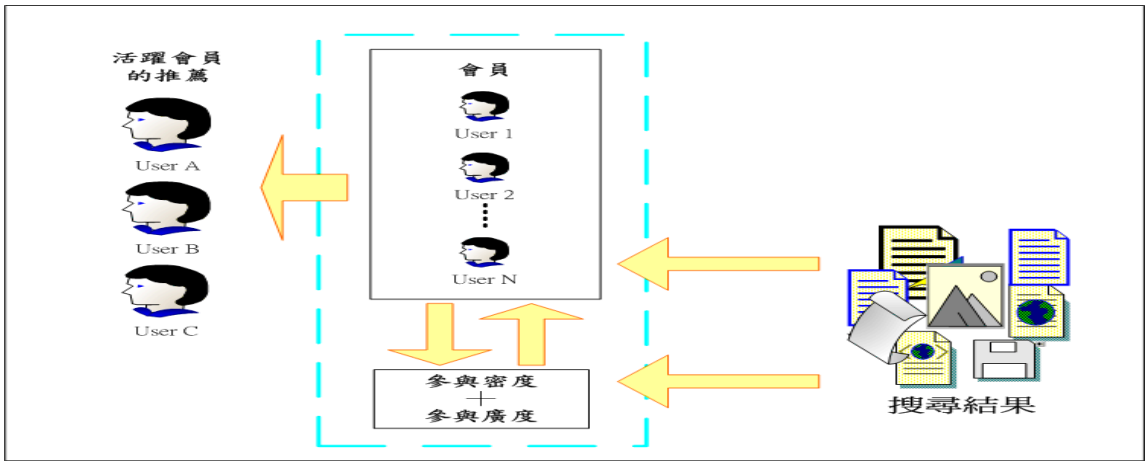


圖三：時間趨勢分析模組

呈現給查詢者觀看(圖三)。

◎推薦活躍會員模組

RAS 系統是屬於封閉式的社群，然而並非每個成員彼此都了解各自的專長的領域和經驗，尤其隨著研究生的迭替，有時並無法輕易了解其他 RAS 會員所作的研究。因此在本研究中設計的改良式搜尋機制，也加入了在搜尋結果中提供活躍會員作者的考量。讓成員在瀏覽查詢結果之際，亦可同時得知系統所推薦的活躍會員，讓搜尋者可以很快的掌握可供諮詢或討論的對象。此模組主要是根據搜尋結果的統計，將每位列於其中的會員，根據其於搜尋結果中的統計資料，分列運算出每位會員的「參與密度」(density of participating， DP)與「參與廣度」(width of participating， WP)。並根據這兩項分數相加以代表使用者對於該查詢的相關度。最後將前三名的會員於搜尋結果中推薦給查詢者(圖四)。



