天主教輔仁大學圖書資訊學系碩士班碩士論文

指導教授：陳舜德 博士

網路分類資源自動化擴展系統之研究

A Research into Automatic Expansion System of Classified Resources on Internet

研究生：詹姿潁 撰

中華民國 103 年 08 月

摘要

有別於現今網站利用人力資源將相關知識網站進行分類整理的作法，希望能提供一個協助平台來幫助使用者在資訊搜尋上能更快速正確；本研究將在現有的網際網路環境下，透過搜尋引擎結合自然語言處理、關鍵詞擷取、字詞間相似性的計算等相關技術，來建構一個由現有網站中所提供的相關連結資源來進行自動擴展服務，進而延伸搜尋到其他相關連結的機制；使得網站中所提供的參考性連結資源可定期自動的擴展與更新，期望藉由此網路資源自動擴展機制，協助使用者能夠更加容易地查找到自身所需的相關資訊並搜尋到其他眾多的相關網頁連結，而不僅止於網站內原先所建立的既定連結資源，這樣的機制將可協助使用者獲得更多樣、廣泛且高相關性的網頁資源做為其檢索資訊的參考性網站。

本研究期望透過開發之自動資源擴展推薦的方法，讓使用者在使用網際網路進行資訊檢索時，能提供一個更多元、多樣化的網頁資源擴展服務，使其可以更進一步的幫助讀者與潛在使用者，減少因為資訊素養與數位落差而造成的知識弱勢族群更難與其他人競爭的情形。

關鍵字：查詢擴展；搜尋引擎；數位落差；資訊檢索

Abstract

Unlike today's website will use human resources to manage which to do classified and sorted. Desirable to provide a platform to help users in searching for information on more quickly and correctly. In this work, under the current Internet environment, combined with Natural Language Processing (NLP), keyword retrievals, calculation of similarity between words and other related technologies through search engines, to construct a related Links resources provided by the existing website to the automatically expansion service, and then extended searching to other relevant link mechanism. Making the website provided referential link resources can automatic expansion and update regularly. Expected by this network resource automatically expansion mechanism, help users to more easily find relevant information for their own required and find many other related web links, and not only limited to within the site originally established link resources. Such mechanism would help users get more kinds, widely and high relevance web resources as reference website for information retrieval.

In this work, expectations of by developing automatically resource expansion of the recommended method. When users using the Internet to do information retrieval, to provide a more diversified and variety of the web resource extension service. It can still further help readers and potential users. Reducing knowledge disadvantaged groups harder to compete with others situations caused by Information Literacy and the Digital Divide.

Keywords: Query Expansion, Search Engine, Digital Divide, Information Retrieval

目次

圖表目錄

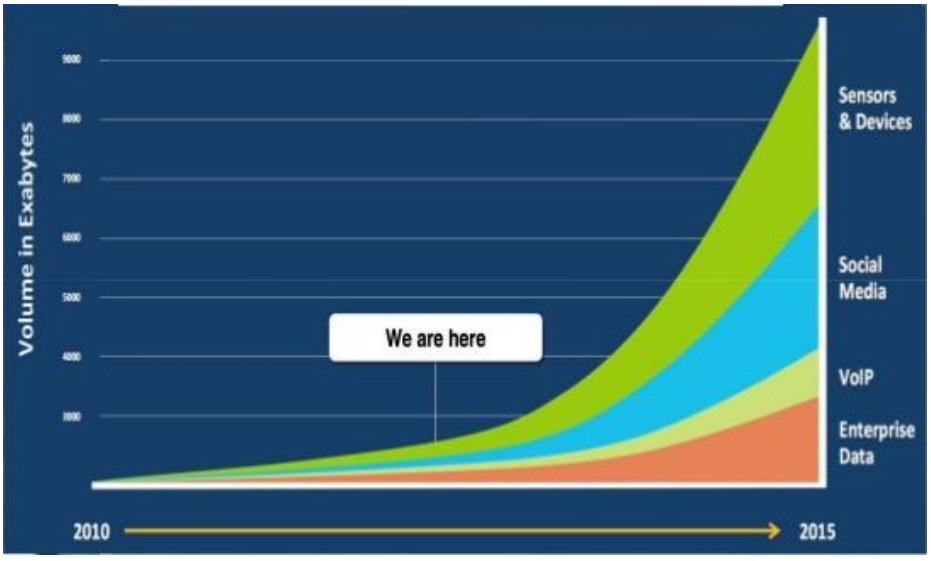
第一章　緒論

隨著網際網路和全球資訊網(World Wide Web，簡稱 Web)的快速發展，Web page 已然成為網際網路的重要應用方式，更是人們日常取得資訊管道的重要來源。網際網路所蘊含的資訊內容持續不斷地擴展與膨脹，我們要由這麼龐大的資訊儲存量中找尋到使用者所需要的資訊，是相當困難的工作；因此當網際網路成為人們獲取資料的重要管道時，如何有效率地進行處理、整合與再利用這些資料，就成為重要的議題。

第一節　研究背景與動機

隨著資訊科技的進步與網際網路的蓬勃發展，使得網路上網頁數量呈爆炸性的成長；使用者也習慣使用瀏覽器(Browser)來進行網路資訊的瀏覽，然而使用者想要在網路世界中快速的找到所想要的資訊，就必須透過搜尋引擎的力量才能達成 (吳典恩, 2007)。1998 年，NEC 的研究調查顯示，當時的全球資訊網已擁有三億兩千萬頁文件資料，相當於一百多萬本的紙本書籍，就好比是三個國家圖書館所容納的資訊總量 (王力行, 1998)，於 2011 年時，根據 IDC Digital Universe Study 的監測統計得知數位資料的數量更成長至 1.8 ZB，此相當於 18 億個 1TB 的行動硬碟之大小 (Adamov, 2012)，其發展的速度與涵蓋的資訊廣度已超過我們的想像，至 2012 年 IDC Digital Universe Study 更加指出於 2020 年時，全球將擁有 35ZB 的數位資料量，此統計數據對比 2011 年來說增長了近 20 倍 (John & David, 2012)，由下圖表 1 所示，我們可看出至 2015 年的資料量增長的速度如此之快；且其含括的資訊內容猶如大海之深且廣，就像一棟永遠閱覽不盡的知識殿堂。而資訊內容也從早期單純文字資訊演進至今含括了更多的圖片、聲音、影片...等多媒體資源，網路資源也越來越多樣與複雜。伴隨著網路資訊內容的不斷擴增，使用者從早期「資訊匱乏」漸漸變成「資訊過載」的現象。要從如此巨大的網路知識庫查找出

所需要的資料變得難上加難了；因此使用者對於資訊需求已非早期資訊匱乏時「量」的需求，而是逐漸轉變為對於檢索資訊「質」的要求。為協助使用者能快速查找、搜尋資訊，網路搜尋引擎(Search Engine)應運而生，使用者往往只需輸入所欲查找的關鍵字詞，搜尋引擎就能快速回饋使用者所查詢的資訊，使用者可以在短時間之內擷取到其「所需」的網路資源，所以目前使用者查詢資料已無法脫離使用搜尋引擎。



圖表 1：資料量的增長預測

資料來源: http://www.slideshare.net/clavenke/ss-23923075 (柯皓仁, 海量資料與圖書館, 2013)

網際網路自 80 年代開始快速發展迄今，存在於網路上的網頁資訊已難以計數，使用者面對如此龐大的網路資源已經很難快速而正確的搜尋到其所需之資訊。根據 Delphi Group 於 2004 年所針對知識工作者的調查發現，59%的使用者認為在網路搜尋環境中，取用資訊雖比以往更加快捷便利，但卻有 68%的使用者認為查詢資訊仍然是一個相當具困難度的行為，而其中 62%的使用者對於自身的檢索效率是不滿意的 (Delphi Group, 2004)。其主要原因在於，大多數的搜尋引擎僅提供關鍵字檢索功能，當使用者面對網路巨量的資訊，使用者輸入任一關鍵字，常常就會產生數以萬筆的查詢資料結果，面對搜尋引擎如此巨量的資訊反饋，如果無法有效地去進行資料的分類、篩選，使資料發揮其功用轉變為個人之知識與智慧，

即使擁有再多的資料也是無用。同時地，使用者在利用關鍵字檢索時，若不清楚應輸入哪些合適的檢索詞彙，檢索結果也不一定會是使用者所需要的資訊，因此如何有效的處理如此巨量的資料更形重要。

然而資訊科技的接觸機會與使用經驗更會因為使用者的性別、種族、階級、甚至是居住的地理區域等個人背景因素不同而有所差異 (曾淑芬, 2002)，也因為這樣的差異形成所謂的「數位落差」(Digital divide)，而「數位落差」的產生使得原本經濟貧困階層的民眾，更加難以取得與一般人對等的資訊，也相對造成這些弱勢族群在知識使用上「質」與「量」的相對剝奪感，在目前以知識經濟為導向的社會中，更讓這些社經地位處於弱勢的族群更難以翻身；這樣的發展造成另一種形式的人權剝奪，所以對於弱勢族群的「資訊人權」應該更加予以重視 (葉俊榮, 2006)。因此學校、公共圖書館或一些政府單位更應協助提供弱勢族群基本的資訊設備與網路環境，以降低「數位落差」所造成新的「知識階級」落差；藉由搜尋引擎(Search Engine)可以協助使用者與知識弱勢族群跨越資訊技術鴻溝，得以快速獲得其所需之資訊 (吳典恩, 2007)。而搜尋引擎的使用就是運用了資訊檢索(Information Retrieval，IR)的技術來達成幫助使用者快速查找所需的資訊，為此搜尋引擎提供了三種主要目的如下 (Gordon & Pathak, 1999)：

1. 將網際網路上各種領域的網頁資料當作一資料集合，讓使用者可以進行擷取資訊的動作。

2. 將上述所說資料集合中的各式網頁以一致的格式提供給使用者進行瀏覽。

3. 使用者進行查詢後所回饋的資料集合，搜尋引擎運用資訊擷取的方法將這些資料集合中與使用者所查詢之最相關資訊提供給使用者。

目前大多數搜尋引擎的做法，是將使用者輸入的關鍵詞彙透過全文檢索技術，搜

尋網際網路中相關網頁，並將搜尋到的網頁連結與摘要內容列表回傳，並未進一步針對文件內容做進階的處理與篩選，所以使用者仍須人工逐一過濾篩選相關訊息；但因回傳網頁連結頁面資訊相當繁多，「資訊過載」的現象更讓使用者在資訊搜尋耗費更多心力，還不確定是否能搜尋到使用者本身所需的資訊，所以在資訊檢索的過程中，使用者絕大部分的時間是用在瀏覽搜尋引擎回傳的檢索結果，而非一直在輸入關鍵字檢索 (卜小蝶, 2007)。因此，有一些網站將相關知識網站進行分類整理，希望能提供一個協助平台來幫助資訊能力較弱勢的族群在資訊搜尋上能更快速正確；然而透過人工整理畢竟耗時費力，而且網際網路上網站頁面的更新頻率與生命週期均不固定，造成耗費相當時日所整理的網站資源連結資訊，一段時日後就變成「老舊」，有些資訊甚至無法連結。本研究是在現有的網際網路環境下，透過搜尋引擎結合自然語言處理、關鍵詞擷取、字詞間相似性的計算等相關技術，來建構一個以現有網站中所提供的相關連結資源來進行自動擴展，進而延伸搜尋到其他相關連結的機制；藉由這樣的機制我們希望能建置一個知識入口網站，這樣的知識入口網站所提供的參考性連結資源可定期自動的擴展與更新，讓使用者能藉由此知識入口更加容易查找到所需的相關資訊。也期望藉由網路資源自動擴展機制，能協助使用者找尋到所需的多樣性網頁連結，而不僅止於網站內原先所建立的既定連結資源，這樣的機制對於因年齡因素與資訊素養較低的使用者，將可協助其獲得更多樣、廣泛且高相關性的網頁資源做為其檢索資訊的參考性網站。

本研究期望透過開發之自動資源擴展推薦的方法，讓使用者在使用網際網路進行資訊檢索時，能提供一個更多元、多樣化的網頁資源擴展服務，使其可以更進一步的幫助讀者與潛在使用者，減少因為資訊素養與數位落差而造成的知識弱勢族群更難與其他人競爭的情形。從圖書館的角度更應讓讀者更快速便捷的查找到所需之資訊，落實圖書館服務功能之普及。

第二節　　研究目的與問題

在網際網路中一些過時的新聞報導、過舊的資訊內容等不穩定的因素有著各種不同的特徵值或關鍵字，所以查找結果的回傳值有可能是一些過時甚至是不存在的文件資訊，這些條件造成了資訊需求與搜尋結果間的差距，此種問題至今仍是資訊檢索(Information Retrieval，IR)時所需克服的一大議題。本研究主要的目的在於探討網際網路服務所提供的各式網站中所提供的網頁資源擴展之服務的便利性與易達性，希望透過此研究，彌補網站在相關網頁資源服務的提供上所可能造成的資源不足與狹隘性之缺陷，進而使多樣化的網路平台資源可以發揮最大之效能，使提供使用者相關網頁需求之服務，可透過本研究所開發之自動化擴展推薦系統，進行可用性高的網頁資源服務平台的篩選與擴展之推薦，使讀者與使用者可以有效地獲得其他相關的網頁資源服務。方法建立後，本研究最後將進行實驗，評估字詞擴展的效果。更期望本研究之貢獻可以做為各式網站中所提供的相關網頁資源服務之推薦或是各式服務管道上的相關網路資源之平台擴展推薦的依據。

本研究所提出的網頁資源擴展之方法，主要希望達到以下幾點目的：

1. 線上即時更新的功能。本研究的進行以入口網站為主要設計之架構，希望以知識入口網站為平台並透過本研究提供的處理機制來調整關鍵字詞間的關聯性，每當目標網頁中有新的關鍵詞產生時，即可以以此作為擴展之依據，並確保其即時性。但隨著時間的變化，有些關鍵字會變得不合時宜，所以此處我們結合 Google 的線上搜尋引擎機制，去代替傳統人工的文件收集過程，使系統的參考資料來源可以維持其時效性。

2. 希望透過此研究，彌補網站在相關網頁資源服務的提供上所可能造成的資源不足與狹隘性之缺陷，進而使多樣化的網路平台資源可以發揮最大之效能。

針對本研究之研究目的，所設計的研究問題如下：

1. 是否減少知識弱勢族群查找知識資源時的不易性？
2. 資訊查找的便利性是否增加？
3. 資訊查找的正確媒合性是否增加？
4. 是否能如預期有助於檢索效益的提升？
5. 使用此網頁資源擴展之服務是否能夠得到較高的精確率？

本研究之研究流程首先說明研究之動機與目的，再者蒐集本研究所提到的相關研究主題進行研究探討，如入口網站、數位落差、資訊擷取、查詢擴展、特徵權重的計算方式等，並予以整合。透過描述使用之方法並結合相關工具予以進行系統實作實驗。最後針對研究的實作結果做出結論並對本研究未來研究方向提出建議。

第三節　研究範圍與限制

針對研究問題與研究目的所進行設計，本研究所提出的研究方法及實驗結果，並不一定適用於所有的領域以及其他相關查詢擴展之應用，研究範圍及限制說明如下：

1. 目前僅針對經整理具分類類別的網站連結資料進行實驗。由於網路資源的種類繁多，本研究無法全部蒐集之，故以小規模的資料集合代替，

並以實驗結果表示其有效性。對於大規模的資料集合而言，並不保證仍保有其實驗結果之效度及信度。

2. 限定於某一應用領域類別。要準確地進行網頁資源的擴展，首先必須分析其應用之領域，包括各種應用的知識與目標，由於過於廣泛與多樣，本研究無法全部一一分析之，故以一領域代替之。

