在超連結環境下針對資訊分類相關權威網頁之探勘

Mining the Authoritative pages for Information Categorization in a Hyperlinked Environment

何裕琨
國立成功大學電機工程學系
台南市大學路 1 號
ykho@eembox.ncku.edu.tw

曾耀順
國立成功大學電機工程學系

摘要

隨著網際網路(Internet)科技的進步，與資訊量快速的成長，如何快速找到契合需求的資訊相對愈顯困難。網際網路之超文字（hypertext）是利用超連結（hyperlink）來建立相互之關係，超連結的結構中隱含著代表人類高程度的抉擇，當網頁與另一網頁建立連結，即代表此網頁對連結至的網頁在主題相關性上的肯定。在傳統之全文檢索模式無法有效率提供有關主題資訊的情況下，利用超連結之關係作資訊檢索是一個可行之辦法。

本論文利用網際網路的超連結（hyperlink）之結構，以有向圖（direct graph）的概念將網頁文件視為節點（nodes）及超連結視作有向邊（direct edge），以便利用有向圖形來計算超連結與相關網頁之關係，藉以判斷某網頁文件在某一主題上之重要性。透過連結的鏈結數分別設定權威權重（authoritative weight）與發散權重（hub weight），並利用權威網頁文件的平均權重值當作權重的臨界值（weighting threshold），藉以過濾掉相關主題性較低的權重，當作是否取捨網頁文件與進行資訊分類的依據。

利用真實網站分類目錄之實驗分析，證實本論文提出之以權重臨界值當作主題相關程度的判定之方法，可以有效進行主題分類的功能。透過空間向量模式（space vector model ）來計算查詢關鍵字與檢索結果網頁文件的相似度，發現透過連結權重計算之方式，的確能分類出具有較高主題相似度的網頁文件，並可達到良好的檢索效果。

關鍵字：資訊尋找、資訊分類、超連結、權威權重、權重臨界值。

一、簡介

在網際網路盛行及資訊大量增長的今日，由於網路包含的資料種類繁多而且變動性高，倘若無法有系統的整理這些資料，日後要查詢時會變的相當困難。但是在網際網路眾多的資訊中，使用者並非所有資訊都有需要，而因此如何幫助使用者找尋到重要的資訊，是一個值得探討的問題。

應用於網際網路的資訊檢索(information retrieval)技術[1,5,9,13]，便是解決此問題的方法，此種資訊檢索的技術已經存在相當久的時間，藉由網際網路的普及與網路資訊的推波助瀾下，以資訊檢索為基礎的網際網路搜尋技術已成為網路上經常使用之工具。

傳統上之全文檢索是較容易瞭解的一項資訊檢索技術，此種檢索的方式通常是以字串比對為其核心[1]，雖然全文檢索技術可以說是相當完備，但是使用者除了做比對的動作外，還要對比對後產生的資訊做瀏覽搜尋，所以最後找到適合需求的資訊往往不是一開始找到的，而是需要使用者透過瀏覽人工搜尋的方式從中整理所得來。全文檢索技術在使用上也有所限制，比如文件必須含有和查詢條件完全符合的文字才會被篩選出來，其比對方式是將查詢字(query word)和文件集合中每篇文件一一的做比對，然後從中找出所有符合查詢的文件。縱觀整個處理過程耗時費工而效果則可以預期的不甚完美。因此若能加入文件叢聚分類(clusters)的概念[14,15,16]，將包含各自的主題的文件，分類成特定主題的群集，於查詢時再打破每個文件都需要逐一處理的情況，而僅對部份與查詢字串高度相關性的主題分類聚集文件來做比對，如此將可以加速查詢之速度並可以提高查詢的效果。

理想的查詢模式，希望是能夠與文件之間的連結結構相互配合，如以科學類別的檢索系統為例，原始的資料庫是各種：如天文、海

洋、氣候等不同的主題領域的科學資料，但是使用者在做檢索時並不希望傳回整個科學類別的資料庫，使用者僅是希望找到相關喜好的主題資料。例如查詢者想找「天文學」相關的文件，並不希望查詢的結果是「科學」，而是希望查詢的結果為「天文現象」、「太陽系」，甚至是「宇宙」等資料，而此類資料對使用者才具有意義。

所以顯而易見的，以主題分類的資料庫，對查詢使用者而言，其檢索的結果會具有高相似於主題的文件，然而一般以字頻(word frequency)為主的排序(ranking)方法[2]，所檢索出之大量的資料，可能無法提供高相似度的資料。以經驗看來，具有符合查詢者的資料文件，通常排序的分數都不高，所以倘若只瀏覽前 20 到 30 檢索的結果，是無法找到適合的文件資料。

本論文研究之資料檢索技術，則是依據文件與文件之連結的網路結構，藉由分析連結的相互關係，來評估網頁文件的重要性，以探掘出符合主題的分類群組(subject classification)來。Jon M. Kleinberg [3]所提出超文字導引之主題搜尋(hypertext-induce topic search，簡寫成HITS)方法，便是利用連結的分析，檢索出最符合查詢主題的網頁文件[3,12,15,17]。針對此研究方向，本論文將著眼於超文字之超連結關係，透過分析連結的特性，分類出特定主題的分類群集，並將群組環境中的網頁文件做進一步的分類，以增加高相關主題網頁文件的擷取，減少不必要的人工篩選、瀏覽的工作，提高檢索的效率與品質。

本論文於第二節將對研究背景作介紹及資訊檢索與連結結構相關文獻的探討，第三節提出利用權重臨界值進行主題分類的構想，第四節建構出資訊檢索系統，第五節以 Lycos 網站內科學與新聞二大類別的網頁資訊當作主題分類驗證的實驗資料，並針對主題分類結果進行分析與討論，第六節針對本論文進行結論。

二、連結結構

傳統的資料搜尋技術重點是以文字為主，但隨著網路資源的豐富多樣化，除了文字本身以外，超文件中的連結關係亦佔有重要的地位[5,6]，藉以分析文件連結間之重要性，並可利用此連結資訊來協助判斷網頁文件之間的相關性。

