網頁地理資訊檢索與探勘—以民宿主題為例

鄒明城
國立高雄海洋科技大學航運技術系

韓慧林
實踐大學高雄校區資訊管理學系

邱景星
實踐大學高雄校區觀光管理學系

摘要

網際網路上散佈了各式主題與大量的網頁資料，其中隱含了非常多的知識，但是這些內容大多是半結構性，甚至是非結構性的資料，如何能夠有效率的管理這些資料，並且進行資訊與知識的擷取，一直是研究與開發的重點，因此也就有各式各樣的網路搜尋引擎、資料探勘以及網路行銷技術的開發。但是目前一般的網路搜尋技術大多只著重於關鍵字的檢索，對於網頁內容與主題的分析，則仍未盡理想。另外，對於網頁內容中的地理資訊，也未能進行有效的檢索與分析，以致於犧牲了許多內含的地理資訊。

本研究以網頁中的民宿主題為例，使用Google Search Web Service 為網路搜尋的基礎，結合中央研究院詞庫小組開發的斷詞斷字系統與文字資料探勘的技術，對於Google所搜尋到的網頁，進行空間與語意內容的探勘、檢索與排序，找出與所查詢主題在內容與地理資訊上最相關的網頁。接著，透過地理資訊檢索與正規表示式，由這些篩選過的網頁內容中，檢索出有用的地理資訊，再透過Google Map API地址對位的技術，將檢索出來的地理資訊與文字內容結合顯示於Google Map地圖上。以這樣的方式所搜尋出來的結果，將是包含了地理資訊的圖與文，且更貼近需求的查詢結果，將可應用於各種與空間主題相關之內容的查詢、分析、地理資料蒐集與空間知識的管理上。

關鍵字：地理資訊檢索、文字探勘、網頁探勘、正規表示式

Geographic Information Retrieval on Web Pages-Taking Homestay as An Example

Ming-Cheng Tsou
Department of Shipping Technology, National Kaohsiung Marine University

Hui-Lin Hai
Department of Information Management, Shih Chien University, Kaohsiung Campus

Ching-Hsing Chiu
Department of Tourism Management, Shih Chien University, Kaohsiung Campus

Abstract

The World Wide Web (WWW) offers an enormous spread of information and data, and assembles a tremendous amount of knowledge. Much of this knowledge however, comprises either non-structured data or semi-structured data. In order to make use of these unexploited or underexploited resources more efficiently, the management of information and data gathering have become essential direction for research and development. However, at the present moment, the ability of regular search engines to access and use this data, is still far from perfect, since it is limited to the retrieval of basic keywords rather than analysis of the subject matter and content of the webpage itself. In addition, there are limited capabilities for effective retrieval and analysis of implicit geographic information contained within the webpage.

This paper focuses on the task of researching a hostel or homestay by using the Google Search Web Service as a base search engine. From the search results, mining, retrieving and sorting out location and semantic data were carried out by combining the Chinese Word Segmentation System with Text Mining technology in order to find geographic information thatthatthat can be derived from the webpage. The results obtained from this particular searching method allowed users to get closer to the answers they sought and achieve greater accuracy, since the results included graphics and associated textual geographic information. In the future, this method may be suitable for and applicable to various types of queries, analyses and geographic data collection, and in managing spatial knowledge related to different keywords within a document.

Key words: geographic information retrieval, text mining, web mining, regular expression