第四節　名詞解釋

1. 數位落差(Digital Divide)

數位落差的概念隨著科技的日愈進步，不斷的被賦予新的內涵。從一開始只的是評估電腦設備的擁有率之高低差異，直至現今用以評估網際網路資源擁有與查找使用的不平衡現象 (葉俊榮, 2006)。

2. 搜尋引擎(Search Engine)

搜尋引擎指的是一種可以自動從網際網路搜集訊息，經整理後提供給使用者進行查詢的服務系統。網際網路上的訊息成千上萬筆，而且毫無關連性可言，所有的訊息像汪洋上的一個個小島，網頁連結是這些小島之間縱橫交錯的橋樑；而搜尋引擎利用網頁連結的方式為使用者繪製了一幅可一目了然的訊息地圖，提供使用者即查即用 (維基百科, 2009)。

3. 資訊檢索(Information Retrieval，IR)

「資訊檢索」係指搜尋資訊的科學，如在檔案中搜尋資訊、搜尋文件本身、搜尋描述檔案中資料的資料，或是在資料庫中進行搜尋，無論是何種獨立資料庫或是普遍被使用之網路資料庫皆可稱之為資訊檢索，IR 已成為一種不斷發展並和其他領域、技術不斷融合的學科 (Christopher, Prabhakar, & Hinrich, 2012)。

4. 查詢擴展(Query Expansion)

由於網頁資訊過多，使用者在透過搜尋引擎查找時所輸入的關鍵字，因為欠缺廣泛性的考量，造成使用者在查找知識時的不易性，因此無法確切獲得其需求性之資訊；查詢擴展對於資訊檢索領域來說，是一種可以幫助使用者在檢索過程中修正其查詢檢索詞彙並降低其資訊查找的困難度，查詢擴展即是一種用以輔助使用者進行查詢檢索，提升其查詢之效能的方法。

第二章　文獻探討

本研究主要的目的為希望藉由網站中所提供的網頁鏈結資訊，來延伸擴展相關的網路連結資源，可將這樣的技術運用在建置知識資源連結的入口網站，希望可以從現今已含有過多資訊的網際網路中，從已知的相關網頁連結資訊自動擴展延伸出相關的知識連結。透過此研究，能協助以往須由人力來整理相關網站連結資源，也可以彌補因人力不足無法及時提供充足的相關網頁鏈結資源所引起的可得性資源不足之缺陷，進而使多樣化的網路平台資源可以發揮最大之效能，意即在一個「知識入口」網站的介面下，可以提供使用者更加完整的相關性鏈結資訊，讓使用者可以確實地在一個網站中做到跨足網路資源的取得，透過本研究所開發之自動化擴展推薦系統，進行可用性高的網頁資源服務的篩選與擴展之推薦，滿足使用者的需求並提供一個正確的擴展機制，使讀者與使用者可以有效地獲得與之需求相關之網路資源。

本研究之文獻探討，由入口網站開始談起，接著探討數位落差、資訊擷取、查詢擴展，最後探討特徵權重的計算，希望透過描述使用之方法並結合相關之工具予以進行系統實作實驗，期望本研究之貢獻可以做為各式網站中所提供的相關網頁資源服務之推薦或是各式服務管道上的相關網路資源之平台擴展推薦的依據。

第一節　入口網站

隨著網路的快速成長，網際網路縮短了人與人之間溝通的距離，也造就了地球村的概念。「搜尋引擎」使人們在彈指之間便可獲得所需資訊，是現代人查詢資訊時不可或缺的工具 (林仁奎, 2011)。就使用者而言，入口網站就是使用者瀏覽網際網路的進入點，該入口並提供搜尋引擎、電子郵件、資訊內容、聊天室、線上

購物等眾多的服務 (Clarke & Flaherty, 2003)，並協助使用者得到符合且適切的資訊內容 (謝水木、黃敬仁, 2008)。許多入口網站同時扮演著一個知識入口的角色 (Zahir, Dobing, & Hunter, 2002)，幫助使用者蒐集、整理相關的網站資源，並依網站本身資源性質加以分類，讓使用者可以透過入口網站快速地查找到所需的網站資源 (劉冠宏, 2009)。

因為使用者個人知識背景的不同，或伴隨著個人經驗、居住地區...等各種因素，而產生不逕相同的知識落差 (Eszter, 2002)，間接造成在搜尋引擎使用上的認知差異；一般使用者面對網路巨量的資訊，不一定了解應輸入甚麼檢索詞彙才能搜尋到所需的資訊結果，往往使用者輸入任一關鍵字，搜尋引擎就會回應數千甚至上萬筆的資料查詢結果，這樣的檢索結果讓使用者無力去確認回傳的訊息哪些是正確的。搜尋引擎可以讓使用者快速獲得如此多的資訊，但如此大量未經組織的資訊反而造成使用者的負擔。由上述我們可得知，僅依賴關鍵字的檢索功能，是無法滿足使用者的多元檢索需求，亦無法解決資訊過載問題。因此，我們希望能夠使用本研究所提供的機制來協助類似「國家圖書館的參考資源選介網」或是「台北市政府知識管理網」等提供使用者知識的入口網站，能快速自網際網路蒐集相關類別的知識網站連結，藉由這樣的「知識入口網站」幫助使用者進行查找檢索所需的資訊，減少其因知識落差所造成的查找不易性。

當入口網站成為使用者進入各式網路平台的常用首頁畫面時，入口網站另類的成為了「知識入口網站」，「知識入口網站」是一個能夠提供使用者可以快速獲取他們所需知識的網站，使用者也能在此空間上分享他們的知識與心得 (邱俊銘, 2010)。在我們的研究中，我們希望透過此類的知識入口網站，為使用者提供符合其資訊尋求之偏好的資訊，便可以在最快時間捕捉住使用者的目光，並且正確地滿足使用者之需求，並透過本文所提供的自動擴展機制，使「知識入口網站」所提供的資訊內容保持為最新的狀態，使得使用者在大部分的情況下都能藉由「知

識入口網站」找到他要的資訊。此種方式可供企業了解使用者的偏好，或是強化、改善入口網站的服務品質，並可期望縮短使用者認知的差距。

第二節　數位落差

數位落差是一種隨著資訊科技發展越蓬勃，對於資訊弱勢族群所形成的另一種社會問題，數位落差隨著資訊科技的變遷其內涵也持續性的轉變，初期數位落差的衡量是由判斷「資訊設備的有無」，指的是能夠擁有使用電腦及網路能力者與無法擁有使用電腦及網路能力者之間的差異性 (廖秀紋, 2007)，這樣的差異性可能對於弱勢者會產生資訊取得不易、教育機會少、收入偏低、工作機會少等層面的相對剝奪感，隨著數位科技的發展愈盛，數位落差指的已不僅只是會使用電腦與否如此簡單了，它已成為現在社會不平等階級的另一種現象。

在現今的資訊時代裡，資訊本身即為一種無形的資本，資訊的運用成為了一種必備的技能，是否擁有資訊能力成為了現今衡量能否獲益的重要因素 (Haywood, 1998)。1999 年美國商務部國家通信及資訊委員會(National Telecommunications and Information Administration，U.S. Department of Commerce，NTIA)指出「數位落差」是資訊擁有者與資訊匱乏者間的落差情形所造成的社會不公平現象 (葉俊榮, 2006)。世界經濟合作與開發組織(Organization For Economic Co-operation and Department，OECD) 於 2001 年的公開報告中道出「數位落差」是個人、家庭、企業和地理區域、經濟環境等因素，再加上網路活動的利用方式不同而呈現差距的現象 (OECD, 2001)；直至今日的知識經濟時代，誰能快速掌握資訊，誰就較具有競爭力。當有一群使用者在電腦及網路的使用機會及能力的取得優勢形成「資訊社會菁英階層」稱為「資訊富者(Information-rich)」；而有另一群受限於經濟能力、教育程度、居住環境更甚是性別差異，致使無法取得資訊的技術及知識，成為「資訊弱勢階層」即「資訊貧者

(Information-poor)」 (高永煌, 2010)，這兩群人隨著資訊科技的快速發展使得知識的差距持續性的擴張，原本經濟貧困(The Disadvantaged)的階層，更加無法容易的取得資訊技術及知識，造成知識弱勢的族群在資訊使用上的質與量相對困難(曾淑芬, 2002)，資訊貧者在資訊化的社會，無法享受到資訊科技對生活與工作所帶來的便利，更因為缺乏資訊科技的基本應用及使用能力，以致於失去了與他人公平競爭的機會，造成新的社會問題及經濟發展上的隱憂，而這種資訊貧富不均的現象就是目前所謂的「數位落差」。

綜合上述所說，數位落差的廣泛定義意指在一個資訊科技的發展過程中，因為各項因素而造成資源分配不均所產生的一種在人與人之間的數位知識落差的現象 (Koss, 2001)。

第三節　　資訊擷取

資訊擷取是一種從根據使用者查詢行為所找出的資料結果集合中進行擷取並給予使用者相關文件的一種技術。資訊擷取的定義為將使用者所感興趣的主題、需求使用文字表達後，進而找尋出合適的文件與相關結果，亦即希望在最低成本之下取得與使用者之需求最相關的資訊結果 (Klir & Yuan, 1995)。搜尋引擎即是資訊擷取(Information Retrieval)技術的另一種呈現，分別扮演著資訊搜集者與資訊提供者的角色，使用者使用文字輸入所需要的需求後，搜尋引擎搜集了網際網路中各處的相關資訊文件，將之進而彙整後提供給使用者一個解析後之結果 (吳典恩, 2007)，因此資訊擷取主要概分三種主要功能如下 (Gordon & Pathak, 1999)：

1. 將網際網路上各種領域的網頁資料當作一資料集合，讓使用者可以進行擷取資訊的動作。

2. 將上述所說資料集合中的各式網頁以一致的格式提供給使用者進行瀏覽。

3. 允許使用者進行查詢，並且運用資訊擷取的方法將這些資料集合中與使用者所查詢之最相關資訊提供給使用者。

在資訊擷取領域中存在著許多種資訊擷取的模式，以下章節將進階探討較傳統的兩種模式：布林模式(Boolean Model)、向量模式(Vector Model)來闡述資訊擷取中關聯法則的作法。

關聯法則是 1993 年由 Agrawal 等學者所提出之概念方法，主要目的是從大量資料中找出資料之間的關係。有時我們並不會知道資料庫中各種資料間的關聯，即使知道也是具有不確定性的，因此希望關聯分析所產生的規則具有可信度，如果某一集合符合一定的可信度，也必具有一定普遍性的規則。關聯法則即希望透過分析紀錄計算出資訊集合中的支持度(Support)與信賴度(Confidence)，衡量找出的規則是否具有意義。

一、 布林模式(Boolean Model)

布林模式的檢索方式是一傳統的檢索方法，此方式透過集合理論(Set Theory)與布林代數(Boolean Algebra)為基礎進行運算，是資訊擷取模式中最為簡潔的一種方法，此種模式的概念相當的直覺性並且能將查詢的結果已具有精準語意的布林表示式(Boolean Expression)表達之 (蔡育欽, 2005)。在此種模式中，只需判斷檢索字詞是否存在於文件資料中，此檢索結果的表示只有兩個值，為非 0 即 1 的二分法(Binary Decision)的方式，1 的話表示此檢索字詞有存在文件資料中，反之則無。即在進行布林模式檢索時，會針對每個相關資訊文件分為相關與不相關兩者狀態，對於查詢的條件是無法進行部分比對(Partial Match)或有相似之處的，因此造成檢索結果的不精確。

因此，此模式的主要優點在於它是所有資訊檢索模式中最為簡潔的方式，

但無法進行部分比對的方式也成為此模式最為不足的缺點。

二、 向量模式(Vector Model)

向量空間模型是由 Salton 所提出的一種為了彌補布林模式中二元權重的不足，而利用關鍵字進行資訊檢索時所產生的二維向量來表示檢索字詞與文件的關係，進而計算出布林模式中所無法辨別的相似程度之方法。1998 年向量空間模型經過些微幅度地修改後，而用於文件自動分類的向量空間模型則是將關鍵字所產生的二維向量分別用來表示類別與文件 (Salton, 1988)。藉由向量空間來計算相似度的方式被廣泛的運用在文件的分類、群聚與檢索等應用。常見的向量相似度公式有內積、Dice 係數、餘弦、Jaccard 係數等算法(Salton, G.、Buckley, C., 1988)。其中 Dice 係數與 Jaccard 係數比較偏向集合中的交集、聯集觀點，兩向量關鍵詞交集數越高，其相似度越高。又相同交集數之下，Dice 係數傾向於比 Jaccard 係數給出更高的相似度值。餘弦與內積是比較偏向量夾角觀點，兩向量夾角越小，其相似度就越高。

由於兩者皆表現成主題向量時，兩主題向量之相似度可是為兩者之匹配法，進而利用物件分群化後的向量相似度進行比較分析，物件分群化(clustering)是基於以相似度作為分群之依據，以資料探勘(data mining)為本，向量相似度為輔之方法。在相似度計算方面一般常見的計算方式整理為下圖表 2：

Similarity Measure Sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner Product	$\|X \cap Y\|$	$\sum_{i=1}^{t} x_i \times y_i$
Dice Coefficient	$2 \cdot \frac{X \cap Y}{\|X\| + \|Y\|}$	$\frac{2\sum_{i=1}^{1} x_i \times y_i}{\sum_{i=1}^{t} x^2{}_i + \sum_{i=1}^{t} y_i{}^2}$
Cosine Coefficient	$\frac{X \cap Y}{\|X\|^{\frac{1}{2}} + \|Y\|^{\frac{1}{2}}}$	$\frac{\sum_{i=1}^{1} x_i \times y_i}{\sqrt{\sum_{i=1}^{t} x^2{}_i \sum_{i=1}^{t} y^2{}_i}}$
Jaccard Coefficient	$\frac{X \cap Y}{\|X\| + \|Y\|}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x^2{}_i + \sum_{i=1}^{t} y^2{}_i - \sum_{i=1}^{t} x_i \cdot y_i}$

圖表 2：相似度計算公式(本文整理)

此向量模式有以下幾點優點，一為檢索字詞的計算權重方式改善了擷取效能；另一為能夠處理部分比對的相似處，使擷取到的文件更為接近查詢條件，而其主要之缺點則是在此向量空間中將所有字詞視為互相獨立(Mutually independent)，然後實際上必須考慮到檢索字詞的相依性(Dependency)，由於許多檢索詞彙擁有局部相依的特性，這使得應用向量模式於文件集合中會降低整體的擷取效能(Ricardo & Beithier, 2002)，此種自動分類方法有著以下的限制 (黃嘉宏, 2008)：

1. 不適合處理過長的文件

   由於此種方式須將查詢文件與被查詢文件轉換成「文件-特徵詞彙」與「類別-特徵詞彙」之向量空間模型。當文件的字數越長，所涵蓋的特徵詞彙會越多，所佔的向量空間也會更大。當向量空間越大時，系統在進行相似度計算所耗費的時間便會越多；為此利用向量空間模式進行文件自我分類時，多半會將資訊負載量低或者較無意義的虛詞過濾掉。