在網頁文件連結之研究中 [4,6,8]，是利用超文字中連結間關係，把相關文件進行集合並呈現給使用者。研究的方法是透過圖形概念，將網際網路表示成一有向圖(direct graph)，網頁文件代表圖形的節點（nodes），連結則代表成有向性的邊（direct edge），利用圖形的架構作連結方面的計算分析來尋找網頁文件間的相關性。

在 J. Carriere and R. Kazman [6]之研究中針對網頁重要性的排序法，指出一篇網頁的重要性相當於其引用其他網頁的連結數，與被其他網頁所連結的數目和。而 Jon M. Kleinberg [3]則指出傳統的檢索結果總是無法符合使用者的需要，使用者不只是要和查詢關鍵字相關的網頁文件，而是希望從中找出最切合查詢關鍵字主題的網頁資料，而通常檢索後的結果數量相當大，乃是因為自動檢索索引中缺乏人類的判斷。

超文字之中事實上已隱含了大量的人類的抉擇，當網頁文件與另一網頁文件建立連結，即代表此網頁文件對另一網頁在主題相關上的肯定，所以在連結關係中，網頁文件 A 對於某種主題的重要性，就等於其他網頁連結到此網頁文件 A 的肯定。而網頁文件 A 則可以當作提供某種主題資料的來源網頁，因此稱網頁文件 A 為一權威網頁（authority page）。

倘若網頁文件 A 為相當通用的網頁，將會存在著許多連結至該網頁之文件，但網頁文件 A 未必是對特定主題是相關的，為了避免這種情況發生，經由觀察，一個連結到權威網頁的網頁本身，應該是個很好的發散網頁（hub page）。權威網頁和發散網頁之間包含著相互共利而生的關係：亦即一個好的發散網頁將是連結到許多好的權威網頁，而一個好的的權威網頁則是被許多好的發散網頁所連結的。因此權威網頁和發散網頁有相互加強關係的存在。圖 1 可見權威網頁和發散網頁相互連結關係。據此，每個網頁均應可定義兩個值，而亦可用此兩值判定網頁對特定主題之重要性。下列為權威權重與發散權重之定義：

- ＊ 權威權重（authority weight）：網頁 A 之權威權重值為所有連到它的網頁發散值之和。
- ＊ 發散權重（hub weight）：網頁 A 之發散權重值為所有它連結至網頁的權威權重值之和。

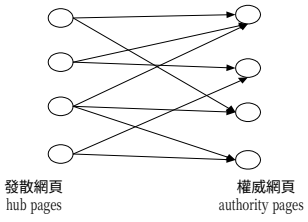

圖 1 權威網頁和發散網頁連結關係

David Gibson[18]之研究中利用連結做網頁分析的工具，而不需要牽涉到文字的部分，當系統輸入一網頁後，輸出的為一群與輸入網

頁相關的網頁群。其主要探討的是輸入不同的網頁文件、不同的語言等參數的選擇對於結果上的差別。作者根據連結的結構定義出個別的權威值與發散值，並依據此做運算。最後發現單單使用連結的結構來作網頁文件的分析，已經可以對網頁做出正確的判別，肯定連結結構存在有利於使用者之查詢。

三、主題分類

(一) 超文字導引之主題分類

Jon M. Kleinberg[3]所提出之超文字導引之主題搜尋（hypertext-induce topic search，HITS）是以權威網頁和發散網頁為基礎，針對連結關係，對特定主題作搜尋的方法。HITS 的方法中包含了二個主要的部分，第一是取樣（sampling），尋找當做起始網頁文件之集合，盡可能的建構切題、有關的網頁集合。第二是權重增值 (weight-propagation)計算，在網頁文件集合中，評估計算權威權重與發散權重。

在取樣（sampling）的步驟中，是利用傳統關鍵字為主的檢索模式去收集檢索大約 200 筆的網頁文件，然後將收集的資料集合建構成一名稱稱為根集合（root-set）的有向圖，再利用根集合延伸成基礎結合（base set）。

在權重增值 (weight-propagation)之計算步驟中，是計算基礎集合中，每個網頁文件的權重值。權重值的計算目的是為了對網頁文件之排序，及定義對於某種主題的相關性，因此需要反覆計算基礎集合網頁文件之權威權重和分散權重。下面數學式子表此二權重之計算：

$$Authority\_weight = \sum_{\forall q;q\to p} Hub\_weight$$

q → p 代表網頁 q 連結至網頁 p

$$Hub\_weight = \sum_{\forall p;p\to q} Authority\_weight$$

p → q 代表網頁 p 連結至網頁 q

由上列之權重計算式子可知，某一網頁倘若被大量網頁文件所連結，其權威網頁權重值將會提高，或者是被許多高權重的發散網頁連結，也會提高網頁的權威權重值。相反的，假如網頁文件連結至許多網頁文件或者是連結至高權重的權威網頁，其發散網頁的權重也會有所提高。

在本論文研究中，我們亦將透過分析連結結構，針對主題（topics）分析方面計算出網頁文件之權威權重和發散權重的值，並依照權重數值的大小，來判定網頁文件與主題關係之相關程度，並藉此判別分類的群組。

(二) 主題間相關程度之計算

另一種計算網頁文件權威權重和分散權重的方式是利用矩陣的計算方式[4,8]。將基礎集合中包括的網頁文件 $\{1,2,3\cdots,n\}$ 加以編號，並定義一 n ＊ n 的相鄰矩陣 A 之元素 $a(i,j)$。

$$a(i,j)=\begin{cases} 1 & ;if\ \ page\ \ i\ \ link\ \ to\ \ page\ \ j \\ 0 & ;otherwise \end{cases}$$

並將所有網頁文件集合的權威權重改寫成向量 W＝（w_1，w_2，…….，w_n）相同地亦將網頁文件發散權重寫成向量 H＝（h_1，h_2，….，h_n），再將加總計算權重的方式改寫成 $W \leftarrow A^{\mathrm{T}}H$ 與 $H \leftarrow AW$，將二式子展開可得到：

$$W \leftarrow A^{\mathrm{T}}H \leftarrow A^{T}AW = \left(A^{T}A\right)W$$
$$H \leftarrow AW \leftarrow AA^{T}H = \left(AA^{T}\right)H$$