壹、前言

網際網路上散佈了各式主題與大量的網頁資料，已成為人們每日獲取訊息的重要來源之一，其中隱含了非常多的知識，如何有效的搜尋與檢索其中的資訊，有賴於網頁搜尋引擎，但因這些網頁內容大多是半結構性，甚至是非結構性的資料（Mitra & Acharya 2003），而目前一般的網頁搜尋技術大多只著重於關鍵字的檢索，對於網頁內容與主題的分析能力，則仍未盡理想，因此，如何能夠有效率的管理這些資料，並且進行資訊與知識的擷取，一直是研究與開發的重點，於是有各式各樣的資訊檢索以及資料探勘技術的研究，只是這些技術大多只針對語意內容而開發。但是根據估計，人們在對搜尋引擎下達的搜尋條件中，有五分之一具有空間上的敘述，例如，找餐廳、找戲院、找學校，而在這些查詢中有百分之八十包含有地點名稱（Kornai & Sundheim 2003 ; Souza et al. 2005），例如，找尋高雄的美食，或是墾丁的民宿。礙於現有的發展，一般搜尋引擎對於網頁內容中的空間資訊，大多未能進行有效的檢索與分析，針對「墾丁民宿」這樣的搜尋條件，搜尋引擎只是進行「墾丁」與「民宿」這二個關鍵字的拆解，然後再到資料庫中將包含這二個關鍵字的網頁資料擷取出來，這樣的過程其實只做關鍵字的檢索，缺乏主題語意與空間敘述（spatial context） 上的處理，因此所得到的結果，有時往往會與查詢需求差距甚大（Buyukkokten et al. 1999）。

地理資訊系統雖然具有很強的空間處理能力，但對於空間資訊的存取大多限於以空間座標型態所構成的幾何空間表達，只是一般人對於日常生活的空間表達，大多是以文字敘述或是以地名的方式來表達空間概念，例如，地名、地標、地址甚至是電話號碼，來描述事件、關係或產品，這些都具有空間上的位置資訊，少有人會使用明確的空間座標來描述（Jones et al. 2001）。網頁的內容正是這種現象的具體展現，人們把對於空間資訊的表達書寫於文字敘述中，透過這些內隱（implicit） 的空間資訊，我們可以將網頁與某一空間位置連結一起，目前地理資訊系統的分析功能很難與文字資訊的分析進行整合，雖然現有一些所謂的網路地理資訊系統（Web GIS），但是他們的功能仍然不脫傳統地理資訊系統的處理範疇。

針對上述的需求，單是依靠文字探勘技術或地理資訊系統，都是不足的，一個能處理空間資訊查詢的搜尋引擎，應該同時兼具處理主題語意上與空間敘述（spatial context） 上的解析與度量，並且可以將查詢條件拆成空間與主題二部份進行處理，彼此互補。在這樣的需求下，地理資訊檢索（Geographical Information Retrieval - GIR） 也就成為滿足這方面查詢問題的重要研究課題，目前也正逐漸受到學術界與商業界的重視（Byrd & Ravin 1999；Jones et al. 2002），它結合了文字探勘（Text Mining）、資訊檢索（Information Retrieval-IR） 與空間處理、空間認知等相關領域的研究，目標是能更精確且有效的從文字資料中檢索出與空間相關的資訊。

正因人類活動中約百分之七十五到百分之八十的訊息與地理空間位置有關，而「旅遊」又是一項極具空間特性的人類活動（李俐槿等 2007）。隨著週休二日的休假政策已

行之多年，間接帶動了國人利用假日出遊的風氣，許多不同類型的休閒農業蓬勃發展，其中又以民宿獨占鰲頭迅速增加。民國92年觀光局統計資料顯示台灣合法民宿僅有124家，直至民國95年3月觀光局資料顯示申請登記的民宿有1,845家，而合法登記營運的民宿則高達1,250多家，總法定登記房間數為5,120間。但市場上亦有資料顯示，國內未合法登記的民宿至少超過3,000多家。保守估計全台灣民宿不管登記有案與否大大小小超過萬家以上，隨著旅遊風氣的盛行和住宿習慣的改變，民宿每年以200家左右的速度增長著，記錄著民宿資料的網頁也迅速遞增。因此，只要上網搜尋相關民宿資料往往可以獲得相當大量的資訊，其中有些與搜尋主題無關，有些則僅有文字描述，欠缺較具體的空間資訊展現，使用者往往需要耗時逐一點選消化內容，缺乏一個整體性的資料展現，因此透過地理資訊檢索的功能，可以提供一個更具符合查詢要求與整合相關資訊的搜尋功能。