2. 無法辨別語彙間之關聯性。

由於此種方式於現今的網際科技中尚未達到可以自由判定語意的階段，因此在進行運算時，檢索字詞必須相符才會被系統判別為關聯；當出現檢索詞彙上有關聯之語彙時，更是會因為字詞的不相符而影響相似度計算之結果。於本文中，我們使用中央研究院的斷詞系統配合關聯規則技術，來尋找檢索詞彙間的關聯加以利用來提升向量空間模式的文件群集計算 (黃嘉宏, 2008)。

因此，本研究應用向量模式的檢索模型的作法，分別使用以下步驟列表示之：

1. 將查詢關鍵字詞及擴展後文件轉換成維度(Dimension)相同的向量表示法。

   q=$[w_{1,Q}, w_{2,Q}, \dots\dots w_{n,Q}]$，代表查詢關鍵詞向量

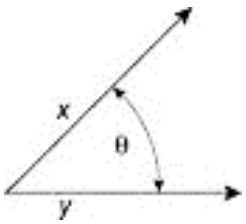
   d_j=$[w_{1,d_j}, w_{2,d_j}, \dots\dots w_{n,d_j}]$，代表擴展後文件向量

   將查詢關鍵詞與擴展後文件進行向量之轉換，接著使用量化的方式進行處理與分析，並可計算其兩者間的相似程度，本研究中使用 COSINE 餘弦作為運算之法則。

2. 使用 COSINE 餘弦作為計算方式，進而計算兩向量之夾角值，其值介於 0 至 1 之間。公式如下：

   Similarity Measure Sim(X,Y)

   = Cosine Coefficient $= \frac{\sum_{i=1}^{1} x_i \times y_i}{\sqrt{\sum_{i=1}^{t} x^2{}_i \sum_{i=1}^{t} y^2{}_i}} = \frac{\mathrm{X} \cdot \mathrm{Y}}{|X| \times |Y|}$



圖表 3：向量模式示意圖。資料來源： (蔡育欽, 2005)

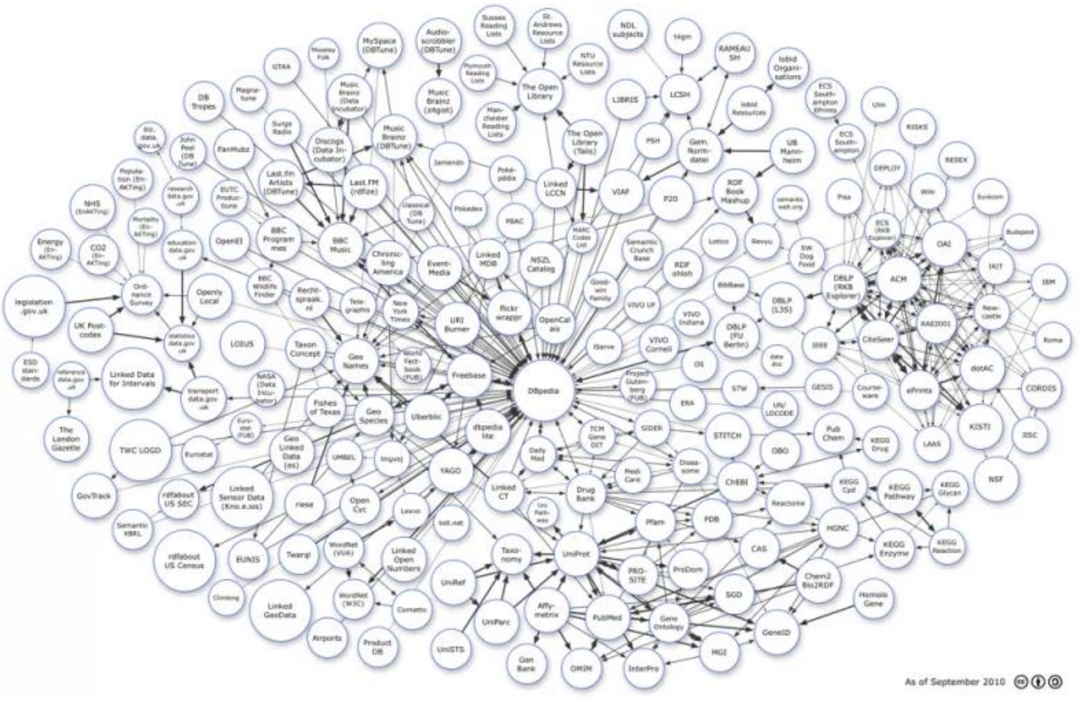
藉由運算後所得出代表相似度之值，當其夾角值愈小直至趨近於 0 時，其 Cosine 值會趨近於 1，此時相關度最高，而當兩向量夾角趨近於 90 度垂直時，其 Cosine 值趨近於 0，相關度亦即為 0 最小；最後再將所有文件以相似度加以排序整理，得到一相似程度之排名。

第四節　　查詢擴展

使用者在網際網路中搜找資料時，使用適當的查詢關鍵詞彙表達自身的資訊需求，對於一般的使用者來說是不容易的，隨著各自的知識背景不同，每一位使用者或知識專家對於相同主題的描述可能會使用不同的檢索詞彙，因此使用者會不斷的嘗試修正其所下的查詢關鍵詞，希望能查找到其真正想要的檢索資料。「查詢擴展」即是一種由同義詞、查詢語意相近語彙來擴展使用者所輸入之查詢關鍵詞上的缺漏，使得檢索成效可以提升的方法。

查詢擴展的方式常出現在資訊檢索系統中，此方法是一種可以幫助檢索者在檢索搜找的過程中修正其所下的檢索關鍵詞彙，並可以降低檢索者在檢索的過程中所面臨的繁雜的步驟 (蔡育欽, 2005)。而產生擴充詞的方式在查詢擴展中又可以分為「人工」與「自動」兩種，舉例來說，「人工」即為傳統圖書館中索引典的方式，舉例來說，國家圖書館的網站中所提供的參考資源連結即屬於此例，此種方式，有時會過於依賴個人認知；「自動」則是可以依據檢索結果或是相關文件擷取出相關詞彙來作為擴充詞，借以此法所自動化產生的擴充詞彙，可以減去傳統人工方式所產生的垢病。此方法可以運用所有得到的資料文件進行建立一索引典之架構以定義出詞與詞之間的關係，並可以找出與檢索詞相關的字詞作為擴充詞彙以加入初始查詢中。本研究使用的方式即建立一關聯詞庫，以該關聯詞庫所提供的詞彙作為擴充詞的來源，以達到查詢擴展的目的。

而 WEB 之父 Tim Berners-Lee 提出的語意網(Semantic Web)之概念(Berners-Lee, 2000)即為全球資訊網經擴展後的結果，為了具體的實現文件間的擴展性，Tim Berners-Lee 於 2006 年提出了所謂的「鏈結資料」(Linked Data)的概念，企圖賦予全球資訊網結構化的資料，使得這些資料內容可以在網際網路上利用鏈結資料進行分類並自動彙整出同一關鍵詞彙的相關資訊 (柯皓仁, 2013)。當越來越多的組織和個人採用鏈結資料的原則將資訊內容發佈到網際網路中時，這些彼此串聯的資料將構築出一個全域的資料空間，下圖表 4 即 W3C Linking Open Data(LOD)計畫所繪製的鏈結資料雲圖之範例。而在本文中的範例經過自動的斷字、斷詞與產生關聯詞後，將關鍵詞的詞彙拿來作為擴充詞彙後所致使的假定預測結果將與之相同。



圖表 4：Linking Open Data(LOD) Cloud (2011/8)

資料來源：http://www4.wiwiss.fu-berlin.de/lodclouds/state

第五節　　特徵權重計算

利用資訊擷取的方法擷取出頁面內容後，亦須利用特徵擷取的方法擷取出頁面內容的文章特徵，並根據該特徵於所屬文件的重要性給予一權重值 (黃嘉宏, 2008)，意即為將具有意義的文章詞彙做一個機率與統計的計算，並去除一些具有意義但是不足以代表文章資訊的內容詞彙，進而形成一個維度較小較容易計算的新特徵空間 (任炳魁, 2008)。特徵選取的目的在於從文字資料中擷取出具有語意的詞彙，以作為分類方法從而建立與頁面內容相關的特徵集合 (陳信源、葉鎮源、林昕潔、黃明居、柯浩仁、楊維邦, 2009)，此種特徵選取的作用在於過濾一些與本文所分析之問題無關、具有偏差或是擁有重複性的資料維度，更甚是對分類類別具有較小的影響力的詞彙或會干擾分類類別的雜訊詞，透過此方法可以降低特徵詞的數量與整個查詢擴展所需花費的時間，並可提升系統的效率及可獲取知識的準確性。因此，挑選出適當的特徵詞並進行特徵權重的計算，於本文中進行資料分析時扮演著非常重要之角色。

在特徵權重的計算部分，常用的特徵挑選方法包含了 TF (Term Frequency，TF)、DF(Document Frequency，DF) 或是 TF-IDF(Term Frequency Inverse Document Frequency，TF-IDF)等方法。TF 指的是特徵詞於文件中出現的頻率，DF 指的是特徵詞出現的文件數量之頻率，上述兩者常作為權重分析的重要考量，而最常使用的方法為 1983 年由 G. Salton 所提出的 TF-IDF 公式；TF-IDF 是一種統計方法，用以評估一特徵字詞於一個文件集合的其中之文件的重要程度，特徵字詞的重要性隨著它在文件中出現的次數成正比增加，進而比較決定特徵與類別間的關係為何 (任炳魁, 2008)，如某一特定文件內的特徵詞彙的頻率愈高，加上其在整個文件集合中的文件頻率越低，則 TF-IDF 的權重值會愈高，因此，TF-IDF 常用於過濾常見之用語，從而保留具重要性之詞彙，其公式如下 (Salton, G. and McGill, M. J., 1983)：

$$w_{ij} = TF_{ij} \times IDF_j \quad \text{(公式 1)}$$

w_{ij}：代表文件 i 中出現詞彙 j 的權重

TF_{ij}：文件 i 中出現詞彙 j 的次數

1. TF(Term Frequency)：詞彙頻率，指的是某一特徵詞彙出現在一篇文件中的次數，當此特徵詞彙出現次數越多通常代表其重要性越高，愈能成為此文件的代表性特徵詞。對於在某一特定文件裡的特徵詞彙 t_i 來說，它的重要性可表示為以下公式：

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{k,j}} \quad \text{(公式 2)}$$

n_{ij} ：該特徵詞在文件d_j中的出現次數

$\sum_k n_{k,j}$：分母指的則是在文件d_j中所有特徵詞的出現次數之和

2. DF(Document Frequency)：文件頻率，指的是某一特徵詞彙在一文件內容集合中所出現的內容篇數，當此特徵詞彙所出現的篇數愈多，表示此特徵詞彙較不具差異性。
3. IDF(Inverse Document Frequency)：反向文件頻率，1988 年 Salton 等人認為除了詞彙出現在文件的頻率 TF 之外，詞彙出現在文件中的數量也應該是評估詞彙重要性的指標之一，現今已是一種普遍性衡量重要性之指標。針對某一特定特徵詞t_i的 IDF 之計算公式為以下：(Salton, G.、Buckley, C., 1988)

$$idf_i = \log(\frac{N}{df_i}) \text{ (公式 3)}$$

N：整個文件集合中的總文件數。

df_i：出現詞彙 j 的文件數量。

即 $idf_i = \frac{|D|}{|\{j:t_i \in d_j\}|}$ (維基百科, 2013)

$|D|$：指的是訓練語庫的內容文件總數。

$\left|\{j: t_i \in d_j\}\right|$：指的是有包含特徵詞$t_i$的文件數目（即$n_{ij} \neq 0$的文件數目）

如果該特徵詞彙不在訓練語庫中，就會導致被除數為零，因此一般情況下使用 $1+\left|\{j: t_i \in d_j\}\right|$。

在特徵詞彙選取上應注意以下幾項要點原則 (李伯毅, 2004)：

1. 應選取特徵詞彙的語意資訊涵蓋較多，且對文件的表達能力較強的語言單位作為特徵項。

2. 在實際應用上，常常被採用的字、詞與短語來作為特徵詞彙之項目。

第六節　　檢索的成效評估

在傳統的成效評估方法中，我們常使用查全率(recall ratio)與精確率(precision ratio)兩種評量指標作為檢視成效的方法，且在資訊擷取上常應用於數位圖書館或搜尋引擎中為確保其系統的檢索效率，作為一項成效驗證之用的評估準則 (范瓊文, 2005)。

查全率可以用來瞭解系統找回所有相關資料的能力，或是系統遺漏相關資料的情形，其又稱召回率、回收率；指的是在資訊檢索的過程中，查找出相關資料的筆數與應有的相關資料總筆數的比值，稱為查全率。例如：假設在某一次檢索需求中，相關文件的總篇數應該有 50 筆，而在系統回傳的 100 筆結果中，只有 20 筆相關文件，則此次檢索的查全率為 20/50=0.4 即 40% (曾元顯, 2012)。而精確率可以用來瞭解系統查找結果的正確性程度高低，或是鑑定系統過濾不必要資料的能力，又稱查準率、求準率等；在進行檢索的過程中，系統查找出相關資料的筆數與系統找出資料之總筆數的比值，稱為精確率。例如，假設在某一次檢索需求中，系統回傳 100 筆檢索結果，其中有 20 筆被判斷為相關，則此次查詢的查準率

為 0.2 即 20% (曾元顯, 2012)。

在衡量系統檢索成效時，查全率與精確率通常是一起並用的，單獨檢視其中一項指標可能會有所偏差，而 F-measure 方法(調和平均值)就是同時考量精確率與查全率也是最常用來作為評估檢索品質好壞的評量準則，針對檢索文件結果可被劃分如下表所列的四種情況之一 (曾元顯, 2002)。兩者間，當系統儘可能將找到的所有相關資訊給予檢索者時，查全率會越高，但若查找結果包含不相關的資訊越多時，其精確率則會下降。

	系統分為該類 (相關)	系統不分為該類 (不相關)	總和
屬於該類別 (被檢索出)	a	b	a + b
不屬於該類別 (未被檢索出)	c	d	c + d
總和	a + c	b +d	a+b+c+d

圖表 5：文件數量分佈

依上列檢索結果分類表，可列出精確率(P，precision)與查全率(R，recall)的計算公式如下：

精確率(P)　　$\text{Precision} = \frac{a}{a+b} = \frac{\text{檢索所得具相關之筆數}}{\text{所有筆數}}$　　(公式 4)

查全率(R)　　$\text{Recall} = \frac{a}{a+c} = \frac{\text{檢索所得具相關之筆數}}{\text{所有具相關之筆數}}$　　(公式 5)

公式中符號 a,b,c,d 分別表示意義如下：

a：表示相關文件被檢索出之筆數

b：表示不相關文件被檢索出的筆數

c：表示未被檢索出的具相關之文件筆數

d：表示未被檢索出的不具相關之文件筆數

而對於評估檢索結果成效，我們通常會同時考量精確率與查全率，常使用的法則為 F-measure(調和平均值)，其公式如下所示：

$$\mathrm{F} = \frac{2PR}{P+R}，0 \leq \mathrm{F} \leq 1 \quad (公式\ 6)$$

當 F 值愈趨近於 1 時，表示其檢索成效愈佳，反之則愈劣。

第三章　研究方法與設計

本研究希望能夠協助使用者在處理單位機構網頁所提供的分類項目之鏈結時，能藉由本文提出的方法從網際網路蘊藏的網頁資源中，自動地將原先所建立的網路鏈結資訊與網路其他網頁透過相似度計算,能自動將相關的網頁連結彙整收錄，如此將可大幅減少由人力過濾篩選相關網站資源來提供使用者查找的負擔。本研究以網頁中所提供的分類連結作為擴展之基準，並以線上搜尋引擎作為資料收集介面來幫助使用者去進一步發掘具相關度的資訊，進而降低其因資訊超載所造成的資訊查找困難度以利其學習。