因此觀察上列之式子，其展開相乘計算產生的結果與 $A^{\mathrm{T}}A$ 幕次方展開相乘計算相同。線性代數[7] 證明向量 W 經過展開計算後，會收斂至 $A^{\mathrm{T}}A$ 的主要特徵向量（principal eigenvector）。相同的，向量 H 經展開計算亦會收斂至 AA^{T} 的主要特徵向量。依據分別計算 $A^{\mathrm{T}}A$ 與 AA^{T} 的特徵向量與特徵值，在 $A^{\mathrm{T}}A$ 與 AA^{T} 具有相同的特徵值情形下，其特徵值表示對某主題的網頁之群集。$A^{\mathrm{T}}A$ 的特徵向量代表權威權重網頁文件，而其特徵向量中數值愈高代表網頁文件的權威權重也愈高。相同的，AA^{T} 的特徵向量代表發散權重網頁文件，而其特徵向量中數值愈高代表網頁文件的發散權重也就愈高。

最後產生的結果則是列出向量 W 和向量 H 中具有最高權威權重和分散權重，並依據權重數值的大小加以判定，就可以判斷網頁文件對特定主題間的遠近程度和高權重值網頁文件中連結的密度。

因為每個網頁文件包含著不同總數的發散出（outgoing link）和連結（incoming link）的連結個數，因此以超文字文件的結構特性與連結的建立，可透露利用連結建立連結的網頁文件，其對於包含相同主題的機率大於沒有連結的二網頁文件[10]。

據此結果，在本論文研究中我們將界定網頁連結的個數並計算平均權威權重值以作為權重臨界值（weighting threshold），藉以過濾掉主題性較低的網頁文件，以區分出高關聯主題性的權威網頁文件。此一構想將有助於降低查詢結果的數量，並增加主題相關性高網頁文件之聚集效果，以提高檢索結果的精確度。

(三) 利用權重臨界值之主題分類

在本論文之主題分類系統中首先於網際網路上搜尋網頁文件集合（如於分類目錄建立方式的網頁文件資料庫找出起始的 URL），當選擇好起始的網頁集合 S（initial set）後，對於所選擇的網頁文件和其對應之 URL 則依據人工至網站進行搜尋網頁文件的集合。例如科學科技、運動體育與休閒天地等等。此網頁文件及其 URL 係透過人建立於分類目錄中並提供使用者瀏覽查詢用，在忽略網頁文件中內部瀏覽連結外（如回首頁或回上一頁的連結等），這些網頁文件本身並不包括發散的連結（outgoing link），因此由起始網頁文件集合中所選擇的網頁文件視為權威網頁。

完成收集了起始網頁文件集合 S 後，接下來需要找出連結至起始網頁文件中的發散網頁文件，這部份將利用現行的搜尋引擎，這些搜尋引擎儲存連結的資訊，而且提供"哪些網頁連結至查詢者所指定 URL"的查詢功能[17]，例如：可以在查詢欄中鍵入"link:http://www.ncku.edu.tw"，便可以檢索出連結至 www.ncku.edu.tw 網站的網頁文件資訊，透過此種功能找出連結至起始網頁集合的發散網頁，此種查詢功能，Google[21]與 Altavisa[20]皆有提供。

然後在起始網頁文件集合 S 中，找出 m 個連結至起始網頁且具有高數值的發散網頁文件並且將這些網頁文件加入於發散網頁集合 H 中，在完成找出權威網頁和發散網頁後再分別計算其權威網頁文件權重值：

$$Authority\_weight = \sum_{\forall q; q \to p} Hub\_weight$$

$q \to p$ 代表網頁 q 連結至網頁 p，與發散網頁文件權重值，

$$Hub\_weight = \sum_{\forall p; p \to q} Authority\_weight$$

$p \to q$ 代表網頁 p 連結至網頁 q。

計算完成權重值後再求其權威權重值的平均當作分類主題的權重臨界值(weighting threshold)。在進行分類時，倘若其所發散網頁文件連結的權威權重值大於權重臨界值代表其此發散網頁主題相關聯性較高將予於留下進行分類，否則將之過濾並排除在分類動作外。

由連結分析可知假如發散網頁權重沒有超過一定數量的權威網頁文件或其發散網頁權重數值過低，其與主題的相關性就隨之降低，所以將這些相關主題性低的網頁文件刪除，可加強網頁文件對於主題的相關程度。在分析計算在發散網頁集合中的網頁文件後，倘若發散網頁集合 H 中網頁文件連結至起始網頁集合 S 中的網頁個數沒有超過 x 個或其連結權威網頁的數值沒有超過全部權威網頁權重的平均值（權重臨界值）之上，就可將此網頁文件由發散網頁集合 H 中刪除。

接下來經過分析計算所得的發散網頁文件集合 H 和起始網頁文件集合 S，建立一個 n 維度的相鄰矩陣 A，矩陣維度 n 的大小是起始網頁文件的數目與發散網頁文件的數目之和，其中假若網頁文件 i 連結至網頁文件網頁文件 j 則 a（i，j）=1，否則 a（i，j）=0。最後計算 A^TA 的特徵值和特徵向量，每一個特徵值代表一個分類集合，並且對應於特徵值的特徵向量代表所屬分類網頁文件集合。圖 3-1 說明主題分類的分析步驟。透過主題分類後的網頁文件與對應之 URL 為進行網頁文件索引（indexing）重要的資訊內容。