本研究以網頁中的民宿主題為例，使用Google Search Web Service 為網路搜尋的基礎，結合中央研究院詞庫小組開發的斷詞斷字系統與文字資料探勘的技術，對於Google所搜尋到的網頁，進行空間敘述（spatial context） 與語意內容的探勘、檢索與排序，找出與民宿主題在內容與空間資訊上最相關的網頁。接著，透過地理資訊檢索與正規表示式（Regular Expression），由這些篩選過的網頁內容中，檢索出有用的地理資訊，如地址，再透過Google Map API地址對位的技術，將檢索出來的空間資訊與文字內容結合顯示於Google Map地圖上。以這樣的方式來改善搜尋引擎的不足，讓所搜尋出來的結果，包含了空間資訊的圖與文，更貼近使用者的需求，且可進行各種與空間主題相關之內容的查詢、分析與空間資料蒐集，將可應用在資料探勘與空間知識的管理。

貳、相關研究探討

過去在進行有關網頁空間資訊檢索的研究常可分為以下二類，一是以網頁實體位置為基礎型（context entity-based），另一個則是以網頁內容為依據（content-based）（Amitay et al. 2004 ; Martins & Silva 2005）。以網頁實體位置為基礎主要是根據網頁的IP位址以及網域服務（DNS），找出伺服主機或網頁本身所處的可能位置。Gtrace tool（Periakaruppan & Nemeth 1999） 就是類似的研究工具，Buyukkokten et al.（1999） 則根據網址由網域資料庫中找到網域管理員的資料，再由管理員的電話資料的區域號碼或郵遞區號，判斷出該網址的可能空間位置，藉以找出網站與網站間的空間關聯性。McCurley（2001） 根據網頁的IP位址、網站管理員的電話號碼，對應網頁的空間位置，並且提供一個網址與網頁對應的地圖使用介面。這些作法所找到的位置只是一個可能位置，相當不精確，而且只能代表該網址，與網址內的網頁內容無關，無法找出網頁內容所描述的空間位置，再者，於主機代管的運作模式下，以及行動通訊日漸普及的情況下，網址位置也不再能代表空間敘述了，僅能提供參考而已。而以網頁內容為依據的搜尋方式，則是根據搜尋的條件，對網頁內容分別進行主題相關與空間敘述上的檢索，透過這樣的方式，更能夠傳達該網頁所代表的空間意義以及滿足使用者的查詢需求（McCurley 2001；Vogel et al. 2005）。限於網頁實體位置只能提供相當有限性的空間資

訊，本研究將強調在網頁內容上空間資訊檢索的探討。

一、相關研究計畫

GIPSY（The Goereferenced Information Procession System） 可以說是第一個對於空間資訊檢索進行研究的計畫，目的在於根據文件內容的解析，賦予每份文件一個空間位置（Vestavik 2008），該計畫先建立好一個地名詞典（gazetteer），在這個詞典中，包含有地名與顯著的地標名，然後使用一個三階段的演算法進行索引，1）首先由被索引的文件中將具有空間敘述（spatial context）的地名或片語找出來，2）接著再將它們與地名詞典進行比對，將詞典中所記錄的空間座標位置，賦予每一個地名或地標，3）最後根據所找到的位置，進行位相的計算，產生多邊形，當面積超過某一設定值時，就把這個位置當做是該文件的代表位置（Woodruff & Plaunt 1994）。除了以地名詞典進行對應的方式外，另有些計畫，如SPIRIT（spatially-aware information retrieval on the internet），則以建立地名詞庫（thesauri） 或是包含地名文字描述與空間階層結構概念的本體論（ontology） 來進行空間資訊的檢索，它使用了一組空間查詢運算子（Vaid et al. 2005），例如, near運算子用以判斷距離上的鄰近關係，東、南、西、北用以判斷方向的空間關係，該做法與Larson（1995）所提出的概念類似，透過這樣的方式，用以改善在網際網路上的搜尋能力（Purves 2002）。geoXwalk 計畫是一個提供英國與愛爾蘭地名辭典服務的計畫，該服務透過網路應用程式介面（API） 提供查詢的介面，使用者可以透過該服務API上傳文件來進行解析，將文件中所內含的空間地名解析出來，再根據其地名辭典，對所解析出來的地名進行對應查詢，找出各地名的空間座標位置（Reid 2003）。 GeoVSM（Geographic Vector Space Model） 整合文件內容的座標資料與文件關鍵字，建立文件的向量空間模型，分別根據空間敘述與主題語意進行相似度的度量，並綜合其度量結果，以期找到符合空間與主題條件的搜尋結果（Cai 2002）。