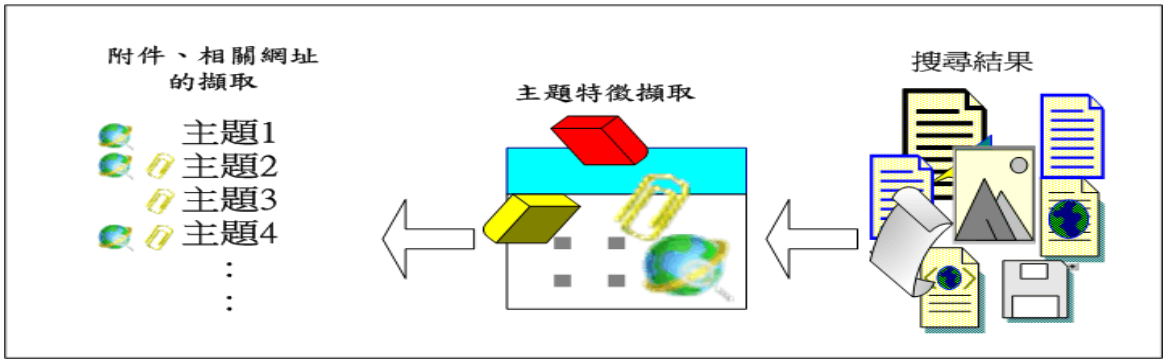
圖四：活躍會員的推薦

◎主題特徵擷取模組

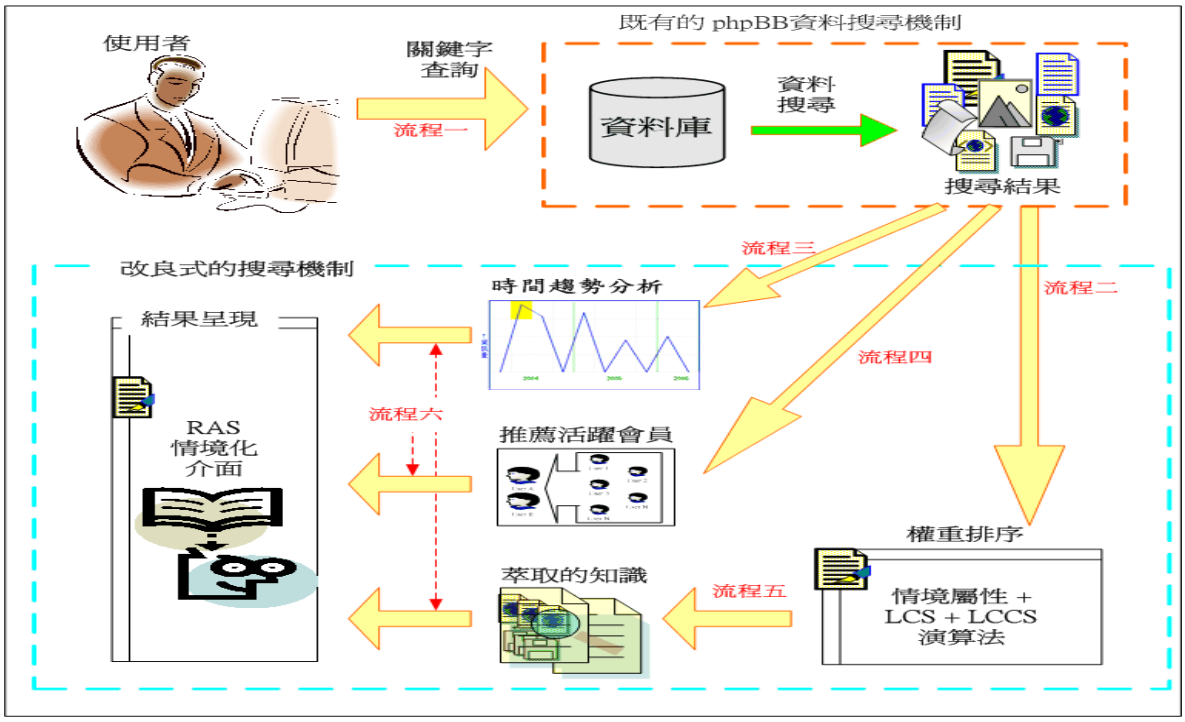
「主題特徵擷取模組」(retrieval of topic trait)是用來擷取搜尋結果的「附件」(attached document)與「網址」(URL)屬性特徵。在 RAS 常中，有許多討論文章是資源的分享或是研究的心得。當會員在 RAS 系統上發表文章時，往往因爲網頁上輸入功能的限制，或網頁內容本身呈現的因難，因此會員常會於發表的文章中，另外提供了相關資訊的超連結網址(hypertext)或是上傳成附件(如 doc, ppt, pdf 等文件)以做爲參考資源。因而在此模組中(如圖五)，會針對搜尋結果的各個主題，將討論內容中所包括的附件和文章中所提到的網址擷取出來，以做爲後續結果呈現的來源之一。

◎搜尋結果呈現模組

搜尋結果呈現的模組，是把前面四個模組所得到的輸出，根據 RAS 上的需求，設計出適合的呈現介面。依權重排序各主題、顯示各主題所含的相關網址和附件檔案描述，並且提供直接於搜尋結果的畫面中直接做超連結或檔案下載。同時也將搜尋結果依時間上的統計繪製趨勢圖，並且讓使用者可以透過趨勢圖上的分析，對搜尋結果做時間範圍的篩選，再重新依權重排序。



圖五：主題特擷取模組



圖六：改良式搜尋功能的系統流程

最後，透過活躍會員的推薦，查詢者則可以直接在結果的畫面上，透過電子郵件向推薦會員請教，或是讓搜尋者直接針對該會員做搜尋結果的主題做篩選，屬於該會員的主題排在前面，然後再重新依原先主題的權重做排序，以加快針對某重要會員的文章篩選。