Step 1.檢索起始網頁文件集合 S。
Step 2.擷取 m 個連結至起始網頁（initial set）S 且具有高數值的發散網頁文件並且將這些網頁文件加入於發散網頁集合（hub set） H 中。
Step 3.計算起始網頁文件集合中權威權重與其平均值(average)當做為權重臨界值。

$$Average\_authority\_weight = \left(\sum_{\forall j \to i} Hub\_weight\right) \Big/ n;$$

j is in hub set， *n*：*is number of authority pages*

Step 4.倘若發散網頁集合中網頁文件連結至起始網頁集合 S 中的網頁個數沒有超過 x 個或其連結權威網頁的權重值沒有超過全部權威網頁權重的平均值之上，就將此網頁文件由發散網頁集合 H 中刪除。
Step 5.計算 起始網頁文件數目與發散網頁文件數目之合 n。
Step 6.建立一 n 維度的相鄰矩陣 A。
Step 7.計算 A^TA 特徵值。

圖 3-1 主題分類分析步驟

四、資料檢索系統

(一) 資訊檢索系統之架構

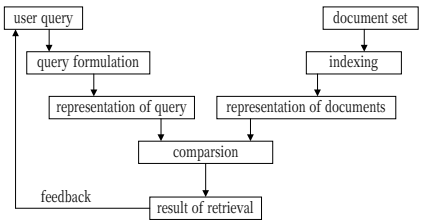

圖 4-1 資訊檢索模式

圖 4-1 之資訊檢索模式是以文件資訊為例來說明整個資訊檢索的處理過程。右邊之文件資訊依據適當的組織加以集合，將文件資料集合轉換成對應的代表特徵加以呈現；左邊則將使用者的需求轉換成查詢。然後將文件資訊對應的代表特徵與使用者的查詢來比對，以便檢索出符合的文件資訊。最後則將檢索的結果回饋給使用者，使用者若有需要亦可參考送回之結果對於查詢需求再加以修正。

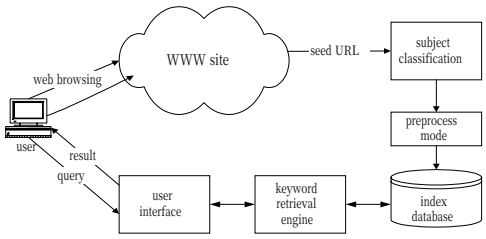

圖 4-2 資訊檢索系統架構圖

圖 4-2 為本論文之檢索系統之架構，依據系統建構方式可分為成三個主要部分，第一部份為第三節所描述主題分類網頁文件部分，第二部份為使用者介面，第三部分為網頁文件的處理與建立索引檔。以下就對各個部分加以說明：

在每一次檢索查詢處理過程中，使用者首先利用使用者查詢介面（user interface）取得查詢的首頁並鍵入要求，每一個查詢首頁上的輸入項目代表查詢的要求變數，當查詢者鍵入查詢字和選擇要求選項後，將會利用查詢首頁的搜尋按鍵送出搜尋要求，使用者介面則透過 CGI 程式取得要求的變數然後送至關鍵字檢索引擎（keyword retrieval engine），關鍵字檢索引擎依據輸入的變數如關鍵字 ”and”（全部）、“or “（任一）及”except”（除了）的關係彙整後再與索引資料庫（index database）的索引檔之資料作關鍵字的比對，比對後的查詢結果內容，再透過使用者介面顯示給查詢者，以完成一次的檢索。索引檔的建立方式則是至網際網路各個網站搜尋出種子 URL（seed URL）並對網頁文件作摘要描述，並將這些資訊傳送至主題分類模組（subject classification），分類或特定主題相關的群組，並同時建立索引檔(index files)。

圖 4-2 之資訊檢索系統各模組之功能，分述如下：

1.使用者介面（user interface）
提供 Web 使用者單一、友善、易操作的介面。

2.關鍵字檢索引擎（keyword retrieval engine）
以使用者鍵入的關鍵字查詢為主，透過使用者介面輸入的查詢字與索引資料庫中的反向索引檔的索引作比對，將比對後符合檢索的結果顯示於查詢者使用介面。

3.主題分類（subject classification）
主題分類模組將傳至的種子 URL，利用連結至種子 URL 的網頁和由種子 URL 連結出的網頁作集合，透過連結結構分析做主題分類的動作。

4.前置處理（preprocess mode）
前置處理是要找出文件中較能代表該文件，相對而言比較重要的關鍵字字彙，關鍵字在資訊檢索中是代表一種表現文件資訊特徵的方法。因此在索引過程中，需先對網頁文件作有條件的區分，刪除某些對檢索過程毫無幫助的字彙，保留有用之關鍵字。

5.索引資料庫（index database）
透過前置處理後的文件接下來再依據適當的組織加以集合，並建立索引，其建立索引的方式和傳統檢索系統建立索引方式類似，稱為索引資料庫或稱作反向索引（inverted index file） [11]。

(二) 資訊檢索運作流程

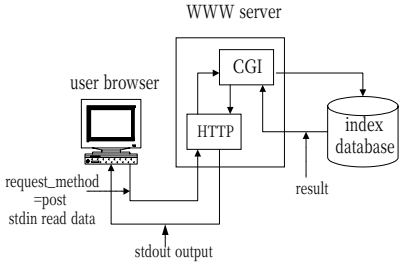

圖 4-3 資訊檢索運作流程

圖 4-3 為資訊檢索運作流程，首先在使用者端，查詢者透過使用者介面上的 “搜尋” 按鍵，按下並進行查詢，告知使用者瀏覽器（user browser）完成輸入，使用者瀏覽器則將查詢者端輸入的資訊傳送至 WWW 伺服器並包裝成 HTTP 要求的訊息，隨後伺服器啟動指定的程式並將包裝過的參數傳入，接著程式依照傳入的參數與索引資料庫中索引檔作比對查詢，當檢索查詢完成後，傳回符合檢索結果至 WWW 伺服器，網站伺服器再將結果包裝成一個 HTTP 回覆的訊息傳回給查詢使用者，完成資訊檢索流程。

在整個傳遞資訊過程，為了要從伺服器傳遞有關輸入資訊給 CGI 程式，WWW 伺服器將不同的資訊轉換為各種環境變數以供程式使用，而這些環境變數是在 WWW 伺服器執行 CGI 程式時被設定的。有關資訊檢索程式如何將使用者經由使用者介面輸入的資料進行處理，並送至檢索查詢程式。