二、現有搜尋引擎在地理資訊檢索之提供

目前各著名的搜尋引擎也開始提供所謂的在地網頁搜尋（Local Web Search）的服務，Google、Yahoo!、Ask Jeeves、 MSN等等，均提供這樣的服務，以地圖結合超連結的方式來呈現搜尋結果。Google的在地網搜尋服務（Google Local） 即透過Google Map來呈現地圖（http://local.google.com），它的地理資料庫中，包含了相當多的旅館、商店、餐廳等商業方面的資料，與Google 原先搜尋引擎不同的是，他相當仰賴與提供城市資訊資料庫的網站結合，如CitySearch（http://www.citysearch.com）、 WCities（http://www.wcities.com）與台灣的Kingway（http://www.metamap.com.tw），也就是說，Google Local其實更像是一個提供城市導覽的網站，而不是一個地理資訊檢索的網站，此外，它也提供了以Web 2.0觀念，供大眾上傳與分享空間訊息的分享平台，不過仍然沒有包含網路上的網頁資料。MSN City Guides（http://local.msn.com） 以及Ask Jeeves Local也提供類似的城市導覽網站。Google Earth則是另一種不同方式的網路地理資訊服務，它結合了非常精細的衛星影像與地理資訊，所不同的是，他並沒有直接連結到網頁內容，而是提

供一個Web 2.0觀念下，供大家上傳分享地理訊息的載台。雅虎（Yahoo）的Local Maps（http://local.msn.com） 也是一個與Google Map類似的服務。美國線上（AOL） 的AOL Local 則把本身網站所擁有的豐富資訊整合進來，包含了當地事件以及電影訊息（http://localsearch.aol.com）。另外，人們常常會將他旅遊或日常生活與空間相關的訊息，記錄在部落格中，這些內容正好可提供我們了解某一地區的重要消息來源，因此，就有像DC Metro Blogmap以及NYC Bloggers這樣的服務，提供由部落格中檢索出當地相關的文章。

三、結合語意資訊檢索與空間資訊檢索的相關應用

李俐樑等（2007） 以中文斷字斷詞技術及Web-GIS，由特定旅遊網站的旅遊行程文章中擷取出景點名稱，轉換得各景點的位置，再根據文章中的時程，最後再結合Google Map系統與旅遊網站資源完成景點空間對位，產生一張張的旅遊行程地圖。該研究係針對特定的旅遊網站，內容有一定的固定格式，並未對個別主題語意進行分析，且空間位置係風景區位置，無法找尋個別的點位置，只能由文字找空間語意資料，無法由特定空間範圍檢索出符合主題的語意及空間資料。Tezuka 與 Tanaka（2005） 由認知地理學的觀點出發，以地標（landmark）資訊的擷取為案例，他認為人類的空間行為，主要來自對於空間的認知意像（cognitive image），而不是實體的空間結構，地標往往代表了人們對於地物在空間認知上的顯著性，而網際網路中無數的文件，就相當於人類社會的縮影，因此藉由文件頻率（document frequency）、地名同時出現數（regional co-occurrence summation）以及地名同時出現變異數（regional co-occurrence variation）來衡量由網頁中所找出地標的重要性排序，取代過去在認知地理學研究上以發放問卷為主的調查方法。Tezuka et al.（2006）延續之前的研究，從旅遊部落格網站中，擷取網頁內容中地點、時間、行動以及活動等內容。該系統在空間資訊處理的部分，主要藉由文件頻率、地名同時出現數等方法，以擷取文件中的地名為主，並判斷地名的重要性。然後再藉由地名、時間、行動與活動等四個要素建立關連法則（association rule），以便根據使用者的基本資料與空間資訊進行空間上的判斷與關聯法則推論，主動提供與位置資訊相關（LBS）的旅遊網頁內容之服務。