貳、系統流程運作

逐步說明本研究所改良的搜尋功能，其內部各個模組的運作流程，如圖六所示。

流程一：

使用者輸入關鍵字，從系統的資料庫中，以全文比對的方式擷取出符合條件的文章。使用者亦可同時使用多個關鍵字，或是用 And、Or、Not 的基本邏輯運算來限制多個關鍵字。此部份是依循原先 PHPBB 的搜尋方式，並未改變其全文比對的搜尋演算法。因此，對於相同的查詢，內建式和改良式的搜尋結果，其所得查詢結果的涵蓋率(Recall)值是相同的，亦即查詢結果所傳回的總筆數相同。

流程二：

經由流程一所得的搜尋結果和原資料庫系統中，取出各篇主題的屬性做爲分析各主題權重的運算依據，把這些屬性交給權重排序模組來處理，算出對應的積分，以做爲排序的依據。

流程三：

流程一所得的搜尋結果資料，同時會送交給時間趨勢模組處理，算出文章在時間維

度上的分佈趨勢。透過日期時間的轉換和數量的統計分析，繪製成圖形化的趨勢圖。

流程四：

搜尋結果的資料，同時會經由會員推薦模組處理，以決定在該次搜尋，所要推薦的活躍會員是哪些。

流程五：

搜尋結果經由主題特徵擷取模組的處理，取出每個討論主題的內容特徵，也括附件和討論中所提及的相關網址。

流程六：

將以上各流程處理後的資料結果，交給搜尋結果呈現模組處理。先依主題的權重積分排序、顯示出文章的時間趨勢、每個主題的內容特徵以及被推薦的活躍會員，以產生計適合 RAS 社群行爲特性的情境化呈現介面。

四、研究結果

本研究提出改善「非同步網路論壇」內的搜尋機制有二：搜尋結果的排序方式與搜尋結果的呈現介面。如此一來，即可讓成員依其意圖調整參數而快速取得意義層次的內容。

1.搜尋結果的排序方式

表一：計算排序權重時所使用變數及函數名稱與定義

變數及函數名稱	變數及函數定義
QStr	使用者的查詢關鍵字(Query String)
Ws,i ：	主題 i 的討論主題權重
Wa,i	主題 i 的文章附件描述權重
Wf,i	主題 i 的文章頻率權重
SSi	主題 i 的標題
ASi	主題 i 的文章附件描述
Ci	主題 i 的文章總數
Ci(Str)	主題 i 包含字串 Str 的文章總數
LCS(S1,S2)	最長共同子序列
LCCS(S1,S2)	最長共同連續子序列
LCCSR(S1,S2)	length(LCCS(S1，S2)) / length(S1) 即 Str 和 QStr 的最長相同字串比例

爲了讓使用者在 RAS 論壇上搜尋資料時，不會因傳回過多的資料而浪費時間逐一過濾，因此必要設計一個在 RAS 網路論壇上適用的搜尋結果排序方法。先了解 RAS 網路論壇的運作特色，同時分析各項文章的「屬性特質」(attribute property) ，設計適合的權重計分方法以做爲排序的依據，將相對於搜尋關鍵字較重要的資料排在前面的部份，以減少使用者過濾搜尋結果的時間。

本研究中採用三個權重設計，分別是 1.討論主題(Subject); 2.文章附件描述(Attachment); 3.含關鍵字的文章頻率(Frequency of posts)。此外，定義在計算排序權重時所使用的變數及函數如表一。