在網頁表格（form）中，每個資料輸入欄中必須包含對應的 value 參數值，再被轉換成”name=value”的型式，而不同的資料輸入欄，之間會以”&” 作為間隔，成為”name1=value1&name2=value2”之型式。如果參數中間有空白字元會將以”+”代替，有特殊字元則以”%XX”代替，其中”XX”為特殊字元的 ASCII 碼，這些特殊字元包括”&”、”=” 及 ASCII 為 128 以上的字元。資料經過處理包裝後，再利用 HTML 中”post”的方法（method）發出查詢檢索 CGI 要求，然後資料則以標準輸入串列（standard input ，縮寫為 stdin）的方式送至查詢檢索的程式。

當查詢檢索完成後，如果要將輸出資訊傳給使用者端，則查詢檢索程式會將檢索的結果資料利用標準輸出串列（standard output 縮寫 stdout）以資料流的方式傳回給 WWW 伺服器， WWW 伺服器則負責把資料流以超文字文件傳輸協定（HTTP）的形式包裝起來，並利用 HTTP 將檢索結果的資料流轉換給使用者的瀏覽程式（browser）並呈現檢索結果給使用者。

(三) 前置處理

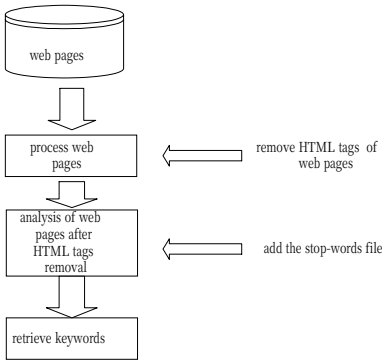

圖 4-4 網頁文件前置處理

網頁文件資料其中包括了網頁本文內容和 HTML 的標記（tags），而使用者感興趣的部分則是網頁本文內容，倘若要將 HTML 標記作索引（indexing），將是沒有必要也是耗工費時且浪費儲存空間的。因此前置處理的首要步驟就是作去標記的動作，接下來就是對去 HTML 標記的網頁內容作詞彙的處理。在做詞彙處理前需建立一個名為 stop-words 之檔案，這個檔案中儲存的詞彙是一些需要過濾的字彙，如 a、an、is 與 the 等等，不過 stop-words 檔案的內容通常比較偏主觀的判定，因此此檔案的建立端看設計者設計而定。將一些不必要的字彙去除後，建立索引的量就可以有所減少而不會影響查詢的結果。

圖 4-4 為前置處理的子模組，將網頁文件資料首先作去除 HTML 標記（tags）的處理工作，由於此系統無支援大小寫搜尋的功能，因此經過去除 HTML 標記的網頁文件後，將將文件資料轉換成小寫，接下來則是需要從去除標記的網頁文件找出關鍵字，其過程是將除去 HTML 標記後的網頁文件與 stop-words 檔案中所列出的字彙做比較後，將網頁文件中出現如 stop-words 檔案所列的字彙去除後，所得的字彙便是關鍵字，然後以便將來作進行索引（indexing）動作的資料。表 4-1 所列是較不具代表性的詞彙，如 when、will、someone 與 sometimes 等等，亦是 stop-words 檔案之內容。

表 4-1 stop-words 檔案所列字彙

a list of stop-words a 、all 、also、an、as、at、and、any、are、about、be、by、but、can、for、from、have、here、i、if、in、is、it、no、not、of、on、or、our、

(四) 索引資料庫

若以文件為主的角度來看某一文件包含了哪些字彙，若在每一列中該列的每一欄位以 1 代表文件中包含該行所代表的字彙，以 0 代表該文件不包含該字彙，依此方式可以建構出一個文件對應字彙的陣列，再將此陣列進行轉置（transpose），每一列以字彙為單位，分別紀錄了該字彙於所有文件中出現的相關資訊，則此這每一列可稱為反向索引檔案（inverted index file），在實際儲存每一個反向索引檔案時，為了避免稀疏（sparse）的情況，可以改成儲存所出現的文件編號。在系統中前置處理後的網頁文件內容將和其 URL 建立一個反向索引檔（inverted index file），其檔案內容結構為

（Document_item ；Document_number，
Documen_item_frequency）

，如圖 4-5 所示: 其中 Document_item 代表索引資料庫的索引欄位，亦即是網頁文件出現的關鍵字，Document_number 代表出現此關鍵字 item 的網頁文件編號，
Document_item_frequency 為此關鍵字 item 出現頻率或者出現在網頁文件的位置。

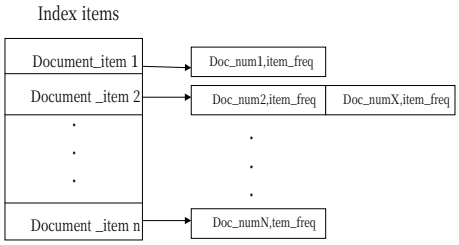

圖 4-5 反向索引檔

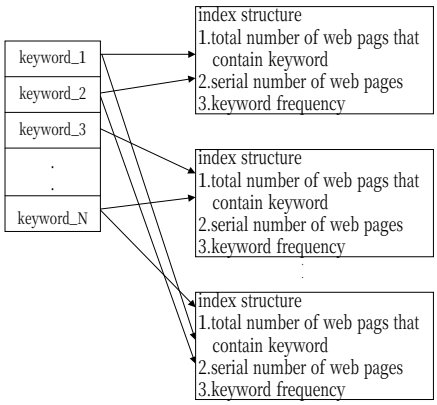

圖 4-6 索引檔結構

為了讓反向索引檔可以儲存更多的資訊，以簡化檢索的處理或提供檢索時設定查詢條件之用，除了文件編號外，還可以包含該字彙在文件出現的位置、出現的頻率或權重等等。索引包含的資訊越多所能提供查詢的資訊也越多，當然所使用的索引空間也愈大。