參、研究方法

本研究所要處理的資料，包含了文字資料與空間資料二大部分，但如何由最原始的文字資料衍生出語意資訊、空間資料、空間資訊以及語意與空間的整合，是我們所要解決的重點。不同於以往，本研究從資料的取得、資料的處理分析到最後成果的展現資料，均是透過網際網路與Web 2.0的方式來建置完成，針對文字語意的部分，除了使用典型中文文件探勘的技術，如中文斷詞系統進行斷詞、向量空間模型來代表文件、以BM25函數來計算文件相似度外，另外，還建置了主題的關鍵字庫以供查詢。而屬於空間資料處理的部分，由之前的相關研究來看，大多只著重在空間語意的分析與擷取，未能再加入地理資訊系統處理空間資料的觀念，對於語意資訊及空間資訊與實際空間位置的整合

仍然有限。為此，本研究首先建立相對應地名詞庫，內容除了地名與編號外，不同地是，再加入一般在地理資訊系統上用來進行空間查詢時常用到的空間物件的MBR座標資料，而非僅是單一點的座標，並且建立地名空間物件間彼此的階層關係，以便藉由正規表示式來擷取地址資料，實踐地名同時出現數的觀念，藉以確認語意的空間關係。最後則透過空間對位與地圖顯示，直觀的將語意資訊與空間資訊整合展示。我們認為透過本研究，可以同時達到文字資料探勘、視覺化資料探勘與空間資料探勘的效果。以下就文字資料處理以及空間資料處理所應用到的各方法在本研究所扮演的角色、關聯以及整合進行說明，研究設計架構如圖1。

一、文字資料的處理

文字資料檢索的部份是以中文語意為基礎，透過文字探勘，計算所擷取的網頁內容與查詢主題的相似度，篩選出語意相似度較高的網頁，接著才能網頁內容中的空間資訊進行檢索與篩選。研究方法除文字探勘相關理論方法如向量空間模型、BM25文件相似度計算函數外，還包含進行中文斷詞的處理。

（一）主題關鍵字庫之建立

探勘的目的與檢索的主題有關，如找民宿、找餐廳、找旅館等，不同的主題，也就會有其對應的重要關鍵字。本研究以民宿為例，經訪談大學觀光相關科系有關教授民宿管理之實務老師與相關網站搜尋後，整理歸納出最能代表各種主題的關鍵字組，並且一一建立進入主題關鍵字庫中（如表1），以民宿為例，可能會包含了民宿、交通、住宿、旅遊、風景、食宿、訂房…等關鍵字，當使用者選取某一主題進行查詢時，也就相當於查閱出該主題所對應的一組關鍵字，提供做為文件與查詢主題在語意相似度上的度量之用。由於關鍵字對於搜尋結果影響甚鉅，目前挑選的方式仍然相當主觀，未來有需要做進一步的探討。

表1：主題關鍵字串範例

主題	關鍵字串
民宿	民宿、位置、交通、住宿、旅遊、風景、食宿、訂房…
旅館	旅館、飯店、位置、交通、會議、設施、訂房、旅遊、美食、套房、預約…
餐廳	餐廳、交通、料理、美味、套餐、價格、營業時間、海鮮、牛排、火鍋、吃到飽…

（二）CKIP中文斷詞技術

進行文件探勘的首要任務在於找出文件中的詞彙，特別是重要的關鍵詞，由於我們的標的為中文網頁，中文不同於歐美語系的文句中，詞彙與詞彙間有空白格開，因此對於詞彙的擷取，相對上比較容易，而亞洲如中文、日文則因詞彙間彼此並沒有明確的空白間隔，很容易造成詞彙辨識上的模稜兩可（Kanada 1999），因此就資訊檢索工作而言，相對上比較麻煩，而資訊檢索的首要工作，如VSM、TF、DF、IDF與BM25等的計