接著針對三個權重設計的演算法進行介紹。

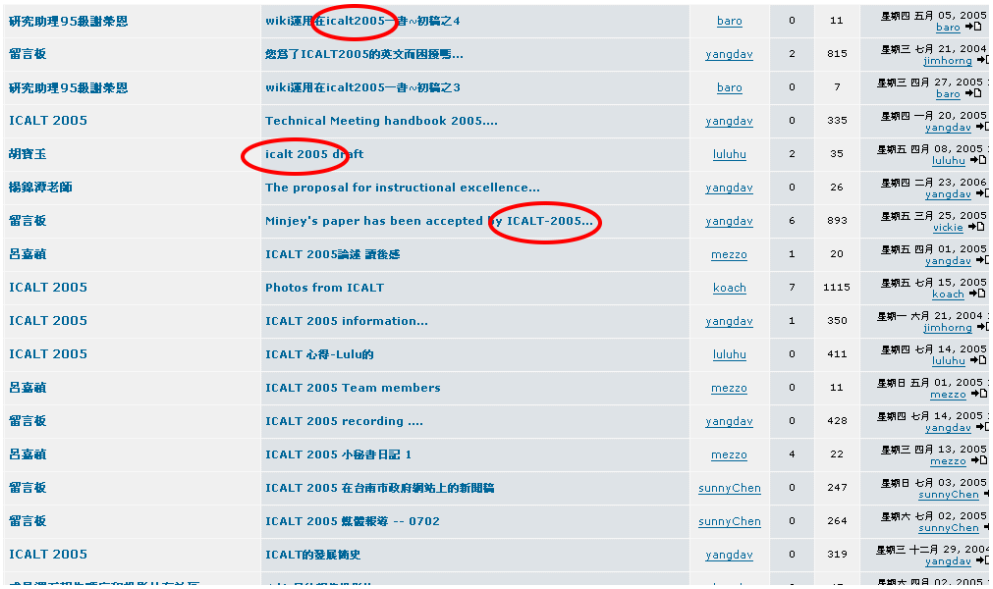
A.討論主題的權重(Ws,i)

將討論主題的標題文字(SSi)取出，若主題已包含查詢關鍵字(QStr)，則計爲 1 分。如果沒有出現，則判斷 QStr 和 SSi 的最長共同連續子序列比例 LCCSR(QStr,SSi)，若其值超過 0.5，則再進行 LCS 的處理，以增加相似文字的容許程度(劉洋，民 94)。最後以 LCS 長度和 LCCS 長度除以(查詢關鍵字長度*2)所得之值做爲主題部份的權重，如下公式(6)。

$$W_{s,i} = \begin{cases} 1 & LCCSR(SS_i) = 1 \\ \dfrac{length(LCS(QStr, SS_i)) + length(LCCS(QStr, SS_i))}{(2 * length(QStr))} & LCCSR(SS_i) \geqq 0.5 \\ 0 & LCCSR(SS_i) < 0.5 \end{cases} \quad (6)$$

結合 LCS 和 LCCS 的演算法的用處，以輸入”ICALT2005”的查詢關鍵字時爲例。有時在 RAS 上的討論文章，會因爲每個人的習慣不同，而採用不同的描述方法，如圖七所示，有人以”ICALT 2005”或爲”ICALT-2005”來表示。

此時如果直接以”ICALT2005”做比對，則會將這兩種不同的描述方法判斷爲完全不同而得到相似度爲 0 的權重。然而事實上，不管是”ICATL2005”、”ICALT 2005”或”ICALT-2005”指的都是同一個名稱，因此在本研究中，爲了解決此種情況，採取結合 LCCS 和 LCS 同時處理的方法。”ICALT-2005”的 LCCSR，會得到 5，超過原查詢關鍵字 ICALT2005 長度的 0.5(9*0.5=4.5)，因此再進行 LCS 的處理，以增加對相似字的容許能力，LCS 值得到 9，則最後該篇主題的主題權重可得到(5+9)/(2*9) ≒ 0.8，而不會因爲直接比對而產生 0 的權重。