經過前置處理後的網頁文件和其 URL 所建立之反向索引檔 （inverted index file），其結構如圖 4-6，其中還包括了索引檔的索引部分，此部份是由網頁文件中所擷取的關鍵字所組成，其後所對應的則是出現該關鍵字的文件編號、出現關鍵字次數的頻率與出現關鍵字的網頁文件的筆數。

五、主題分類之測試

(一) 分類之結果與分析

本系統係利用 Lycos 網站分類目錄資料庫中的網頁文件內容作為研究系統分類實驗測試的資料，藉以評估本系統主題分類的群集。所擷取的測試資料總共包括由分類目錄資料庫中找出的科學類別與新聞類別的網頁文件，藉由分類目錄資料庫中包含豐富主題性的網頁文件資料，並針對主題分類中網頁文件連結不同的個數予於研究，因為連結的個數，在分析計算網頁的權威權重和發散權重會有直接的影響，而且高權重數值的權威網頁文件必定被眾多的發散網頁文件所連結，和高權重值的發散網頁也會連結至眾多的高權重值的權威網頁文件。所以在分類實驗測試中分別設定發散網頁文件連結至權威網頁的連結個數並依據連結特性的分析測試分類的效果。

首先在第一個分類測試實驗中，由分類目錄擷取約 1000 筆科學類的網頁文件，分別包括 400 筆生物學主題的網頁文件、300 筆物理學的網頁文件與 300 筆天文學的網頁文件等等不同類別的網頁文件。在測試主題分類模組裡，分別設定連結至起始網頁的發散文件個數 100 筆，與每個發散網頁文件連結至權威網頁文件的最小連結數目，分別設定為 2、3 與 4。在表 5-1 中發散網頁文件連結至權威網頁文件的最小連結數目設定為 2，在表 5-2 中發散網頁文件連結至權威網頁文件的最小連結數目設定為 3 與在表 5-3 中發散網頁文件連結至權威網頁文件的最小連結數目設定為 4。

表 5-1 分類科學類別資料實驗結果（一）

網頁文件連結最小個數為2 平均權威權重值為90.21				
科學類別	生物學	物理學	天文學	最多分類網頁文件數
主要特徵值 (84.20)	212	32	56	212
特徵值 (62.73)	174	65	61	174
特徵值 (42.86)	15	77	208	208
特徵值 (40.55)	39	77	184	184
特徵值 (24.79)	17	188	95	188
特徵值 (8.28)	48	103	149	149

表 5-2 分類科學類別資料實驗結果（二）

網頁文件連結最小個數為3 平均權威權重值為85.04				
科學類別	生物學	物理學	天文學	最多分類網頁文件數
主要特徵值 (78.66)	190	36	74	190
特徵值 (74.53)	184	51	65	184
特徵值 (66.22)	27	97	176	176
特徵值 (40.34)	33	97	170	170
特徵值 (24.77)	41	172	87	172
特徵值 (10.19)	52	153	95	153

表 5-3 分類科學類別資料實驗結果（三）

網頁文件連結最小個數為4 平均權威權重值為82.53				
科學類別	生物學	物理學	天文學	最多分類網頁文件數
主要特徵值 (60.11)	134	92	74	134
特徵值 (43.87)	49	83	168	168
特徵值 (24.31)	13	183	104	183
特徵值 (10.97)	94	127	79	127
特徵值 (10.22)	102	103	95	103

表中第一行代表經由主題分類模組方式透過連結分析計算所得到的特徵值，第二行至第三行代表分類網頁文件的筆數，第四行表示最多分類網頁文件的總數。而在計算特徵值和特徵向量時，由於會有不同的特徵值產生，在此實驗部分將會產生二個以上特徵向量的特徵值不予考慮，原因是因為它對同一個的特徵值產生了不同特徵向量。而在進行分類時，依據特徵值所得到的特徵向量，並透過特徵向量中每一個項目的數值大小判定主題性的相關程度並將所對應高數值大小的網頁文件輸出，倘若一個特徵值有二個以上的特徵向量，在進行篩選分類將會有不合理的結果產生，所以對此種情況不予考慮。

觀察 表 5-1 與 5-2 中顯示生物學類別的網頁文件具有很好的分類的效果，在特徵值為 84.20 與 62.73 時，均可以分類出高於其他二種主題類型的生物學類別之網頁文件，在其餘的特徵值中可以發現天文學和物理學類別的網頁文件有分類在同一個集合的現象，其原因是物理學和天文學在理論的相關性上比天文學與生物學網頁文件的相關性較為相近，其中天文學的發散網頁通常也連至物理學的網頁文件，因為天文學與物理學網頁文件之間具有高相似主題的關連程度。

接下來針對新聞類別的網頁文件集合做主題分類的實驗，我們同樣的將發散網頁文件連結至權威網頁文件的連結最小個數分別設定 2、3、4，實驗的網頁數目 800 筆新聞類型的網頁文件資料，其中包括 200 筆運動類、200 筆天氣類、200 筆健康類以及 200 電腦科技類的四種不同類型新聞性網頁文件資料。連結至起始網頁的發散網頁個數設定為 100。

表 5-4 分類新聞類別資料實驗結果（一）

網頁文件連結最小個數為2 平均權威權重值為76.58					
新聞類別	運動	天氣	健康	電腦科技	最多分類網頁文件數
主要特徵值 (60.20)	109	54	18	19	109
特徵值 (45.38)	8	20	62	110	110
特徵值 (40.50)	7	22	53	118	118
特徵值 (25.27)	7	20	40	133	133
特徵值 (13.81)	9	22	66	103	103
特徵值 (6.65)	109	54	10	27	109

表 5-5 分類新聞類別資料實驗結果（二）

網頁文件連結最小個數為3 平均權威權重值為74.20					
新聞類別	運動	天氣	健康	電腦科技	最多分類網頁文件數
主要特徵值 (61.55)	109	53	15	23	109
特徵值 (52.37)	8	22	59	111	111
特徵值 (42.10)	14	20	60	106	106
特徵值 (30.84)	8	24	56	112	112
特徵值 (20.16)	7	21	57	115	115
特徵值 (8.89)	104	48	24	24	104