第一節　研究方法

本研究將使用「實驗研究法」(Experimental Method)進行，藉由實驗的程序來發現各實驗階段的因果關係或比較各個變項(Variables)之結果,並進行多次的實驗觀察其結果，尋求現象中的因果關係。在科學研究上,「實驗」一詞指的是造成某一變項或控制某一種情況，以觀察兩個以上變數之共變的過程;「實驗法」則是指藉實驗的程序以發現因果關係或比較各種變遷之結果的方法 (莊大衛, 2005)。

實驗法主要是在觀察探討自變項(Independent Variables)與依變項間的因果關係，為了解兩者變項間的因果關係，因此本研究在實驗之前，先將分類系統的訓練資料分為實驗組與對照組分別作為自變項與依變項之實驗結果，使這兩組訓練資料的各種條件相等，然後針對可能會受影響的依變項做出適當的控制後，對實驗組進行實驗處理(Experimental Treatment)，對照組則為不進行實驗處理之結果，最後比較實驗組與對照組在依變項統一控制的實驗結果下是否具有差異性，即可以得知兩者間的關係與結論。

本研究實驗的素材取自國家圖書館知識資源參考服務中的網路資源選介所提

供的鏈結資料(參見圖表 6)。針對所選取實驗之原始網站鏈結資料，擷取其網站名稱標題、網站內容描述以及其鏈結內容等欄位內容作為測試欄位資料(參見圖表 7)。實驗中我們假設各分類中的網站資料對於所蒐集的分類主題是具代表性，所以我們從類別主題所提供的連結中選取 10 筆鏈結網站資料作為測試文件訓練語料庫之用。

圖表 6：國家圖書館知識資源參考服務「網路資源選介」

圖表 7：擷取網站名稱標題、網站內容描述以及其鏈結內容等欄位內容作為測試欄位資料

第二節　　實驗設計與流程

本研究所提出的文件自我擴展演算法，是可以針對每一種主題類別的連結資訊，從現有的訓練文件(網頁連結資訊)中擷取其網站內容描述欄位經過中文斷詞器(中央研究院詞庫小組)，進行斷詞與詞性標記的處理，實驗中我們忽略單字詞，只考慮長度在兩個字元以上的詞彙，再統計詞彙出現的頻率與詞性標記特徵，藉此擷取出可代表該網站連結摘要內容的關鍵詞集，有關此內容描述資料(包含網站名稱標題、網站內容描述)斷詞前後之全文統計資料如圖表 8 所示。藉由該主題所選取的 10 筆鏈結網站文件(依實驗假設為具代表性)，擷取出的關鍵詞的集合，作為該主題的特徵詞集，再將這些代表個別主題特徵詞集之關鍵詞，再反饋利用搜尋引擎進行擴展檢索，將搜尋引擎查詢後回傳的網站 URL 進行相似度統計，以作為本研究進行網路資源擴展的依據。

斷詞前總字數	單種分類最多字數	單種分類最少字數	斷詞後總關鍵詞個數	單篇網站名稱標題與網站內容敘述最多關鍵詞個數	單篇網站名稱標題與網站內容敘述最少關鍵詞個數
2808	176	19	295	46	4

圖表 8：資料斷詞前後之全文統計資料

一、 前置處理

前置處理包含了斷詞切字(Tokenization)，詞性標記(POS Tagging)及刪除停用字[1](Stop-word Removal)等步驟。分析十篇頁面內容敘述，並擷取所有可能的詞彙作為其分類學習的特徵。

斷字切詞的目的在於從文字中擷取出具有語意的特徵詞彙。本研究藉由中央研究院所研發的中文斷詞系統，將網站名稱與網頁內容敘述資料進行斷詞與詞性標記後，文件內容範例如圖表 9。

圖表 9：詞性分析後的文件內容使用範例

[1] 大多數的自然語言(natural language)，有一些字詞屬於功能性、連結性的；就像英文中的冠詞、介系詞、連接詞，這些詞性的字詞出現頻繁，但本身沒有甚麼意義，所以在自然語言處理時常會將這類字詞忽略不予處理，這類字詞通常稱為「停用詞」(Stop Words)。

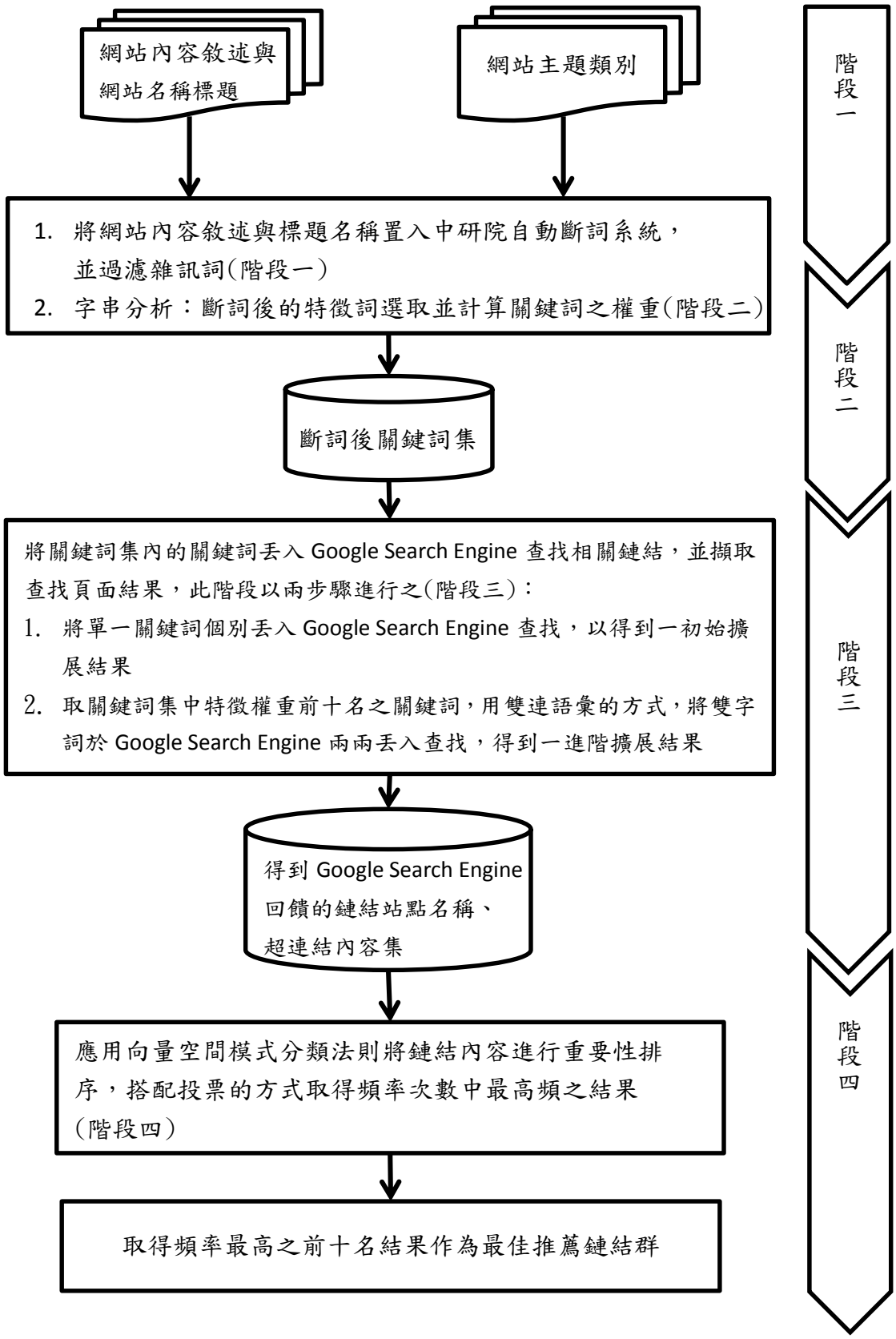
從前述斷詞與詞性標記中，將標點符號等不具有語意的符號進行刪除；在中文詞彙的使用經驗上，大部分的單字詞較不具有明顯之語意，因此本研究僅保留兩字元以上之多字詞語彙，而單字詞予以省略不處理，在處理上對於自然語言中「停用字」也進行過濾刪除，因為「停用字」出現頻率雖高但卻不具分類區別的價值 (陳信源、葉鎮源、林昕潔、黃明居、柯浩仁、楊維邦, 2009)，這類詞彙包含介系詞、指示代名詞、連詞、助詞等，這類高頻但不具語意上實質意涵的詞彙容易造成混淆擴展分類之效果，因此在本實驗中必須先行去除，以利實驗的進行。

二、 特徵詞彙選取

在對全部文件內容進行斷詞過後，因為過多的特徵詞彙容易耗費過多的計算時間與記憶體空間，因而導致系統效率不佳，所以此階段在不影響分類效果的條件下,利用特徵選取的方法可以過濾對分類具較低影響力的詞彙或干擾之雜訊詞，以降低特徵詞彙的數量，進而提升系統的效率。本研究特徵詞的權重計算方式採用 TF-IDF 方法，其公式請參見第二章公式 1 ～ 公式 3 之方法。

三、 資訊擷取與查詢擴展

本文採用向量空間模式(VSM，Vector Space Model)作為實驗的文件自動分類方法。實驗中我們利用現有網站分類「類別」訊息與其所蒐錄的網頁文件連結的「文件內容摘要」兩個特徵，其中「類別」向量與「文件內容」向量的特徵權重計算採用TF ×IDF法計算(參見第二章之公式 1)，並在取得特徵權重後進行正規化(參見第二章之公式 1～公式 3)，用以調整「類別」向量或「文件內容」向量之維度。相似度計算部分,則採用向量餘弦法則(參見第二章圖表 3)進行相似度之運算。詳細的實驗設計流程如下圖表 10 所示：



圖表 10：實驗架構流程圖

1. 使用中研院開發之中文斷詞器進行分類主題網頁集之特徵關鍵詞集擷取。

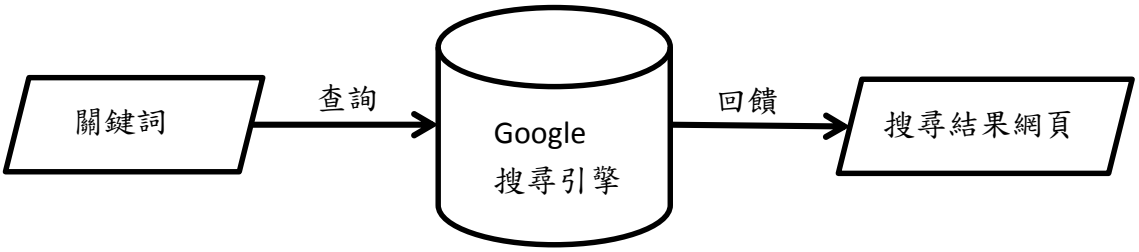
2. 導入擴展文件特徵之策略方式，進行自動分類：

   研究中設計了一擴展文件特徵之特徵擷取方法，針對此部分我們利用國家圖書館知識資源參考服務中的網路資源選介所提供的十筆鏈結資料之摘要內容斷詞後的關鍵字詞組作為搜尋引擎的查詢語句組，搜尋引擎會將與之相關的查詢結果作為回饋(Feedback)，因此我們將擷取搜尋引擎所回饋的網頁標題、超連結與部分描述結果來擴展文件特徵(參見圖表 11)。

圖表 11：Google 搜尋引擎之搜尋結果截圖

在實驗中我們藉由 Google 搜尋引擎來擴展相關網路資源，實驗步驟如下：

(1) 首先利用前階段斷詞與詞性標記篩選之關鍵詞集中的關鍵詞作為查詢語句，藉由會產生反饋的 Google 搜尋引擎進行檢索，截取搜尋引擎所回饋的搜尋結果網頁(參見圖表 12)



圖表 12：利用 Google 搜尋引擎進行擴展文件特徵之步驟示意圖

(2) 為了避免較不相關的查詢結果被採納入本研究之擴展範圍，實驗中針對每次檢索僅採計 Google 搜尋引擎回饋結果中關聯性較高的紀錄，此步驟分為兩階段進行之：

i. 階段一：進行雙字詞個別之詞彙擴展，在此僅採納回傳檢索結果頁面第一頁中前五筆之網頁連結資訊，得到一初始擴展結果集，如下圖表 13 所示之。

	KEYWORDN	NEWWEB	NEWWEBLINK
2	供應	共同供應契約 - 台灣銀行	www.bot.com.tw/procurement/procure_supply
3	供應	共同供應契約一覽表 - 台灣銀行	www.bot.com.tw/procurement/procure_supply/.../default.aspx
4	供應	供应链- 维基百科，自由的百科全书	zh.wikipedia.org/zh-tw/供应链
5	供應	供應商- MBA智庫百科	wiki.mbalib.com/zh-tw/供应商
6	供應	政府電子採購網- 行政院公共工程委員會	web.pcc.gov.tw/pishtml/pisindex.html
7	窗口	[高雄新興] 窗口土司| 宵夜| 熊寶寶的美食天地	gourmet.homeycat.tw/2014-02-13/窗口土司
8	窗口	窗口土司- 高雄市- 宵夜餐廳| Facebook	https://zh-tw.facebook.com/food296
9	窗口	深夜的小確幸- 窗口土司@ 吃心絕對- 高雄美食&高雄旅遊部落格:: 痞客	ksdelicacy.pixnet.net/blog/.../56040150-深夜的小確幸---窗口土司
10	窗口	窗口- 维基百科，自由的百科全书	zh.wikipedia.org/zh-tw/窗口
11	窗口	國內窗口- 外交部領事事務局全球資訊網	www.boca.gov.tw/lp.asp?ctNode=706&CtUnit=150...15...
12	全國	9937 - 全國 (臺灣證交所)	www.google.com.tw/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&
13	全國	ejob全國就業e網~安心找工作，免費找人才！	www.ejob.gov.tw/
14	全國	歡迎光臨全國電子	www.ejob.gov.tw/
15	全國	全國繳費網	www.elifemall.com.tw/
16	全國	全國加油站	https://ebill.ba.org.tw/
17	水利	經濟部水利署全球資訊網Water Resources Agency, Ministry of ...	www.wra.gov.tw/
18	水利	「水利」的新聞搜尋結果	自由時報電子報

圖表 13：初始擴展結果集

ii. 階段二：進行雙連語彙之擴展，此階段的關鍵詞集取其特徵權重前十名之關鍵詞進行查找，並採納檢索結果頁面第一頁中的前七筆之網頁連結資訊，得到一進階擴展結果集，如下圖 14 所示之；此部分的特稱權重計算將以 TF*IDF 為主要評斷之方法，加權後會得到一加權後關鍵詞集如下圖 15 所示之，