研究助理95級謝榮恩	wiki運用在icalt2005一書~初稿之4	baro	0	11	星期四 五月 05, 2005 baro
留言板	您寫了ICALT2005的英文有困擾嗎...	yangdav	2	815	星期三 七月 21, 2004 jimhorng
研究助理95級謝榮恩	wiki運用在icalt2005一書~初稿之3	baro	0	7	星期三 四月 27, 2005 baro
ICALT 2005	Technical Meeting handbook 2005....	yangdav	0	335	星期四 一月 20, 2005 yangdav
胡寶玉	icalt 2005 draft	luluhu	2	35	星期五 四月 08, 2005 luluhu
楊錦潭老師	The proposal for instructional excellence...	yangdav	0	26	星期四 二月 23, 2006 yangdav
留言板	Minjey's paper has been accepted by ICALT-2005...	yangdav	6	893	星期五 三月 25, 2005 vickie
呂嘉穎	ICALT 2005論述 讀後感	mezzo	1	20	星期五 四月 01, 2005 yangdav
ICALT 2005	Photos from ICALT	koach	7	1115	星期五 七月 15, 2005 koach
ICALT 2005	ICALT 2005 information...	yangdav	1	350	星期一 六月 21, 2004 jimhorng
ICALT 2005	ICALT 心得-Lulu的	luluhu	0	411	星期四 七月 14, 2005 luluhu
呂嘉穎	ICALT 2005 Team members	mezzo	0	11	星期日 五月 01, 2005 mezzo
留言板	ICALT 2005 recording	yangdav	0	428	星期四 七月 14, 2005 yangdav
呂嘉穎	ICALT 2005 小秘書日記 1	mezzo	4	22	星期三 四月 13, 2005 mezzo
留言板	ICALT 2005 在台南市政府網站上的新聞稿	sunnyChen	0	247	星期日 七月 03, 2005 sunnyChen
留言板	ICALT 2005 媒體報導 -- 0702	sunnyChen	0	264	星期六 七月 02, 2005 sunnyChen
ICALT 2005	ICALT的發展簡史	yangdav	0	319	星期三 十二月 29, 2004 yangdav

圖七：查詢關鍵字”ICALT2005”得到的部份結果畫面

B.文章附件描述的權重(Wa,i)

在 RAS 系統中的討論內容中有相當高的比例含有附件，其中約有 50%的主題，包含至少 1 個以上的附件檔案。因此對於搜尋結果的排序方式，本研究亦將附件列爲權重考量的因素之一。

對於討論主題中的附件，主要可以用來處理的欄位是檔案名稱和摘要描述。雖然由資料庫記錄顯示出，上傳檔案時僅有很少的比例會真正去填寫摘要性的敘述(comment)，然而本研究仍同時將這一部份的內容取出，所以對於附件描述的部份，是包括了附件檔名和摘要敘述兩部份。

處理的方法是先進行檔名和摘要敘述文字的合併，得到一連續的附件描述文字，再將此描述文字和查詢關鍵字串進行 LCCS 和 LCS 的比對方式，方法如同上述討論主題的權重處理過程，最後將計算所得到的值做爲附件部份的權重，如公式(7)所示。

$$W_{a,i} = \begin{cases} 1 & LCCSR(AS_i) = 1 \\ \dfrac{length(LCS(QStr, AS_i)) + length(LCCS(QStr, AS_i))}{(2 * length(QStr))} & LCCSR(AS_i) \geqq 0.5 \\ 0 & LCCSR(AS_i) < 0.5 \end{cases} \quad (7)$$

圖八：查詢關鍵字"ontology"所得結果的其中一個主題內容

例如當查詢關鍵字爲"ontology"時，所傳回的主題內容中（圖八），該主題中包含了一個附件，檔案名稱爲"ontologies in personalized Mobile"。在此例中，文章附件描述的權重將會以"ontologies in personalized Mobile"和"ontology"做比對，最後依公式(7)計算，得到(7+7)/(2*8) = 0.875 的權重。

C.含關鍵字文章頻率的權重(Wf,i)

在網際網路搜尋引擎所使用到的技術之一，是採用查詢關鍵字在文章中出現的頻率[10]。然而在 RAS 系統上，常會出現研究生將程式碼或是程式執行錯誤的傳回訊息張貼於系統上，以和其他研究生做交流或尋求協助。而這些文章，通常會出現大量重覆性的文字，如圖九所示，而這些重覆性的文字有時並不代表即爲該文章的重要關鍵字。因此若單純以關鍵字出現的頻率來計算，則這一類的文章將會產生較大的權重值。

在本研究中使用的方法是根據關鍵字頻率的概念而修改設計的，我們並不直接採用關鍵字在文章的出現次數，而是採用包含關鍵字的文章次數。因爲在網路論壇中的文章特性，是以討論主題爲單位，每個討論主題是可以由多篇的討論文章所組成。

因此本研究採用在同一個討論主題中，出現關鍵字的文章數做爲頻率的計算依據，亦即如果討論主題中，其中某一篇回覆文章不管出現了是 10 次或 1 次的關鍵字，則均代表 1 篇文章爲包含關鍵字。但如果其中有 3 篇回覆文章均包含關鍵字，則其出現關鍵字的文章頻率即爲 3，其權重會比只有 1 篇包含關鍵字的主題來得重。這樣設計的好處，是強調在網路論壇中，使用者的回覆通常會圍繞在主題的重點上，或是把主題文章的重要部份做引用(quote)，因此這些關鍵字將會重覆出現在同主題中的不同的文章內。