表 5-6 分類新聞類別資料實驗結果（三）

網頁文件連結最小個數為4 平均權威權重值為71.34					
新聞類別	運動	天氣	健康	電腦科技	最多分類網頁文件數
主要特徵值 (57.39)	91	50	25	34	91
特徵值 (44.62)	21	29	42	108	108
特徵值 (30.16)	16	31	53	100	100
特徵值 (25.43)	8	32	64	96	96
特徵值 (10.65)	21	28	59	92	92
特徵值 (7.65)	96	40	42	32	96

於表 5-4 進行主題分類結果發現，一般使用者常常透過網路了解天氣的狀況，也因此在每一個對應的特徵值中皆有保持約 20 筆以上的群聚網頁文件，其分類數目能平穩保持在一個數量以上而且變化幅度業沒有其他主題大，在天氣這類別就具有一定數量的群聚性。運動群集的網頁文件中也明顯的與健康和電腦科技有明顯的區別，在健康與電腦科技這二個主題的網頁文件有分類於同一集合的現象，其原因是醫學技術的進步，不管是利用電腦輔助醫學或者是時下流行發展的生化科技，將健康與電腦科技的相關程度拉近了不少，當然無形中也加強了彼此的相關連性。

由實驗分類新聞與科學類別資料的結果分析發現，將發散網頁文件連結至權威網頁文件的連結最小個數分別設定為 2、3 與 4，觀察實驗的結果發現在連結最小的連結個數為 2 和 3 時，進行的分類群聚數目較為相近，而當我們將連結的最小個數設定在 4 時，會發現聚集的效果不如最小連結的個數為 2 與 3，探究

其原因，雖然連結愈多可以代表愈多人對此網頁文件的肯定與其受歡迎程度，不過將具有連結個數為 2 或 3 的網頁文件刪去，因為這些網頁中包含具有對於某種主題相關性較高的權威網頁或發散網頁，雖然其連結的個數小於 4，但其對於某些主題的相關程度，在計算權威權重和發散權重時往往是具有高數值權重的，所以倘若將這此連結數目少，不過對於分類主題具有影響效果的網頁文件排除在外，則會減少主題分類中網頁文件的數目，而且其主題的相關性也會隨之減少，所以依據實驗結果討論可以得知發散網頁文件連結至權威網頁文件的連結最小個數較佳值約 2。

(二) 網頁文件相似度之評估

經由實驗的分析可知利用主題分類方式可以有效的群集網頁文件的資料，在本系統曾利用主題分類的方式分類網站所擷取的網頁文件 4000 筆，透過主題分類的方式進行主題分類為書籍、新聞、商業、遊戲、科學、電腦通訊以及醫療等等共 1000 筆網頁文件之資料，並將主題分類後的網頁文件的集合送至前置處理子模組處理網頁文件並進行索引，在經由使用者透過使用者介面作關鍵字的查詢，比對索引檔中的索引並找出檢索結果。

在前一部份之分類實驗結果證實了主題分類可以有效的分類主題相關類型的網頁文件，但是卻無法評估經由查詢分類網頁文件集合所產生的結果與使用者查詢關鍵字的相似度，因為透過分類進行網頁文件的方法是增加主題相關的網頁文件的群聚，並未對網頁文件的全文內容進行分析統計，所以我們將利用空間向量模式（space vector model ）[12]來計算關鍵字與檢索結果網頁文件的相似度的評估。

在資訊檢索過程對於網頁文件本身內容進行分析以幫助檢索的進行，這個分析過程就是系統所進行的索引處理。在建立索引過程中透過分析文件內容，找出一組屬性具有足夠的資訊用以代表此文件，空間向量模式的應用，則是利用文件內容的字彙的權重或出現的頻率作為代表該文件的屬性，而統計字彙的權重或頻率之工作則是在索引資料庫中建立索引檔時完成。由於網頁文件是由許多的字彙所組成，因此可以利用文件中有意義的字彙組成文件向量，查詢關鍵字作文件相似度之估計。

進行方式則是利用文件向量與查詢本身所組成的二向量之間的接近程度作衡量標準。在使用者利用關鍵字查詢後的檢索結果，可利用用空間向量模組中向量夾角的餘弦值以式（1）[12]來評估我們利用主題分類所檢索出的結果與查詢關鍵字的相似程度，而查詢向量與網頁文件向量之間的夾角愈小，代表其相似度愈高。

$$\cos Similarity(D_i, Q) = \frac{\sum_{i=1}^{t}(d_{ik} \bullet q_k)}{\sqrt{\sum_{i=1}^{t} q_k^{\,2}} * \sqrt{\sum_{i=1}^{t} d_{ik}^{\,2}}} \quad \text{.....(1)}$$

其中 d_{ik} 是網頁文件 i 中字彙 k 的權重，q_k 是查詢向量中字彙 k 的權重。

在相似度之評估中，本文將利用 5 個不同的查詢關鍵字於本系統檢索出網頁文件資料，並利用式（1）計算檢索結果網頁文件與查詢關鍵字相似程度，所取樣的檢索網頁文件為全部傳回給使用者的網頁文件，而非前單 20 或前 30 筆網頁文件，利用關鍵字相似度的計算藉以評估利用分類方式做主題篩選分類的網頁文件是否能高相似於使用者所檢索的關鍵字。

表 5-7 為利用檢索結果的網頁文件向量與關鍵字向量相似度的計算結果。實驗中並加入利用現在普遍被使用者使用的搜尋引擎，Google、Yahoo 與 Kimo 作關鍵字的查詢檢索，由於利用此類型搜尋引擎所檢索結果是極為大量，而此類型的檢索結果將排序值較高網頁文件當作傳回給使用者瀏覽的排序（ranking）標準。因此在此部份實驗取每個檢索出結果的前三十筆網頁文件，作為關於查詢關鍵字相似性的計算。藉以評估利用連結分析分類方法檢索出的網頁文件是否能高相似於使用者查詢的關鍵字。

表 5-7 向量相似度的計算結果（一）

查詢關鍵字	檢索出網頁文件總數	平均餘弦值
health	31	0.571
news	82	0.694
java	4	0.362
network	29	0.575
economic	22	0.648