算，都必須先能夠順利的將文件中的詞彙找出來。為了提高擷取的正確率與開發成本，本研究關於詞彙擷取的部分，以中研院資訊科學研究所詞庫小組（CKIP） 所開發的中文斷詞系統為工具，該系統為結合詞庫式斷詞法及統計式斷詞法之優點的混合式斷詞法，將使用者所輸入之文章或文句自動斷詞後，再標示出每個詞的詞類標記。根據實驗數據顯示，其系統斷詞的精確度為百分之九十以上。此一系統包含一個約拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題（May & Chang 2003）。 而CKIP除了提供線上展示系統供一般大眾使用外，該斷詞服務亦開放API供一般用戶撰寫程式叫用，資料的交換方式採用 XML。本研究撰寫TCP/IP Socket 連線程式，傳送驗證資訊及文本至中研院的斷詞伺服器上，伺服器經過處理後經由原連線傳回結果。此線上服務為上述提及之斷詞系統的簡化版本，僅提供以基本詞典進行斷詞，並輸出簡化之詞類標記。 舉例將一句子：「我下禮拜將前往墾丁參觀海生館」，以XML格式送往其伺服器，將會回應：「我（N）下（DET）禮拜（N）將（ADV）前往（Vt）墾丁（N）參觀（Vt）海生館（N）」，括號中的N代表名詞、DET代表特指定詞、ADV代表副詞、Vt代表動詞。有了這些資訊之後，才有可能進行後續的詞彙分析。

（三）向量空間模型（Vector Space Model-VSM）

有了文件中的詞彙組合後，接下來須對每一文件進行編碼的動作，透過自然語言處理，將文件中的重要詞彙找出來，並且賦予每一詞彙不同重要性的權重值（Boguraev & Neff 2000；Gey et al. 2006），以向量空間模型來代表一組網頁，也就是說，每一份網頁都是由一組具有權重值的關鍵字向量來代表（Salton & Buckley 1988），表示的方式如下：

$$D=(t_i, t_j, ..., t_p)$$

D代表被查詢的文件，每一個t_k為代表這份文件的詞彙。同樣的，查詢的部份可以視為

$$Q=(q_a,q_b,...,q_r)$$

Q代表欲查詢的文件或查詢條件，每一個q_k為代表欲查詢文件或查詢條件的詞彙或關鍵字。

接下來進行詞頻（Term Frequency-TF）、文件頻率（Document Frequency-DF）與IDF（Inverse document frequency）的計算，其中，TF指詞彙t_k在文件D中出現的頻率，值越大表示，代表此詞彙對文件D越重要。

另外，如果只用詞句出現的頻率來判斷某一篇文章裡面最重要的關鍵字，我們可能會找到常用字，而不是最重要的字，像是英文裡面的 "the"、"a"、"it"，都是常常出現的字，但是通常一篇文章裡面最重要的字不是這些字，因此會使用IDF進行修正。IDF可由下列的式子表示（Robertson & Zaragoza 2007；Pérez-Iglesias 2008）

$$IDF(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \quad (1)$$

N代表文件集合中文件的數目，$n(q_i)$代表文件集合中有多少篇文件包含q_i這個詞彙，當IDF值越大，表示詞彙在越少的文件出現，對文件D越具代表性。經過這樣的處理

後，可以將每一份網頁轉化為由一組具有權重值的關鍵字向量來代表，以便進行後續的相似度計算與搜尋。

（四）BM25相似度計算函數

藉由將文件以關鍵字向量表達後，即可計算查詢文件與查詢條件間的相似度值以及其相似度的排序，挑選出在主題語意上比較相似的文件，作為查詢的結果。最常見的有以TF * IDF權重值之向量空間模型的cosine函數及BM25相似度計算函數，用來計算文件語意上的相似度。由於本研究是以查詢條件之關鍵字，對檢索文件進行語意相似度的計算，查詢的條件相對較短，若以cosine函數計算，需先進行整篇文件與文件間的交互關係比對之後才再計算權重值的矩陣，將較為複雜，使用BM25將可避免這樣的不便，故在本研究中，關於主題語意與查詢條件間相似度的度量是以BM25做為計算上的依據。BM25為Robertson et al.（1994）所開發，與cosine函數不同的是，它根據的是機率的檢索架構，用於度量檢索文件與查詢條件之相似性的計算與排序，是目前最新與成功的檢索方法之一。其簡化公式如下（Hawking et al. 2004；Lin et al. 2005； Robertson & Zaragoza 2007；Perez-Iglesias 2008）：