	A	B	C
1	KEYWORDNAME	NEWWEB	NEWWEBLINK
2	地形 河川	河川地形之歌- YouTube	www.youtube.com/watch?v=2f7RrA9XcEo
3	地形 河川	地形河川之歌- YouTube	www.youtube.com/watch?v=4HFrNYdcb4k
4	地形 河川	第八章河流地形 - Slideshare	www.slideshare.net/shiangyi/ss-205585
5	地形 河川	河流的歷史	gis.geo.ncu.edu.tw/earth/river/history1.html
6	地形 河川	高一河流地形	moss2007.shinmin.tc.edu.tw:8080/personal/h1210/.../高一河流地形.pd
7	地形 河川	吳舜惠- 河流地形	www5.hwsh.tc.edu.tw/web/wu550210/17
8	地形 河川	地形與河川	www.chukps.kh.edu.tw/elearning/social56_web/01_geo/2.htm
9	地形 河流	第八章河流地形 - Slideshare	www.slideshare.net/shiangyi/ss-205585
10	地形 河流	高一河流地形	moss2007.shinmin.tc.edu.tw:8080/personal/h1210/.../高一河流地形.pd
11	地形 河流	吳舜惠- 河流地形	www5.hwsh.tc.edu.tw/web/wu550210/17
12	地形 河流	「地形 河流」的圖片搜尋結果	gis.geo.ncu.edu.tw/earth/river/history1.html
13	地形 河流	河流的歷史	saturn.cksh.tp.edu.tw/~webadmin/site2/.../200901231100243.doc
14	地形 河流	第十一章河流地形、海岸地形	win2k.klgsh.kl.edu.tw/ftproot/socialweb/geography/.../heji.htm
15	地形 河流	河積地形	elearning.ice.ntnu.edu.tw/km/Data/Teacher/28519/Data/.../1.doc
16	地形 資料	基本地形圖資料庫分組 - 內政部國土測繪中心	bmap.nlsc.gov.tw/
17	地形 資料	最新消息	emap.nlsc.gov.tw/emap25/index.php?option=com...
18	地形 資料	數值地形模型	www.land.moi.gov.tw/law/new%5C349-N2.pdf
19	地形 資料	基本地形圖資料庫分組簡介	www.land.moi.gov.tw/law/new/191-N2.pdf

圖表 14：進階擴展結果集

	A	B	C	D	E	F
1	欄位2	次數(IN篇數間)	IDF反向文件頻率	次數(IN SET)	TF詞彙頻率(IN SET)	TF*IDF
2	水文	3	0.522878745	12	0.022099448	0.011555331
3	河川	5	0.301029996	18	0.033149171	0.009978895
4	河流	1	1	5	0.009208103	0.009208103
5	資料	3	0.522878745	9	0.016574586	0.008666499
6	地形	2	0.698970004	5	0.009208103	0.006436188
7	地理	2	0.698970004	5	0.009208103	0.006436188
8	基金會	1	1	3	0.005524862	0.005524862
9	研究	2	0.698970004	4	0.007366483	0.00514895
10	環境	2	0.698970004	4	0.007366483	0.00514895
11	資訊	4	0.397940009	7	0.012891344	0.005129982
12	說明	3	0.522878745	5	0.009208103	0.004814721
13	社會	2	0.698970004	3	0.005524862	0.003861713
14	台灣	3	0.522878745	4	0.007366483	0.003851777
15	中心	1	1	2	0.003683241	0.003683241
16	作用	1	1	2	0.003683241	0.003683241
17	保護	1	1	2	0.003683241	0.003683241
18	倉儲	1	1	2	0.003683241	0.003683241

圖表 15：加權後關鍵詞庫

於 Google 搜尋引擎回饋結果中，取其網頁標題以及搜尋結果之網站摘要描述進行關鍵字詞擴展特徵的處理。因此，此處文件自我擴展的作法，是針對每一種類別所進行擴展時利用 Google，擷取其回饋的現有網頁連結資訊、網站標題名稱與網站內容摘要組成新的訓練資料，以增加每一

類別的訓練文件數。

(3) 將關鍵字詞與其利用搜尋引擎擴展文件後得到的新特徵項目，作為一訓練文件或測試文件，使用向量空間模式分類方法作為分類依據，並使用投票理論賦予權重。

利用實驗設計所完成結果之資料集合，將鏈結內容進行多數決投票的方式計算，本文以「一人一票」的概念進行投票，即每筆資料結果皆具有相同價值之選票，此投票方式就是讓實驗結果的資料集合自行進行內部的投票，求其重要性排序，並將投票後前十名之結果作為最佳推薦鏈結群。

第三節　成效評估

因為網際網路中相關資訊連結相當眾多，並無法訂定出所欲擴展的網路資源連結集合，所以對於實驗成效評估我們基於認知心理學的方式，邀請具資訊素養且熟悉資訊檢索技能之研究生七人，並請七位研究生來評估實驗中所擴展的網路資源連結是否與原先類別所提供的連結資源是相關，評估過程如下：根據原資源網站某分類資源下隨機選取的 10 筆網站資料作為原始網站連結文本依據；其後透過本研究的系統流程所擷取出的關鍵詞組，利用 Google 搜尋引擎來進行檢索與向量空間相似度計算過程所擴展的網站連結資源為實驗之結果集。首先，將訓練語彙的測試資料集作為原始資料集，拿給研究生進行勾選，選出他或她所認為與主題類別所相同之網站文件項目有那些。接著再給研究生實驗結果所訓練出的擴展後實驗結果資料集，並告知其為根據十筆原始資料所擴展出來的結果，請其勾選與主題類別相關之網站項目。

正因為原始文件無庸置疑是最準確的，因此實驗對象看到訓練語彙用的測試資料集(即原始資料集)後，自然可以知道此網站文件是否為其所需，因此可以做為

測試資料集與擴展後的網站連結資源集(即擴展後實驗結果資料集)兩者間資料比較的依據之用。在傳統的成效評估方式中，本研究依循精確率與查全率的評估準則進行本節之檢索成效評估。

本研究希望透過粗略的初步實驗，可以進行實驗前後所產生的資料集間的分析與比較，並更深入的窺知其差異程度的大小，以及得知本研究之系統所提供的擴展後資訊是否可以作為替代原始資料之文件，作為這兩者間判斷之依據。

第四章　實驗與分析

依據第三章的研究方法與設計所述，本研究所進行的實驗共分為四個階段進行處理與分析，分別為網站相關連結資料前置處理、網站摘要特徵詞彙選取、資訊擷取與查詢擴展以及最終實驗結果評估與成效分析，本章就實際系統實作與評估部分，並對照第三章歸納整理的圖表 10 之實驗架構流程圖對於研究結果進行分析說明。

由實驗架構流程圖中，可以看出我們將實驗分為四個階段，依次為：

1. 網站摘要前置處理之詞彙分析，包含了斷詞切字、詞性標記與刪除停用字。
2. 特徵詞彙選取，依次為特徵詞彙選取與 TIF-IDF 特徵權重分析，用以降低特徵詞彙的數量。
3. 資訊擷取與查詢擴展，將數據利用各種向量相似度計算函式與分群演算法的方式組合分析之。
4. 實驗結果的評估與成效分析，在整個實驗系統進行相似度分析後，選擇出一個最佳分析結果進行人工式的評估，最後再利用問卷式評估結果的成效驗證。

第一節　前置處理之詞彙分析

本研究實驗的素材取自於國家圖書館知識資源參考服務中的網路資源選介所提供的鏈結資料，針對所選取的類別下之原始網站鏈結資料，截取其網站名稱標題、網站內容描述以及其鏈結內容等欄位內容做為測試欄位資料。

此處之類別我們選定自然科學中的「地球科學類」，類別主題中我們取其十篇作為測試文件訓練語料庫之用。此階段於網站文字內容摘要與網站名稱中取得具有語意的特徵詞彙，而因為過多的特徵詞彙容易導致系統的執行效率降低，因此

在不影響分類效果的條件之下，利用特徵選取的方法可以過濾對於分類具較低影響力的詞彙或干擾之雜訊詞，已降低特徵詞彙的數量，進而提升系統之效能。

首先我們將標點符號等不具有語意之符號進行刪除，並去除單字詞的使用，僅保留兩字元以上之詞彙，將其過濾後之詞彙整理出如圖 16 所示，可於表格的斷字內容列看出最後得到的關鍵詞彙。

網站名稱標題	網站內容描述	連結內容	斷字內容
河川入口網	本網站收錄臺灣地區河川水文之資訊，便利相關研究者參考點閱。內容上涵蓋有最新資訊、活動訊息、水文資訊（包括水情資訊、水質監測、洩洪資訊、即時雨量、淨水場水質、地下水位、淹水警示、水情水庫 SWE 資料等）、河川地圖、河川知識、河川法規、河川名詞等單元。此外，也提供河川、教育、親水、生態、水庫、單車等主題網供點選。	http://www.e-river.tw/index.aspx	水文(Na) 河川(Na) …… 水庫(Na) 水情(Na) 水質(Na)
水文地質資料庫	本網站收錄臺灣水文地質概況之相關資訊，同時為一個數位技術與知識管理的資料庫，讓使用者快速找到水文地質圖、水文地質資料等訊息。內容上分為兩部分：(1)岩心鑽探，包括有地理位置、樓層架號、井名井號、上架登錄時間、經緯度等資料，使用者可交叉比對，進而查詢到所需的水文地質資料；(2)研究報告，包括有研究報告查詢、岩心取樣研究。	http://hydro.moeacgs.gov.tw/	上架(VB) 水文(Na) 台灣(Nc) 交叉(VH) 地理(Na) 地質(Na) ……

圖表 16：實驗斷字內容部分結果

第二節　　特徵詞彙選取

一、　特徵詞選取策略

透過自動斷詞系統所標示出的含詞性標記之文件結果作為特徵詞選取之依據。斷詞處理後留下具有動詞與名詞屬性之詞彙，將其作為本實驗之特徵詞性選取標的。

二、 文件特徵權重計算

文件特徵即是在前置處理時做為後續資訊擷取與查詢擴展步驟處理之用。藉由前階段的特徵詞選取策略之後，對於實驗所需的文件資料中，抽取出具代表性之文件詞彙，並給予適當的過濾與頻率統計，再利用 TF-IDF 公式計算出各個詞彙於文件中之權重值，如下圖表 17 所示。

	A	B	C	D	E	F
1	欄位2	次數(IN篇數間)	IDF反向文件頻率	次數(IN SET)	TF詞彙頻率(IN SET)	TF*IDF
2	水文	3	0.522878745	12	0.022099448	0.011555331
3	河川	5	0.301029996	18	0.033149171	0.009978895
4	河流	1	1	5	0.009208103	0.009208103
5	資料	3	0.522878745	9	0.016574586	0.008666499
6	地形	2	0.698970004	5	0.009208103	0.006436188
7	地理	2	0.698970004	5	0.009208103	0.006436188
8	基金會	1	1	3	0.005524862	0.005524862
9	研究	2	0.698970004	4	0.007366483	0.00514895
10	環境	2	0.698970004	4	0.007366483	0.00514895
11	資訊	4	0.397940009	7	0.012891344	0.005129982
12	說明	3	0.522878745	5	0.009208103	0.004814721
13	社會	2	0.698970004	3	0.005524862	0.003861713
14	台灣	3	0.522878745	4	0.007366483	0.003851777
15	中心	1	1	2	0.003683241	0.003683241
16	作用	1	1	2	0.003683241	0.003683241
17	保護	1	1	2	0.003683241	0.003683241
18	倉儲	1	1	2	0.003683241	0.003683241

圖表 17：TF-IDF 權重計算範例

圖表 17 中所計算出的 TF-IDF 值表示了某關鍵詞在原始資料集中所屬的網站內容與標題中所出現的頻率，並可看出在整個原始資料集中此關鍵字出現的頻率高低；如在所屬的網站內容與標題中所出現的頻率愈高，而其在整個原始資料集中每個網站間的內容與標題所出現的頻率愈低，則 TF-IDF 的權重值會愈高；TF-IDF 的值愈高代表此關鍵詞彙在擴展文件階段時所代表的重要性與代表性愈高，因此，本研究取出權重值最高的前十名詞彙，以便在第一次擴展實驗後進行第二次更深入的資料擴展以得到相似程度更高的資料訓練結果集。

經過文件特徵權重計算之後，原應依向量空間模式把所有的文件進行相似度計算，再將所有計算後文件一矩陣排列後形成新的 VSM，本研究實驗中把方法做一個降階簡化的方式，使用投票的方式進行賦予權重後的相似度計算方法。

第三節　　資訊擷取與查詢擴展

本節主要分別討論兩階段的資訊擷取與查詢擴展方法，第一階段為資料集的雙字詞一一丟入搜尋引擎進行個別擴展查找，在前置分析時所得到之詞彙分析關鍵詞集於此處作為「關鍵詞彙集合」，參見圖表 18 所列出之部分雙字詞擴展用之關鍵詞集，即圖表中所表示之斷字內容，並在每筆字詞導入擴展策略後取得搜尋引擎所回饋的前五筆資料結果，以形成一初始擴展資料集。

1	樣本編號	網站名稱	網址	斷字內容
2	1	河川入口網	http://www.e-river.tw/index.aspx	收錄
3	1	河川入口網	http://www.e-river.tw/index.aspx	臺灣
4	1	河川入口網	http://www.e-river.tw/index.aspx	地區
5	1	河川入口網	http://www.e-river.tw/index.aspx	水文
6	1	河川入口網	http://www.e-river.tw/index.aspx	資訊
7	1	河川入口網	http://www.e-river.tw/index.aspx	便利
8	1	河川入口網	http://www.e-river.tw/index.aspx	相關
9	1	河川入口網	http://www.e-river.tw/index.aspx	研究者
10	1	河川入口網	http://www.e-river.tw/index.aspx	參考
11	1	河川入口網	http://www.e-river.tw/index.aspx	點閱

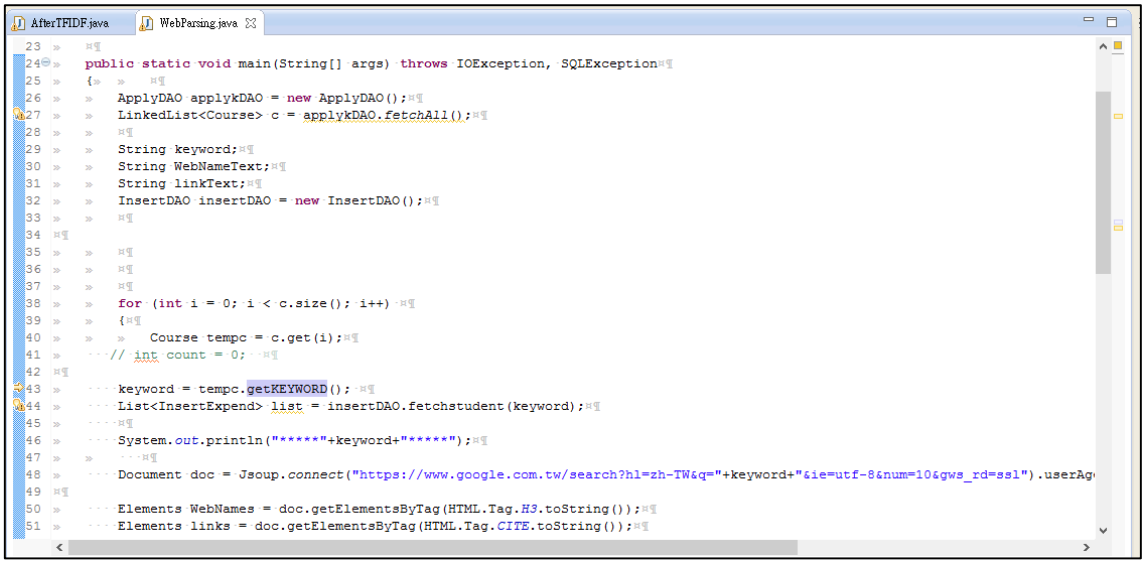
圖表 18：雙字詞個別擴展階段之部分關鍵詞集

第二階段為經由前述步驟的文件特徵計算所選取出代表文件之前十名詞彙後，圖表 19 所列之，將第一階段雙字詞的個別查找轉變為雙連語彙的兩兩詞組丟入搜尋，導入擴展文件特徵之策略方式後並於搜尋引擎所回饋的結果中，取其前七筆資料作為進階擴展結果集之內容。實驗過程中我們設計了一擴展文件特徵之特徵擷取方法，此部份我們藉由 Google 所提供的搜尋引擎進行檢索查詢，進而利用上述步驟所選取出代表文件之關鍵詞彙作為搜尋引擎的查詢詞組，並分別將階段一與階段二的關鍵詞彙集存入 Oracal 資料庫兩張資料表中得以與系統程式設計的部分搭配協調，使其可自動化地進行資料擷取與查詢擴展的查找行為，如下圖 20 所示之