表 5-8 向量相似度的計算結果（二）

查詢關鍵字	Google	Yahoo	Kimo
health	0.613	0.542	0.580
news	0.510	0.462	0.416
java	0.571	0.495	0.547
network	0.590	0.445	0.349
economic	0.676	0.566	0.654

分析相似度計算之結果如表 5- 8 所示，利用主題分類方法所檢索出的網頁文件與查詢關鍵字的相似度餘弦值平均約在 0.50 以上，而餘弦值愈高則表示其二向量之間夾角愈小，相關性也就愈高。不過在結果中也發現關鍵字 java 其檢索的網頁文件與查詢關鍵字的

相似度偏低，其原因在於進行搜尋網頁文件時對於取樣相關類型主題資料的不足，倘若取樣的網頁文件廣度不足，勢必會影響查詢的結果，這是此系統效能的限制。由比較表 5-7 與表 5-8 實驗所檢索的網頁文件內容關於查詢關鍵字的相似性可以接近甚至優於其他三種搜尋引擎所檢索的網頁文件資料。其結果也顯示利用連結分析進行主題分類的方法，可以得到高相似於使用者查詢的關鍵字，縱使主題分類方式並非統計分析字彙重要性而進行分類，實驗結果也證明利用連結分析主題分類的方式可以得到較高之相似度於查詢關鍵字的網頁文件。

六、結論

隨著網際網路快速的成長使得利用網路尋找資料的使用者面臨許多的問題，對於一般使用者來說，如何取得真正需要的資訊是一個最迫切的問題。本論文是針對將主題性高的網頁文件資料進行分類，加強文件彼此對於主題的關聯性，並透過連結結構的分析進行網頁文件資訊的分類，過濾主題相關性較低的網頁文件，提高檢索資訊的精確度以幫助使用者有效率的瀏覽檢索資料。

分類測試實驗的結果證明本論文所提利用連結結構分析進行網頁文件作主題性的分類的方法的可行性。利用 Lycos 網站分類目錄資料庫中的科學和新聞類別的網頁文件當作測試資料之主題分類的實驗亦證明若是界定網頁文件間連結的鏈結數愈小，那麼可以分類出的資料也就愈多，而且分類的網頁文件間將具有高主題相似性。而達到將主題或理論性相關程度較高的文件內容分類聚集之目的並可加強檢索的精確度。透過空間向量模式（space vector model ）計算查詢關鍵字與檢索結果網頁文件的相似度評估，可以發現利用連結結構分類出的資料是具有較高主題相似度的網頁文件，其效果可以接近甚至高於利用文字統計進行索引的查詢檢索系統。

七、參考文獻

[1] Peter G. Anick, Rex A.Flynn, ”Versioning a Full-text Information Retrieval System,” Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.98-111,June 1992.

[2] Budi Yuwono, Dik Lun Lee, ”WISE: a World Wide Web resource database system,” Knowledge and Data Engineering, IEEE Transactions on Vol.8 , No.4 , pp.548 -554,Aug. 1996.

[3] Jon M. Kleinberg,” Authoritative Source in a Hyperlinked Environment,” Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, pp.668–677,1998.

[4] Soumen Chakrabarti, Byron E. Dom, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, and Jon Kleinberg, “Mining the Web’s Link Structure,” Computer ,Vol.32, No.8 , pp.60–67, Aug. 1999 .

[5] Brin S. and Lawrence Page, “The anatomy of a large-scale hypertextual web search engine,” Proceedings of the Seventh International World Wide Web Conference , pp.107-117,1998.

[6] Jeromy Carriere and Rick Kazman,” Webquery:Searching and visualing the web through connectivity,”Proceedings of the Sixth International World Wide Web Conference(1), pp.701-711,1997.

[7] G.Golub,C.C.Van Loan,”Matrix Computations,” Johns Hopkins University Press,1989.

[8] Jon M.Kleinberg,Ravi Kumar,Prabhakar Raghavan,Sridhar Rajagopalan and Andrew S.Tomkins,”The Web as a graph: measurements , models and methods,” Proceeding of the Fifth International Computing and combinatorics Conference,1999.

[9] K. Bharat and A. Broder, " A technique for measuring the relative size and overlap of public web search engines," Proceeding of the Seventh World Wide Web Conference, pp.379-388, 1998.

[10] Monika R.Henzinger, “Hyperlink analysis for the web” IEEE Internet Computing ,Vol. 5 , No.1 , pp.45–50 , January/February 2001.

[11] Gerard Salton.”Automatic text processing : the transformation, analysis ,and retrieval of information by computer,”Addison-Wesley,1989.

[12] Chia-Hui Chang,Ching-Chi Hsu and Cheng-Lin Hou,“Exploiting hyperlinks for automatic information discovery on the WWW,” Tools with Artificial Intelligence, Proceedings. Tenth IEEE International Conference, pp. 156–163, 1998.

[13] Venkat N. Gudivada, Vijay V. Raghavan, William I. Grosky, and Rajesh Kasanagottu,” Information Retrieval on the World Wide Web,” IEEE Internet Computing, Vol. 1, No. 5, September/October 1997.

[14] Rodrigo A. Botafogo," Cluster analysis for hypertext systems," Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 116-125, 1993.
[15] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Namprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," Proceedings of the Seventh ACM Conference on Hypertext, Washington, USA ,1996.
[16] R.A. Botafogo and B. Shneiderman, "Identifying aggregates in hypertext structures." In Proc. ACM Hypertext '91, pp. 63-74, December 1991.
[17] Ronny Lempel and Shlomo Moran," The stochastic approach for link-structure analysis（SALSA）and the TKC effect,"In Proc. of the 9th International WWW Conferernce,2000.
[18] Gibson D. and J.M.Kleinerg , "Inferring Web Communities from Link Topology," in Proceedings of the 9th Conference on Hypertext and Hypermedia ,ACM Press,1998.
[19] Excite,http://www.excite.com.
[20] Altavisa, http://www.altavisa.com.
[21] Google, http://www.google.com.