圖九：張貼程式錯誤訊息的文章

基於以上的考量，本研究於是採用文章數的設計來處理關鍵字的頻率權重。然而若單純以符合的文章數做爲權重的考量，那麼愈多人的回覆的文章，則愈可能得到較高的分數，而且和較少人回覆的文章相比，可能就會造成倍數的差距。因此爲了避免文章的多寡直接影響權重的差異，於是採用文章比例的設計方式。

文章比例的設計方式，是先算出討論主題內包含查詢關鍵字的文章數，再除以該討論主題的總文章數加 1。在分母部份採用文章數加 1 的原因，是爲了避免當主題僅包含第一篇發表文章而無其它回覆文章時，會因此產生分子和分母均爲 1，而造成該主題即爲滿分(1 分) 的情形。因此將分母的部份加 1，以抑制回覆文章數較少反而可能造成權重較高的情形。關鍵字文章頻率權重的演算法如公式(8)所示。

$$W_{f,i} = \frac{Ci(QStr)}{Ci+1} \tag{8}$$

圖十：搜尋結果依權重等級排序

例如在查詢”rss”時，某一討論主題內共有 8 篇文章，其中有 7 篇均包含有”rss”字串，那麼關鍵字文章頻率的權重(Wf)計算，就會得到 7/(8+1)≒0.78 的分數。

D.搜尋結果的排序

經由以上三項權重數據：1.討論主題的權重(Ws,i)；2.文章附件描述的權重(Wa,i) ；3.關鍵字文章頻率（Wf,i），本研究最後是將三項權重的總和訂爲查詢關鍵字和各主題的相關程度，並依之將對應的主題由高至低排序。圖十所示，圖中顯示”等級”的部份是已經把各主題的得分量化爲 1~5 級分的結果。

2.搜尋結果的呈現介面

在本研究所設計的攻良式搜尋機制中，也將在搜尋結果的呈現介面上，提供使用者做進一步資料過濾的功能。亦即當使用者輸入關鍵字後，縱使搜尋結果無法將使用者想到的資料排序在前面，使用者也可以透過改良式的介面，對搜尋結果做進一步的過濾，將搜尋的結果逐步逼進使用者的需求。

對於搜尋結果進一步的過濾功能，本研究則是基於社群行爲的特性爲考量，設計出適合 RAS 運作的過濾功能，提供關鍵字之外的過濾條件。在本研究中採用了三個可以讓使用者在搜尋結果中，再次以不同方向的條件來過濾，分別是 1.事件時間；2.參與會員；3.參考資源。

A.事件時間的引導

事件時間的引導設計，是希望能透過圖形化的統計介面，讓搜尋者對於搜尋的結果，由整體的觀點來了解搜尋結果在時間的分佈狀況。

圖十一：搜尋結果的時間趨勢圖

在本研究中，先將使用者查詢後結果，以每一季的文章數量統計，製作出圖形化的趨勢分佈圖，讓使用者了解該關鍵字於 RAS 過去的記錄中，在時間上的分佈意義。並且提供使用者能夠在圖形直接點選時間區塊，以過濾出該段時間的討論主題。圖十一所示，使用者點選了趨勢圖中 2004 年第 2 季的區塊，則會把該部份的資料過濾到搜尋結果的頂部。

B.參與會員的引導

RAS 系統是屬於封閉性的網路論壇，每個人都必須先註冊通過才能在系統內發表文章，所以每篇文章都很有很明確的作者記錄。而在論壇運作的過程中，除了會員發表的文章之外，會員間的互動也是一項很重要活動。因而本研究在搜尋結果中，除了提供主題的查詢結果外，亦希望能進一步提供會員間互動的機會。

本研究從查詢關鍵字所得到的結果中，經由會員權重的計算，提供活躍會員的推薦，以利使用者可以針對主題中的參與會員做進一步的資料過濾。並且在 RAS 系統內的會員，因彼此均屬於同一研究團隊，這項功能也可以成爲使用者進一步的諮詢對象。

在這個會員推薦的計算方式上，我們採用了兩項關於參與會員的計分標準，分別是 1.會員的參與密度和 2.會員的參與廣度。

a.會員的參與密度(DP)