部分系統程式設計與基本的參數設定，利用此種方式，搜尋引擎將會把與其相關的查詢結果作為回饋資料，因此在階段一與階段二我們將擷取搜尋引擎所回饋給我們的網頁標題、網站站址與部分描述結果分別取其前五筆及前七筆關聯性較高之資料項目作為擴展後文件之特徵，因而整理出如下圖 21 所示之合併兩階段所得到的擴展後資料集，本研究於此部分得到之資料集所擴展出之站點，利用 Excel 的分析工具計算出每一站點之根網址作為後續分析之用，參見圖 22 所示。

	keyword	次數(IN篇數間)	IDF反向文件頻率	次數(IN SET)	TF詞彙頻率(IN SET)	TF*IDF
2	水文	3	0.522878745	12	0.022099448	0.011555331
3	河川	5	0.301029996	18	0.033149171	0.009978895
4	河流	1	1	5	0.009208103	0.009208103
5	資料	3	0.522878745	9	0.016574586	0.008666499
6	地形	2	0.698970004	5	0.009208103	0.006436188
7	地理	2	0.698970004	5	0.009208103	0.006436188
8	基金會	1	1	3	0.005524862	0.005524862
9	研究	2	0.698970004	4	0.007366483	0.00514895
10	環境	2	0.698970004	4	0.007366483	0.00514895
11	資訊	4	0.397940009	7	0.012891344	0.005129982

圖表 19：雙字詞擴展階段之加權分析後前十名關鍵詞集(由高至低)



圖表 20：部分系統程式設計與基本參數設定

	KEYWORDNAME	NEWWEB	NEWWEBLINK
2	供應	共同供應契約 - 台灣銀行	www.bot.com.tw/procurement/procure_supply
3	供應	共同供應契約一覽表 - 台灣銀行	www.bot.com.tw/procurement/procure_supply/.../default.a
4	供應	供应链- 维基百科，自由的百科全书	zh.wikipedia.org/zh-tw/供应链
5	供應	供應商- MBA智庫百科	wiki.mbalib.com/zh-tw/供应商
6	供應	政府電子採購網- 行政院公共工程委員會	web.pcc.gov.tw/pishtml/pisindex.html
7	窗口	[高雄新興] 窗口土司! 窩夜! 熊寶寶的美食天地	gourmet.homeycat.tw/2014-02-13/窗口土司
8	窗口	窗口土司- 高雄市- 窩夜餐廳 Facebook	zh-tw.facebook.com/food296
9	窗口	深夜的小確幸- 窗口土司@ 吃心絕對- 高雄美食&高	ksdelicacy.pixnet.net/blog/.../56040150-深夜的小確幸---
10	窗口	窗口- 维基百科，自由的百科全书	zh.wikipedia.org/zh-tw/窗口
11	窗口	國內窗口- 外交部領事事務局全球資訊網	www.boca.gov.tw/lp.asp?ctNode=706&CtUnit=150...15...
12	全國	9937 - 全國 (臺灣證交所)	www.google.com.tw/url?sa=t&rct=j&q=&esrc=s&source=\
13	全國	ejob全國就業e網~安心找工作，免費找人才！	www.ejob.gov.tw/
14	全國	歡迎光臨全國電子	www.ejob.gov.tw/
15	全國	全國繳費網	www.elifemall.com.tw/
16	全國	全國加油站	ebill.ba.org.tw/
17	水利	經濟部水利署全球資訊網Water Resources Agency, M	www.wra.gov.tw/
18	水利	「水利」的新聞搜尋結果	www.google.com.tw/search?q=%E6%B0%B4%E5%88%A9&es
19	水利	水利及海洋工程學系 - 國立成功大學	www3.hyd.ncku.edu.tw/

圖表 21：階段一與階段二資料擷取後之部分資料集

NEWWEB	NEWWEBLINK	根網址至/的長度	
共同供應契約 - 台灣銀行	www.bot.com.tw/procurement/procure_supply	15	www.bot.com.tw/
共同供應契約一覽表 - 台灣銀行	www.bot.com.tw/procurement/procure_supply/.../default.a	15	www.bot.com.tw/
供应链- 维基百科，自由的百科全书	zh.wikipedia.org/zh-tw/供应链	17	zh.wikipedia.org/
供應商- MBA智庫百科	wiki.mbalib.com/zh-tw/供应商	16	wiki.mbalib.com/
政府電子採購網- 行政院公共工程委員會	web.pcc.gov.tw/pishtml/pisindex.html	15	web.pcc.gov.tw/
[高雄新興] 窗口土司! 窩夜! 熊寶寶的美食天地	gourmet.homeycat.tw/2014-02-13/窗口土司	20	gourmet.homeycat.tw/
窗口土司- 高雄市- 窩夜餐廳 Facebook	zh-tw.facebook.com/food296	19	zh-tw.facebook.com/
深夜的小確幸- 窗口土司@ 吃心絕對- 高雄美食&高	ksdelicacy.pixnet.net/blog/.../56040150-深夜的小確幸---	22	ksdelicacy.pixnet.net/
窗口- 维基百科，自由的百科全书	zh.wikipedia.org/zh-tw/窗口	17	zh.wikipedia.org/
國內窗口- 外交部領事事務局全球資訊網	www.boca.gov.tw/lp.asp?ctNode=706&CtUnit=150...15...	16	www.boca.gov.tw/
9937 - 全國 (臺灣證交所)	www.google.com.tw/url?sa=t&rct=j&q=&esrc=s&source=\	18	www.google.com.tw/
ejob全國就業e網~安心找工作，免費找人才！	www.ejob.gov.tw/	16	www.ejob.gov.tw/
歡迎光臨全國電子	www.ejob.gov.tw/	16	www.ejob.gov.tw/
全國繳費網	www.elifemall.com.tw/	21	www.elifemall.com.tw/
全國加油站	ebill.ba.org.tw/	16	ebill.ba.org.tw/
經濟部水利署全球資訊網Water Resources Agency, M	www.wra.gov.tw/	15	www.wra.gov.tw/
「水利」的新聞搜尋結果	www.google.com.tw/search?q=%E6%B0%B4%E5%88%A9&es	18	www.google.com.tw/
水利及海洋工程學系 - 國立成功大學	www3.hyd.ncku.edu.tw/	21	www3.hyd.ncku.edu.tw/

圖表 22：利用 Excel 計算根網址

此處文件自我擴展的做法為選定類別中現有訓練資料的關鍵詞集進行網路資源擴展，我們利用 Google 搜尋引擎輔佐之，擷取其回饋的現有網頁連結資訊、網站標題名稱與網站內容摘要組成新的訓練資料，以增加每一類別的訓練文件數，此處實驗原僅有十筆網站站址與 245 筆原始相關斷詞後資料，經由資訊擷取與查詢擴展後，所得到的新特徵項目組成新文件資料，於階段一時文件資料中已含有 1471 筆新資料集合，至階段二時又增加至 1724 筆訓練資料。

本研究在取得特徵權重後會先進行一正規化，用以調整新資料集合之重要程度或相似程度，此階段簡化向量相似度法則的繁雜程度，方法降階後，首先進行依其重要性程度由高至低做一排序，再以投票法則作為判斷其相似程度之方法，以「一人一票」的概念進行投票，即每筆資料結果具有相同價值之選票，使擴展後的資料結果集進行內部的投票，票數愈高者則其相似程度愈高。

第四節　　實驗結果

本節主要討論經上述步驟選取特徵詞彙與權重計算，再經由擴展實驗後所得到之新資料集合進行投票法的搭配計算，藉由運算的結果來分析及觀察實驗之成效。因實驗資料有 1724 筆，所以在進行內部投票前，本研究依資料所附載之網址取出根網址作為投票之根據，並在加入投票法後取前二十筆結果進行分析評估，投票結果如下圖 23 所示。

	根網址站點	投票次數	網站名稱
2	zh.wikipedia.org/	269	維基百科
3	www.google.com.tw/	103	GOOGLE
4	baike.baidu.com/	42	百度百科
5	e-river.wra.gov.tw/	28	經濟部水利署-E河川入口網
6	wiki.mbalib.com/	28	MBA智庫百科
7	www.wra.gov.tw/	27	經濟部水利署
8	www.wra04.gov.tw/	21	水利署第四河川局
9	zh-tw.facebook.com/	20	FACEBOOK臉書
10	law.moj.gov.tw/	17	全國法規資料庫
11	support.google.com/	17	GOOGLE說明
12	www.baike.com/	17	百度百科
13	www.iciba.com/	15	愛詞霸
14	wq.epa.gov.tw/	14	行政院環保署-全國環境水質監測資訊網
15	hydro.moeacgs.gov.tw/	13	水文地質資料庫
16	www.cwb.gov.tw/	13	交通部中央氣象局
17	www.sinica.edu.tw/	13	中央研究院
18	tw.news.yahoo.com/	12	YAHOO新聞
19	www.comc.ncku.edu.tw/	12	國立成功大學近海水文中心
20	gis.geo.ncu.edu.tw/	11	國立中央大學應用地質研究室-工程地質與新科技研究室
21	www.bot.com.tw/	11	台灣銀行

圖表 23：加入投票法實驗後之前二十筆資料部分結果

由圖表進行觀測時，可看出部分結果會因搜尋引擎中多數使用者所查找的使用方式而影響查找結果，因此本階段去除百科類網站、搜尋引擎式入口網站與社群網站等類型之站點資訊，得到新的前二十筆相似資料，如下圖 24 所示。

	根網址站點	投票次數	網站名稱
2	e-river.wra.gov.tw/	28	經濟部水利署-E河川入口網
3	www.wra.gov.tw/	27	經濟部水利署
4	www.wra04.gov.tw/	21	水利署第四河川局
5	law.moj.gov.tw/	17	全國法規資料庫
6	wq.epa.gov.tw/	14	行政院環保署-全國環境水質監測資訊網
7	hydro.moeacgs.gov.tw/	13	水文地質資料庫
8	www.cwb.gov.tw/	13	交通部中央氣象局
9	www.sinica.edu.tw/	13	中央研究院
10	www.comc.ncku.edu.tw/	12	國立成功大學近海水文中心
11	gis.geo.ncu.edu.tw/	11	國立中央大學應用地質研究室-工程地質與新科技研究室
12	www.bot.com.tw/	11	台灣銀行
13	rdc28.cwb.gov.tw/	10	颱風資料庫
14	www.edu.tw/	10	教育部全球資訊網
15	dbi.lib.ntu.edu.tw/	9	臺灣大學圖書館電子資料庫檢索
16	www.airitiart.com/	9	華藝世界美術資料庫
17	www.moeacgs.gov.tw/	9	中央地質調查所全球資訊網
18	lui.nlsc.gov.tw/	8	國土利用調查成果資訊網

圖表 24：去除不相關站點後之部分資料

由於網路相關資源繁多，本研究進行擴展之結果並無標準答案來驗證，所以本文藉由邀請具資訊素養的測試者來協助評估實驗中所獲資源擴展之成效，因此針對原有網站連結與實驗所擴展結果列出圖表 25、26；讓測試者對於所列網站關聯性進行評估，來驗證實驗的成效。

來源網站連結相關性認定			
	根網址	網站名稱	請勾選
1	http://www.hydroinfo.gov.cn/	中國水文訊息網	□相關 □不相關
2	http://hydro.moeacgs.gov.tw/	水文地質資料庫	□相關 □不相關
3	http://gweb.wra.gov.tw/wrhygis/	水文資料網路查詢系統	□相關 □不相關
4	http://proj1.sinica.edu.tw/~vid	台灣社會人文電子影音數位博	□相關 □不相關

	eo/main/water/index.html	物館--主題館	
5	http://gic.wra.gov.tw/gic/HomePage/Index.aspx#	地理資訊倉儲中心	□相關 □不相關
6	http://www.yucc.org.tw/	余紀忠文教基金會	□相關 □不相關
7	https://isp.moe.edu.tw/	我愛河川	□相關 □不相關
8	http://www.e-river.tw/index.aspx	河川入口網	□相關 □不相關
9	http://gis.geo.ncu.edu.tw/earth/river/history.htm	河流	□相關 □不相關
10	http://www.gst.org.tw/	社團法人中華民國地質學會	□相關 □不相關

圖表 25：來源網站連結相關性認定

網路資源擴展結果相關性認定			
	根網址	網站名稱	請勾選
1	e-river.wra.gov.tw/	經濟部水利署-E 河川入口網	□相關 □不相關
2	http://www.wra.gov.tw/	經濟部水利署	□相關 □不相關
3	http://www.wra04.gov.tw/	水利署第四河川局	□相關 □不相關
4	law.moj.gov.tw/	全國法規資料庫	□相關 □不相關
5	wq.epa.gov.tw/	行政院環保署-全國環境水質監測資訊網	□相關 □不相關
6	hydro.moeacgs.gov.tw/	水文地質資料庫	□相關 □不相關
7	http://www.cwb.gov.tw/	交通部中央氣象局	□相關 □不相關
8	http://www.sinica.edu.tw/	中央研究院	□相關 □不相關
9	http://www.comc.ncku.edu.tw/	國立成功大學近海水文中心	□相關 □不相關
10	gis.geo.ncu.edu.tw/	國立中央大學應用地質研究室-工程地質與新科技研究室	□相關 □不相關
11	http://www.bot.com.tw/	台灣銀行	□相關 □不相關
12	rdc28.cwb.gov.tw/	颱風資料庫	□相關 □不相關
13	http://www.edu.tw/	教育部全球資訊網	□相關 □不相關
14	dbi.lib.ntu.edu.tw/libraryList2/	臺灣大學圖書館電子資料庫檢索	□相關 □不相關
15	http://www.airitiart.com/	華藝世界美術資料庫	□相關 □不相關
16	http://www.moeacgs.gov.tw/	中央地質調查所全球資訊網	□相關 □不相關

17	lui.nlsc.gov.tw/	國土利用調查成果資訊網	☐相關 ☐不相關
18	http://www.geo.ntnu.edu.tw/	國立臺灣師範大學地理學系	☐相關 ☐不相關
19	bmap.nlsc.gov.tw/	基本地形圖資料庫分組 - 內政部國土測繪中心	☐相關 ☐不相關
20	e-info.org.tw/	環境資訊中心-台灣環境資訊協會	☐相關 ☐不相關

圖表 26：網路資源擴展結果相關性認定

第五節　　實驗評估

本研究的目的在於協助一般使用者就現有網路相關連結資訊進行擴展，實驗的結果並無標準答案來進行驗證，所以我們藉由讓使用者來評估實驗中所獲擴展資源之相關程度，作為實驗的成效評估，我們邀請具基本資訊素養的七位研究生(U_1~U_7)，針對研究中來源網站相關連結與源此擴展的網站資源實驗結果進行認定，其結果如圖表 27 與圖表 28 所示之：