參與密度的計算方法，是根據搜尋結果中，會員的發表文章總數（包含主題文章和回覆文章）除以會員出現的主題數加 1。例如，在查詢結果中，會員

A 發表了 10 篇文章，分別出現在 4 個主題中，則參與密度即爲 10/(4+1)=2。於計算方法中採用分母加上 1 的設計，乃是爲了抑制會員出現的主題數太少（即分母較小），導致所得到的數值相對較大的問題。

b.會員的參與廣度(WP)

參與廣度的計算方法，是根據搜尋結果中，會員出現的主題數除以該次搜尋結果的總主題數。例如，查詢結果出現的總主題數有 50 篇，而會員 A 出現在其中的 30 篇主題中，則會員 A 針對此查詢結果的參與廣度即爲 30/50=0.6。

最後，每個會員依照上述參與密度和廣度的計算，將兩項的分數相加以代表會員和該次查詢的相關程度，然後依據相關程度篩選出最高分的三位會員，推薦給查詢的使用者，呈現在搜尋結果的畫面中。

使用者可以針對被推薦的會員做進一步資料的過濾，將被點選的推薦會員曾參與的主題，優先排序於搜尋結果的前面部份；而除了被推薦的會員之外，有時使用者亦可能要找尋自己過去曾參與的討論內容，因此在這個部份的設計，也同時加入使用者本身的直接過濾。如下圖十二所提供的按鈕”我曾參與的”，就可以優先過濾出使用者本身曾參與的主題，排序於搜尋結果的頂部。

另外，有時會員並無法從搜尋結果中找到滿足的答案，則亦可從搜尋的結果畫面中，透過直接的連結，發送訊息向被推薦的活躍會員請教相關的問題。

圖十二：搜尋結果中活躍會員的推薦

圖十三：針對包含附件的主題優先過濾

C.參考資源的引導

基於前節中所述的 RAS 使用特性，本研究於搜尋結果的呈現介面中，先取出每個討論主題的所有附件檔案名稱，並提供在搜尋結果畫面中可直接下載的連結。而如果討論主題的附件有很多個，那麼系統預設是以曾被下載過最多次的檔案爲預設連結目標。這項功能可以讓使用者直接從搜尋結果畫面上快速取得各主題內所夾帶的附件，而不必進入每一個主題內瀏覽。

如此設計的好處是，若使用者是爲了尋找某一篇主題內的附件而進行查詢動作時，則可以在搜尋結果的畫面上直接下載，增加使用者對於文件取得的效率。

另外除了討論主題的附件連結，本研究亦藉由文章內容的分析技術，直接取出文章內容中所提供的超連結網址，並將網址連結的功能直接呈現在搜尋結果的畫面上，讓使用者可以從搜尋結果畫面上快速連結到該主題內容中所提供的網址。如下圖 7 所示，其中圖示  代表爲附件，  則代表相關網址。當使用者的搜尋是爲了尋找某篇主題中的附件，則可直接點選相關資源欄位中的“附件”按鈕，即可將所有包含附件的主題優先排序於搜尋結果的頂部，結果如圖十三所示。

五、研究建議

圖十四： Google 對關鍵字的偵錯功能

對未來研究的建議，整理如下：

1.搜尋關鍵字輸入的策略

本研究所使用的搜尋功能，是以關鍵字和簡易邏輯運算(AND、OR、NOT)的輸入方式來做文章內容的比對。然而以關鍵字的輸入方式，有時並不能完全滿足使用者的需求，例如有時使用者可能會想採取自然語言的方式輸入。因此要如何利用論壇內容的特性，以及現存於資料庫的文章資料，對自然語言的輸入內容做解析，取出適當的關鍵詞或相似詞，將是一個可供研究的方向。

除此之外，對於使用者輸入關鍵字的容錯能力(tolerance)，也可以爲使用者帶來好處。例如：著名的網際網路搜尋引擎 Google，便可以對使用者所輸入的關鍵字做自動偵錯或建議的功能(如圖 14)。因而如果能對各式各樣的論壇，利用資料庫裡面有結構的文章資料來處理關鍵字的容錯能力，也可以爲使用者在輸入關鍵字時帶來好處。

2.論壇資料庫的資料取回策略

在資料取回的技術上，本研究並未改變原內建式搜尋功能對於資料的取回能力，主要原因是考量資料庫做全文比對的效率問題。因而對於如何增加論壇內容中的資料召回，例如：利用「同義詞庫」的建立，或是「語意網」(semantic web)的推論技術，取回更多相關的資料，亦是一個值得研究的方向。

3.搜尋結果的排序策略

本研究中對於搜尋結果的排序策略，主要是根據 RAS 的使用特性來設計的，對於若將此一策略套用在其它論壇上，是否能得到同樣的使用者滿意程度，則尚需進一步的案例研究方可得知。例如對於關鍵字出現在文章中的位置、討論文章的發表時間、文章之間的次序關係、發表人的身份別等相關因素，在本研究中並未做任何權重的設計，因而是否能利用上述其它特性進行適當的主題排序設計，則可提供爲進一步研究的參考。

4.搜尋結果的知識擷取

本研究中對於搜尋結果的知識擷取，主要是針對附件描述及內文中參考網址的部份。對於是否能夠從搜尋的結果中，加以統計分析，擷取出其它對搜尋者有幫助的知識，則可以進行在其它案例上更進一步的研究。例如：對各主題進行語意分析或人工智慧技術中的「資料探勘」(Data Mining)，以產生主題自動摘要或分類，也都可以做爲後續有關搜尋過程中用來協助搜尋資料的自動過濾研究。

5.適性化的搜尋機制

本研究主要是根據搜尋關鍵字對搜尋結果做資料搜尋機制的改善，因而對於搜尋者本身的特性未予考量。即當使用者 A 和使用者 B 不同的兩個人，若輸入相關的查詢關鍵字，得到的結果和排序將是相同的，亦即未有適性化或個人化的搜尋服務。

適性化系統的建立，必須先搜集足夠的使用者資料，才能建造出合理的使用者模型(model)，也才能提供有效的適性化功能。然而本研究的實作案例 RAS 中，該系統因屬於特殊目的所建置的封閉式論壇，目前主要僅提供近二十位研究生參與，因此在建造適性化的模型上，有實際上的困難。鑑於此，未來的研究應可利用熱門且開放性的論壇爲研究案例，因爲這類的論壇通常會員的數量眾多，例如知名的網路論壇「酷！學園」(http://phorum.study-area.org/)，該論壇上目前已登錄有一萬二千名以上的會員，若能以此類熱門論壇建立適性化模型，可信度會較高，也較能針對適性化的搜尋策略提出更好的設計。然而論壇適性化的研究，通常會牽涉到會員資料隱私的問題，因而必須先獲取論壇內會員的同意方可進行，此點將是研究過程中需特別注意的一項因素。

參考文獻

[1] R. A. Wagner and M. J. Fischer, ”The string-to-string correction problem,” *Journal of the ACM*, Vol.21 No.1, 1974, pp.168–173.

[2] 呂俊彥, *中文自然語言查詢自動轉換爲SQL之研究*, 大同大學資訊經營研究所碩士論文, 2002。

[3] 劉洋, 劉群, 林守勛, “機器翻譯評測中的模糊匹配,” *中文資訊學報*,第3期, 2005, 頁45-53。

[4] 李宜揚, *以聲音內容爲主的音樂資料庫查詢系統*, 國立清華大學資訊工程學系碩士論文,1999。

[5] A. Banerjee and J. Ghosh, ”Clickstream Clustering using Weighted Longest Common Subsequences” in *Proceedings of SIAM Conference on Data Mining*, 2001, pp.33-40.

[6] S. T. Sun and Y. T. Ching, “Hypermedia browsing pattern analysis,” *International Journal of Educational Telecommunications*, Vol.1, No.2/3, 1995, pp. 293-308.

[7] S. L Chin, I. Jin, S. Thomas and H. Sing, ”Incomplete gene structure prediction with almost 100% specificity,” Texas A&M University, 2004。

[8] 楊博宇, *產品比對的研究*, 國立交通大學電機資訊學院碩士在職專班論文, 2004。

[9] 林俊博, *遊戲論壇搜尋引擎之設計*, 逢甲大學資訊工程所碩士論文, 2002。

[10] I. Lucas, “Defining the Web: The Politics of Search Engines,” *IEEE Computer*, Vol.33, No.1, 2000, pp.54-62.