測試者 / 項目編號	來源網站連結相關性認定								
	U1	U2	U3	U4	U5	U6	U7	不相關總計	相關度
來源網站連結_1	✓	✓	✓	×	✓	✓	✓	1	86%
來源網站連結_2	✓	✓	✓	×	✓	✓	✓	1	86%
來源網站連結_3	✓	✓	✓	✓	✓	✓	✓	0	100%
來源網站連結_4	✓	✓	✓	✓	✓	✓	×	1	86%
來源網站連結_5	✓	✓	✓	×	✓	✓	✓	1	86%
來源網站連結_6	✓	✓	✓	×	×	✓	✓	2	71%
來源網站連結_7	✓	×	×	×	×	✓	✓	4	43%
來源網站連結_8	✓	✓	×	×	×	✓	✓	3	57%
來源網站連結_9	✓	✓	✓	×	×	✓	✓	2	71%
來源網站連結_10	✓	✓	✓	×	✓	✓	✓	1	86%
相關	10	9	8	2	6	10	9	總相關度(百分比)	77.14%
不相關	0	1	2	8	4	0	1	不相關度總計(百分比)	22.86%

圖表 27：研究實驗之來源網站統計資料

測試者 / 項目編號	擴展網路資源連結相關性認定								
	U1	U2	U3	U4	U5	U6	U7	不相關總計	相關度
擴展連結網站_1	✓	✓	✓	×	✓	✓	✓	1	86%
擴展連結網站_2	✓	✓	✓	×	✓	✓	✓	1	86%
擴展連結網站_3	✓	✓	✓	×	✓	✓	✓	1	86%
擴展連結網站_4	✓	×	✓	×	×	×	×	5	29%
擴展連結網站_5	✓	✓	✓	×	✓	✓	✓	1	86%
擴展連結網站_6	✓	✓	✓	×	✓	✓	✓	1	86%

擴展連結網站_7	✓	✓	✓	✓	×	✓	✓	1	86%
擴展連結網站_8	✓	×	✓	✓	×	×	×	4	43%
擴展連結網站_9	✓	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_10	✓	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_11	×	×	✓	×	×	×	×	6	14%
擴展連結網站_12	✓	✓	✓	×	×	✓	✓	2	71%
擴展連結網站_13	×	×	✓	×	×	×	×	6	14%
擴展連結網站_14	✓	×	✓	×	×	×	×	5	29%
擴展連結網站_15	×	×	✓	×	×	×	×	6	14%
擴展連結網站_16	✓	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_17	✓	×	✓	×	×	✓	✓	3	57%
擴展連結網站_18	✓	×	✓	×	✓	×	×	4	43%
擴展連結網站_19	✓	×	✓	×	✓	✓	✓	2	71%
擴展連結網站_20	✓	×	✓	×	✓	✓	×	3	57%
相關	17	10	20	5	11	13	12	前 10 名	79%
不相關	3	10	0	15	9	7	8	前 20 名	63%

圖表 28：研究實驗之問卷統計資料

研究中，我們會評估原先藉由人工蒐集歸類的網路連結資源及實驗中所得到網路資源擴展連結結果是否與所歸類的類別一致，而評估方式中是透過使用者來檢視這些連結與所歸類之知識類別是否確實相關。因為網際網路擁有的資訊相當繁多，所以並無所謂「標準答案集」可供比對實驗結果的正確率；一般使用者藉由搜尋引擎進行檢索時，動輒能蒐尋出成千上萬筆「相關」資訊，「資訊過載」反而造成使用者在資訊取得的困難。

因此本實驗僅能就被擴展出之網站頁面是否與分類類別相關與否來做為系統正確率的評估，實驗中也對原始網站連結進行人工評估驗證，因為網路擴展的資料來源如果錯誤性太高，勢必影響擴展的結果，所以評估分析原始資料之正確性也是能了解擴展結果成效所必須的。

因此分別統計計算正確筆數與資源擴展相關度，資源擴展相關度為計算正確筆數占總筆數百分比即精確率之算法，得出下圖表 29、30 之數量分布示意表；並

將「網路資源擴展結果認定」與「來源網站連結相關性認定」分別進行各項網站連結相關度和各網站相關度的平均值統計計算，並從其相關程度與平均數值中深入探討與觀察，該公式如下所示：

$$網站_i連結相關度 = \frac{標示相關人數_i}{受測總人數} \times 100\%$$

$$網站平均相關度 = \frac{\sum_i 網站_i連結相關度}{網站總連結數} \times 100\%$$

原始網路連結資源認可比例		
	標示度達七成以上	標示度未達七成以上
來源網站(10 筆)	8 (80%)	2 (20%)

圖表 29：來源網站連結相關性認定問卷結果相關程度分布示意表

擴展後網站連結認可比例		
	標示度達七成以上	標示度未達七成以上
取擴展結果前 10 筆統計	8 (80%)	2 (20%)
取擴展結果前 20 筆統計	11 (55%)	9 (45%)

圖表 30：網路資源擴展結果認定問卷結果相關程度分布示意表

如以使用者對於網站連結相關度認定標示度達七成以上來統計相關網站連結個數，我們發現由人工篩選分類的原始網站連結有八成在使用者認知上是相關於分類主題的，而實驗結果所擴展之網站連結，如果取排名前 10 名的連結進行驗證，其網站認可度與由人工整理的來源網站不相上下，然而取前 20 名的擴展網站連結進行評估，其認可度確實滑落至 55%，如圖表 29、圖表 30 所示之。

觀察圖表 27~圖表 30 發現使用者因為個人知識背景的差異對於知識內容的認定會有些許落差，此一現象屬正常，但實驗中觀察到 4 號受測者於受測群體中明顯有認知差異過大的問題，對於半數以上的網站連結均表述為不相關，與實際網

站內容有嚴重差異；因此為使實驗分析結果不受獨特性認知影響，因此將去除此認知落差較為嚴重之受測者(4 號受測者)。去除認知差異大的受測者資料後，雖然網站連結認可相關度比例並無變動，但從網站平均相關度來觀察，原始網站連結資料從原先 77.14%提升到 87%，擴展網站連結如果取前 10 名連結也從 79%提升至 85%，取前 20 名亦提升到 5%達 69%，也有近七成的相關度。

實驗中進一步進行錯誤分析時發現受測者會因知識階層而產生認知落差，在深入詢問受測者後，得知受測者彼此間的認知可能受教育程度、環境影響...等因素影響其作答之結果，可能會將大多數受測者認為具相關性的鏈結站點作答為不具相關性，舉例來說：在地球科學的範疇之下，廣義上可涵蓋至太陽、地球或宇宙也可小至地球中的森林、海洋、河川或水源，如受測者在主題式的引導下已具有狹隘性觀點直指某一含意時，則擴展結果的相關性與否已不具任何可信度了。因此，如提示測試者排除這樣的認知落差，我們重新整理找出此類認知落差影響的項目並用紅色標記✓(如圖表 31 與圖表 32 所示)，若將其歸納為具相關性，則其資源擴展相關度數值的改變如下：

測試者 項目編號	來源網站連結相關性認定							
	U1	U2	U3	U5	U6	U7	不相關	相關度
來源網站連結_1	✓	✓	✓	✓	✓	✓	0	100%
來源網站連結_2	✓	✓	✓	✓	✓	✓	0	100%
來源網站連結_3	✓	✓	✓	✓	✓	✓	0	100%
來源網站連結_4	✓	✓	✓	✓	✓	×	1	83%
來源網站連結_5	✓	✓	✓	✓	✓	✓	0	100%
來源網站連結_6	✓	✓	✓	×	✓	✓	1	83%
來源網站連結_7	✓	×	×	×	✓	✓	3	50%
來源網站連結_8	✓	✓	✓	✓	✓	✓	0	100%
來源網站連結_9	✓	✓	✓	✓	✓	✓	0	100%
來源網站連結_10	✓	✓	✓	✓	✓	✓	0	100%
相關	10	9	9	8	10	9	總相關度(百分比)	92%
不相關	0	1	1	2	0	1	不相關度總計(百分比)	8%

圖表 31：去除認知落差的分析結果集(一)

測試者 項目編號	擴展網路資源連結相關性認定							
	U1	U2	U3	U5	U6	U7	不相關	相關度
擴展連結網站_1	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_2	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_3	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_4	✓	×	✓	×	×	×	4	33%
擴展連結網站_5	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_6	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_7	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_8	✓	×	✓	×	×	×	4	33%
擴展連結網站_9	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_10	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_11	×	×	✓	×	×	×	5	17%
擴展連結網站_12	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_13	×	×	✓	×	×	×	5	17%
擴展連結網站_14	✓	×	✓	×	×	×	4	33%
擴展連結網站_15	×	×	✓	×	×	×	5	17%
擴展連結網站_16	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_17	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_18	✓	×	✓	✓	×	×	3	50%
擴展連結網站_19	✓	✓	✓	✓	✓	✓	0	100%
擴展連結網站_20	✓	✓	✓	✓	✓	✓	0	100%
相關	17	13	20	14	13	13	前 10 名	87%
不相關	3	7	0	6	7	7	前 20 名	75%

圖表 32：去除認知落差的分析結果集(二)

原始網路連結資源認可相關度比例		
	標示度達七成以上	標示度未達七成以上
來源網站(10 筆)	9 (90%)	1 (10%)

圖表 33：去除認知落差後網路資源擴展結果認定問卷結果相關程度分布示意表

擴展後網站連結認可相關度比例		
	標示度達七成以上	標示度未達七成以上
取擴展結果前 10 筆統計	8 (80%)	2 (20%)
取擴展結果前 20 筆統計	13 (65%)	7 (35%)

圖表 34：去除認知落差後來源網站連結相關性認定問卷結果相關程度分布示意表

可從資源擴展相關度的數值中發現，原始選自國家圖書館知識資源參考服務的網路資源選介所使用人力篩選出的站點鏈結資料在網站平均相關度的檢視達92%，與原先不排除受測者認知落差的 77.14%大幅提升了十五個百分點，其實也顯示即使由人工建置維護的網站連結資源，一般人在使用時在認知上也存在著極大的差異，如圖表 31 所示。然而反觀實驗中擴展的網路資源連結，若僅擷取前十名擴展結果其網站平均相關度從 79%提升至 87%，前二十名從 63%提升到 75%，雖然不及人工篩選的結果，但從使用者測試結果來看相差不大，如圖表 32 所示。

而從使用者對於網站連結認可度來看，經人工篩選的原始網站連結有九成的認可度，如以取前 10 名擴展網站連結來看也有八成，前 20 名擴展的網站連結經使用者檢核也有六成五的認可度，但從個別網站連結相關度觀察，相關與不相關鑑別度相當高，模稜兩可的狀況低，代表實驗中擴展而得的網站連結雖然不像原始人工篩選式連結認可度那麼高，對使用者而言很清楚能識別其是否歸屬在該類別，如圖表 33、圖表 34 所示之。

從實驗結果分析，網站擴展成效受到原始網站相關連結建置精確與否、以及是否能正確擷取代表該連結網站的關鍵詞,在這個面向機器不如人為判讀來的精確，而自然語言本身的複雜度以目前電腦處理的技術亦無法完全解決，所以關鍵詞集的擷取勢必會有些雜訊詞混雜其中，此類詞彙會直接影響擴展的效能，其次詞彙本身語意的意涵也會影響使用者對資訊認定上的差異。舉例而言，如以「台灣」一詞進行檢索，廣義來看可得到與台灣地理或人文相關之網站鏈結資料；但於狹

隘的角度來看，在擴展查找的過程中，搜尋引擎也會得到一些台灣本地的政府機關鏈結站點或百科類說明網站站點夾雜其中。因此，這些廣義的詞彙的確間接地影響了擴展實驗的成效。從數據上來觀察，擴展後網站連結資料集之擴展結果其平均網站相關度百分比將近 87%，而由人工整理歸類的原始相關連結資料在平均網站相關度評比上將近 92%，於兩者間的相關程度上進行比較分析，擴展後的網站連結資料集結果略低了五個百分點，但從結果來看，擴展後網站連結集與原始相關網站連結集兩者間的差異並無太大之差距，從使用者對網站認可度若僅考慮擴展連結的前十名，有近八成被判定幾乎完全相關，這樣對自動擴展網站相關連結應有極大的幫助。

第六節 錯誤分析

由上述分析結果得知，排除受測者知識階層落差的情況，透過網路資源擴展得到二十筆網路擴展連結中，從受試者認定的結果中所涵蓋的十三筆網站連結受試者認定為絕對相關，所以透過網站連結相關度進行擴展資源的篩選，可以期待藉由網路資源自動擴展來取代利用人力所歸納建置的網路相關連結方式。

為持續提升網路資源擴展成效，我們觀察連結相關度較低的網站連結，回溯原始擴展後資料集以取得相關內容以進行更深入的錯誤分析，首先，以相關度為17%之網站開始進行回溯其關鍵詞集之內容，如有重複性之詞彙在此處將予以省略用以將表格精簡化；由下圖表 35 之擴展後編號第 11 項網站為例，可看出透過關鍵詞彙在初次擴展的結果中，縱使蒐集到的網站連結位址略有不相同，但其根網址均為同一個網站站址，意即網域名稱相同，由此發現編號第 11 項的網站因為關鍵詞彙所擴展得知的根網站連結為 http://www.bot.com.tw/，而該連結為「台灣銀行」的官方網站，該連結本身具有入口網站的特質，提供了使用者經過蒐集與整理後的某些特定的資訊內容之服務，使得使用者可以透過此網站快速地獲得相關

資訊與知識；且由其回溯的關鍵詞集進行深入性觀察，觀察出其詞集結果在自然科學中的「地球科學類」類別主題中皆較不具代表性，但是又因為此關鍵詞集中所產生之詞彙數量重複性所屬較高，因此仍會成為具影響力之雜訊詞。

網站編號	擴展後網站名稱	取根網址連結內容	關鍵詞內容	擴展後原始連結內容
11	台灣銀行	http://www.bot.com.tw/	供應 臺灣 ……	www.bot.com.tw/procurement/procure_supply www.bot.com.tw/procurement/procure_supply/.../default.aspx www.bot.com.tw/ ……
13	教育部全球資訊網	http://www.edu.tw/	教育 資訊 基金會 資料 ……	www.edu.tw/ www.edu.tw/Default.aspx?WID=3ee9c9ee-f44e-44f0-a431... www.edu.tw/pages/list.aspx?Node=1350&Type=1... ……
15	華藝世界美術資料庫	http://www.airitiart.com/	資料庫 資料 ……	www.airitiart.com/

圖表 35：由相關度為17%之網站回溯至原始擴展後資料集之相關關鍵詞彙

網站編號	擴展後網站名稱	取根網址連結內容	關鍵詞內容	擴展後原始連結內容
8	中央研究院	http://www.sinica.edu.tw/	研究 推動 研究 推動 ……	www.sinica.edu.tw/index.shtml www.sinica.edu.tw/institute.htm www.sinica.edu.tw/~mrpcwww/ ……

4	全國法規資料庫	http://law.moj.gov.tw/	資料 目的 地區 保護 團體 資料庫 措施 範圍 ……	law.moj.gov.tw/ law.moj.gov.tw/LawContentDetails.aspx?id=FL010631 law.moj.gov.tw/LawClass/LawContent.aspx?PCODE=Q0010001 law.moj.gov.tw/LawClass/LawAll.aspx?PCode=I0050021 law.moj.gov.tw/LawClass/LawAll.aspx?PCode=D0050091 law.moj.gov.tw/Index.aspx law.moj.gov.tw/LawClass/LawAll.aspx?PCode=N0030019 law.moj.gov.tw/LawClass/LawContent.aspx?pcode=L0040084 ……
14	臺灣大學圖書館電子資料庫檢索	http://dbi.lib.ntu.edu.tw/libraryList2/	資料 資料庫 資料 資料庫 ……	dbi.lib.ntu.edu.tw/libraryList2/ dbi.lib.ntu.edu.tw/libraryList/ ……

圖表 36 ：由相關度為 33%之網站回溯至原始擴展後資料集之相關關鍵詞彙

網站編號	擴展後網站名稱	取根網址連結內容	關鍵詞內容	擴展後原始連結內容
18	國立臺灣師範大學地理學系	http://www.geo.ntnu.edu.tw/	地理 研究 地形 研究 地理 ……	www.geo.ntnu.edu.tw/ www.geo.ntnu.edu.tw/files/publish/166_dcc78854.pdf www.geo.ntnu.edu.tw/files/archive/46_1e2c110d.pdf www.geo.ntnu.edu.tw/archive/archive.php?class=102 ……

圖表 37：由相關度為 50%之網站回溯至原始擴展後資料集之相關關鍵詞彙

次者，以相關度為 33%與 50%之網站為回溯之對象，整理出圖表 36 與圖表 37 之說明細項，於圖表中可看出網站具有圖表 35 所觀察出之特質，並因其專業性與獨特性導致產生相同、相近之詞彙或是專業性詞彙所擴展出的站點集合皆會出自同一個網站根站址，如果網站站址為政府入口網站則此特色更加顯著；透過此三類觀點的深入觀察得出，現今網際網路上的網站站點，開始趨向於提供使用者一些基本的入口網站功能，且其專業化與豐富程度更勝從前；而也由此三張圖表中的關鍵詞彙中觀察出雙連語彙所佔之比例遠遠低於雙字詞的比重，因此得知雙連語彙的擴展結果較雙字詞的擴展結果關聯性更具代表性。

實驗結果顯示，透過本研究自動化擴展之實驗結果，發現透過自動化資源擴展可減少知識弱勢族群查找知識資源的不易性，也成功地提升網路資源擴展系統之效益，由實驗結果得知經過本研究實驗所產生的擴展後擴展連結資源結果集是具頗佳地系統效益的。

第五章　結論與建議

本研究希望在現有的網際網路架構之下，透過自然語言處理、關鍵詞擷取、字詞間相似性的計算等相關技術，並結合搜尋引擎來建構一個藉由原有網站內的相關連結資源來自動擴展出其他相關網站連結資源的機制，藉此希望來建置知識入口網站；在這樣的機制下可幫助知識入口網站的參考性連結資源可定期自動的擴展與更新，讓使用者能藉由這樣的知識入口更加容易查找到所需的相關資訊；進而幫助高、低年齡層中較沒有檢索能力的使用者，使其可以獲得更多樣、廣泛且高相關性的網頁資源做為其檢索資訊的參考性網站；本文旨在研究如何透過查詢擴展之技術提升資訊檢索之成效,協助使用者檢索出更多符合自身需求之資訊，本章彙整前章所得之實驗結果與觀察之結論及未來研究之建議。

第一節　結論

在文件自動分類處理程序中，文件的特徵擷取是影響分類結果的重要因素之一，所以特徵擷取方法的選擇與設計是具重要性的考量之一(曾元顯, 2002)。為此本研究將實驗分為四個階段，分別為「前置處理之詞彙分析」、「特徵詞彙選取」、「資訊擷取與查詢擴展」與「實驗結果的評估與成效分析」,在這樣的研究架構下，我們可以針對每一階段輸出的結果進行詳實地觀察，以發現並了解影響實驗成果的因素為何，並將其作為後續研究改進的依據。

一、　查詢擴展成效

在去除特異性過大之受測者與受測者個體間的知識階層落差問題後，實驗結果顯示，利用結合網路資源擴展與特徵擷取之方式，在調整權重後所得到之資源擴展相關度或是進行錯誤分析後所得到的結果，確實有效地將原始

網站所提供的連結資料進行擴展。從實驗研究結果發現，利用本研究實驗所產生之二十筆網站鏈結資料，其擴展後相關程度與原先網站以人力方式所建置之結果比較分析後不相上下，且各別以其檢索失誤率相較起來，兩者的失誤率也是伯仲之間，自動化擴展後的實驗結果毫不遜色於由人力篩選後的資料鏈結集，由前章實驗分析中所得到之統計資訊足以判定本研究所採用之實驗方法，已擁有足夠的條件可以成為一個擴展文件之依據標準。

二、 特徵詞彙選取之檢索策略

透過上述實驗結果發現，在特徵詞彙選取階段的人工介入過濾去除雜訊詞，並進行第一階段的雙字詞的一次擴展，將其一一丟入搜尋引擎進行查找而得到的初始擴展結果與階段二採用 TF-IDF 方法進行特徵詞權重計算後取其權重前十名的特徵詞進行雙連語彙的二度擴展，即將雙連語彙於搜尋引擎中兩兩丟入查找，得到一進階擴展結果，由擴展結果集合中發現雙連語彙兩兩詞彙進行查找之成效更勝使用雙字詞個別進行查找之結果且其相似性也更高。

三、 檢索詞彙對於查詢擴展階段之影響

觀察實驗中檢索時所使用的檢索詞彙，當檢索查找的主題網站站點描述過於冗長與複雜時，於「前置處理之詞彙分析」階段所得到的關鍵詞彙會含蓋較多干擾的雜訊詞或對檢索成效較沒有貢獻的詞彙，正因為過多的特徵詞彙容易耗費過多的計算時間與記憶體空間，因而導致系統效率不佳，所以檢索詞彙的雜訊詞或較低影響力的詞彙愈多時，會大大地降低查詢擴展時的檢索成效，因此在不影響分類效果的條件下，本研究利用特徵選取的方式進行過濾與篩選，如去除停用詞、雜訊詞、量詞與副詞等詞彙，以降低此類特徵

詞彙的數量，進而提升系統的效率。

透過本研究之實驗結果觀察出，從一開始的實驗結果至最後的分析結果，擴展連結相關性的大幅提升與檢索失誤率的大幅降低可看出本實驗的擴展結果得到了一個更佳的推薦鏈結集合群，且此集合也精準地符合檢索者的資訊查找需求，因此，這樣的擴展結果已具有一定程度的替代性，意即為原先由人力作業建置的方式提供了一線上即時與自動化擴展之功能；如此一來，透過本研究也可彌補了入口網站中提供相關網頁資源服務時可能產生的資源不足、過時或是狹隘性之缺陷。

第二節　　未來展望

以下針對本研究之實驗結果及研究過程中所遭遇的困難，我們整理了幾個未來研究方向與建議：

1. 提升擴展文件特徵擷取之品質

   本研究所提出的擴展文件所使用的特徵擷取方法，仍未盡完善，當網站內容敘述與網站名稱標題較短且所用詞彙為高頻詞但屬一般通用語彙時，這類的語彙往往會造成搜尋引擎在檢索時較容易檢索出較不相關的網站連結，對系統造成雜訊而導致的系統效能降低之情形；未來的研究中，可利用字詞的搭配進行組合查詢檢索，以縮小擷取搜尋引擎所回饋的資訊範疇，甚至擷取的關鍵詞若能利用語意訊息來篩選，相信也能增加所擴展網站內容的相關性。

2. 增加專家式選詞

   實際進行自動擴展時，可能會因為原始網站提供相關連結不足或所整理的網站的誤失，使得某些詞彙在實驗訓練的過程中權重並不高，或是根本沒

將此類詞彙納入訓練語彙集中，因此即使實驗結果集中有出現相似之網站站點資訊，但因關鍵詞彙的權重不高之因素而被去除，因此，建議未來的實驗中可藉由專業領域中的專家針對斷詞後之關鍵詞彙擷取進行專家選詞來改善上述所提之問題。

3. 查詢擴展之詞彙對檢索成效之影響

在本研究實驗中進行觀察所得到的結果中，是透過斷詞與篩選後的詞彙進行檢索成效之分析，因為並沒有各領域專業詞彙的資訊，所以並無法針對個別詞彙的檢索成效進行充分理解與統計，進而了解哪些詞彙對排序時的提升有幫助、哪些詞彙易造成系統效能之降低等，在後續研究中可再加以分析，不僅對於系統針對詞彙篩選時有所助益，對於提高網路資源擴展的成效應有很大的助益。

4. 擴展後資源網站內子節點的站址探討

本研究檢索成效與擴展品質由本實驗中的錯誤分析中所觀察得出之結果，得知如由擴展後的資源網站內容中所擷取之根網址深入探討至根網址內的各子節點所相對應之站址，發現子節點之站址會有比母節點的實驗分析更加精確的實驗結果，不僅對於檢索成效會有提升之效果，相信對於未來系統的擴展品質也能增加其內容之相關性。

對於網際網路中資訊的快速成長，使用者在搜尋真正所需的資訊，似乎越來越困難，所以建置一個知識入口來幫助資訊能力較薄弱的人能快速有效查找到所需的資訊，也能定期自動擴展與更新相關網路知識連結，對於使用者再利用網際網路進行知識探索是相當重要的,所以希望能更進一步提升自動擴展相關之技術，以提供資訊檢索者一個更佳之檢索環境。

英文參考文獻

AdamovAbzetdin. (2012 年 3 月 17 日). The Explosive Growth in the volume of Digital Data will demand more IT Professsional. 擷取自 Abzetsin Adamov's IT Blog: http://aadamov.wordpress.com/2012/03/

Berners-Lee, T. (2000). Weaving the Web: The original design and ultimate destiny of the World Wide Web. *Harper Business*.

Christopher, D. M., Prabhakar, R., & Hinrich, S. (2012). *Introduction to information retrieval 資訊檢索導論.* (柯皓仁, Ed., & 王斌, Trans.)

Clarke, I., & Flaherty, T. B. (2003). Web-Based B2B Portals. *Industrial Marketing Management*(32), pp. 15-23.

Delphi Group. (2004). Information Intelligence: Content Classification and the Enterprise Taxonomy Practice.

Eszter, H. (2002). Second-Level Digital Divide:Differences in People's Online Skills. *First Monday, 7*(4).

Gordon, C., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management*(35(2)), pp. 141-180.

HaywoodTrevor. (1998). Globle Networks and the Math of Equality:Trickle Down or Trickle Away? In Cyberspace Divide-Equality, Agency and Policy in the Information Siciety.

John, G., & David, R. (2012). *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.* EMC Corporation.

Klir, G. J., & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic:Theory and Applications.* Prentice Hall.

KossF.A. (2001). Children falling into the digital divide. Journal of International Affairs, 頁 75-90.

OECD. (2001). Closing the gap:Securing benefits for all from education and training, in Education Policy Analysis. Paris.

RicardoB.Y., & BeithierR.N. (2002). Modern Information Retrival. Addision-Wesley.

Salton. (1988). Automatic text processing:the Transformation,Analysis,and Retrieval of Information by Computer. Mass.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval.* McGraw-Hill.

Salton, G.、Buckley, C. (1988). Term-weighting approaches in automatic retrieval. Imformation Processing and Management(24), 頁 513-523.

Zahir, S., Dobing, B., & Hunter, M. G. (2002). Cross-cultural dimensions of. *Internet Research*(12), pp. 210-220.

中文參考文獻

卜小蝶. (2007 年 12 月). 網路自動分群搜尋引擎之使用者評估研究. 圖書資訊學研究, 頁 55-80.

中央研究院詞庫小組. (無日期). Auto Tag 中文斷詞工具. 擷取自 Auto Tag 中文斷詞工具: http://ckipsvr.iis.sinica.edu.tw/

王力行. (1998 年 9 月). 資訊、知識、智慧. 遠見雜誌(147).

任炳魁. (2008). 數位典藏庫中資料分群產生之研究-以數位學習詮釋資料為例. 碩士論文, 輔仁大學, 圖書資訊學系.

吳典恩. (2007). 結合本體論以及關聯法則於查詢擴展之研究. 台南: 國立成功大學資訊管理研究所碩士論文.

李伯毅. (2004). Support Vector Machine 技術應用於中文文件自動分類之探討. 碩士論文, 國立高雄應用科技大學, 電機工程系.

林仁奎. (2011). 影響使用入口網站黏著性因素之研究-以 Yahoo!奇摩網站為例. 花蓮縣: 國立東華大學企業管理學系碩士論文.

邱俊銘. (2010). 以 Web 資訊擷取和知識本體融合方法整合領域內容和知識. 南投縣: 國立暨南國際大學資訊工程學系.

柯皓仁. (2013). 海量資料與圖書館. 擷取自 http://www.slideshare.net/clavenke/ss-23923075

柯皓仁. (2013). 鏈結資料在圖書館的應用. (頁 9). 2013 電子資訊資源與學術聯盟研討會.

柯皓仁. (2013). 鏈結資料在圖書館的應用. 2013 電子資訊資源與學術聯盟研討會, (頁 9).

范瓊文. (2005). 主題概念階層模型：概念式搜尋. 碩士論文, 國立中央大學, 網路學習科技研究所.

高永煌. (2010). 台灣地區族群數位落差現象與治理政策之研究. 碩士論文, 國立台北大學資訊管理研究所.

莊大衛. (2005). 文件自我擴展於自動分類之應用. 新北市: 輔仁大學圖書資訊研究所碩士論文.

陳信源、葉鎮源、林昕潔、黃明居、柯浩仁、楊維邦. (2009 年 1 月). 結合支援向量機與詮釋資料之圖書自動分類方法. 資訊科技國際期刊, 第三卷(1).

曾元顯. (2002). 文件主題自動分類成效因素探討. 中國圖書館學會會報(68), 頁 62-83.

曾元顯. (2012 年 10 月). 雙語詞彙、學術名詞暨辭書資訊網. 擷取自 國家教育研究院: http://terms.naer.edu.tw/detail/1678994/

曾淑芬. (2002 年 1 月). 數位落差. 資訊社會研究, No.2, 頁 234-237.

黃嘉宏. (2008). 基於自動分類為基礎的圖書提名特徵截取之研究-以輔助圖書分類系統為例. 新北市: 天主教輔仁大學圖書資訊學系碩士班碩士論文.

楊智元. (2001). 符合使用者個人喜好需求的入口網站介面之研. 碩士論文, 大葉大學資訊管理研究所.

葉俊榮. (2006 年 2 月). 台灣數位落差現狀與政策. 研考雙月刊, 第三十卷(1), 頁 3-16.

廖秀紋. (2007). 一般、客家及原住民族群數位落差探討. 碩士論文, 天主教輔仁大學應用統計研究所.

維基百科. (2009). 擷取自 Wiki 維基百科: http://zh.wikipedia.org/wiki/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E。

維基百科. (2013 年 11 月 24 日). TF-IDF. 擷取自 維基百科: http://zh.wikipedia.org/wiki/TF-IDF

劉冠宏. (2009). I3S-用於建構領域相關知識入口之智慧型資訊系統. 南投縣: 國立

暨南國際大學資訊工程學系碩士論文.

蔡育欽. (2005). 查詢擴展之詞彙摘選應用於主題檢索之研究. 新北市: 天主教輔仁大學圖書資訊學系碩士班碩士論文.

謝水木、黃敬仁. (2008). 整合性數位資訊管理入口網站規劃與建置. Journal of Comercial Modernization, 4(4), 頁 17-28.