假設一個查詢條件 Q, 包含了$q_1,...,q_n$等關鍵字，文件D的 BM25 得分為

$$score(D,Q)=\sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.(1-b+b.\frac{|D|}{avgdl})} \qquad (2)$$

f（q_i,D）為查詢關鍵字q_i在文件D中的詞頻（TF），｜D｜為文件D中的詞彙數目，$avgdl$為所有文件集的平均詞彙個數， k_1與b為自由參數，b介於0～1之間（Robertson & Zaragoza 2007），有關k_1與b之設定，將於5-2節另外作說明。IDF（q_i）為查詢關鍵字的IDF（inverse document frequency）權重值，透過BM25分數的計算，可以賦予每篇文件一個BM25的分數，並且進行排序，挑選語意上最相近的文章。

二、空間資料的處理

存在於文件中的地理名稱或地理觀念若未經空間資訊的處理，則只是一般的辭彙，無法與空間建立關係，如此空間資訊將被隱藏於文件中，為了凸顯文件的空間意涵，我們對於滿足語意相似的文件再進行了下列的空間資料處理：

（一）地名詞庫（Gazetter）之建立

我們先以地理資訊系統（GIS）收集與地名相關的地理資料庫（如台灣地區鄉鎮行政區的GIS圖層），其中包含有用於地理資訊系統空間搜尋的各行政區最小包圍矩形（Minimum Bounding Rectangle - MBR）的左上角的X座標、MBR左上角Y座標、MBR右下角X座標、MBR右下角Y座標。過去的研究對於座標的記錄，大多只記錄地名的單一點座標（例如南投縣就記錄該行政區的中心點座標），這樣的做法不利於空間搜尋以及空間範圍查詢。我們將MBR資料納入地名詞庫中，使該資料庫包含地點的編號、地點名

稱、該地點MBR座標等欄位、該地點所屬的空間結構階層（例如，縣市為1、鄉鎮區為2等）以及該地區之上一層空間結構地點的編號（用以識別所屬的父結構，例如金山鄉，它的空間階層編號為2，它的上一層結構則為台北縣）。有了這個資料庫後，即可以地名來搜尋MBR，或是以目前搜尋範圍的MBR來找出在此範圍上為包含了哪些地點，並且了解該地名在空間階層上所處的關係，用以從網際網路中找尋包含相關位置的網頁資料。目前本研究僅先針對縣市、鄉鎮這二層級進行資料建置，未來可陸續建置更完整的資料庫。

（二）地名同時出現數（Regional Co-occurrence Summation）與空間敘述片語

Tezuka et al.（2006）在針對地名詞彙使用IDF或TF進行字詞分析時發現，現有的方法無法分辨該地名是否應用在具空間敘述的語意上，因此，常常會有像企業分部、大學、或連鎖商店得到了很高的分數，例如新竹銀行、台北大學等，這樣的語彙，其實他們並不具空間的意義，但常常會被誤判為具有空間意義，其他如包含了地名的道路名稱也常易誤判，如天津街，被誤判為天津。為了避免這樣的結果，他認為文件中具空間敘述的地理詞彙，通常都會與他附近的地名或所屬的地理階層共同出現，正如同文件探勘中的應用，詞彙的共同出現頻率，也被視為是一項重要的度量數值（Salton 1968；Rijsbergen 1979），因此提出了地名同時出現數的度量。透過這樣的方式，不僅更能掌握文件中空間敘述的部份，而且對於一些模稜兩可的歧義地名也具有解析的效果，例如，台中縣的東勢與雲林縣的東勢，台北市大安區的大安國小與台中縣大安鄉的大安國小，如果文件能找到相鄰的地名或地理階層關係，那麼就能在空間上區隔出來。另外，他也認為地理詞彙若搭配相關的空間觸發片語（spatial trigger phrase），那麼就能更明確的判斷空間的意涵，改善搜尋結果的正確率。

但由於當初Tezuka et al.（2006）在進行這項度量時，是以地名間彼此的空間距離來做判斷，這樣的做法須仰賴大量地名資料庫的建立，不適用較大的範圍，本研究應用這樣的觀念，但不進行距離的度量，而是從地名的空間階層來判斷，若具有空間階層關係則可視為是空間敘述。文件中的地址資料就包含有這樣的特性，地址中的地名通常會伴隨有其相關臨近的地點或是所屬的地理階層結構，例如，大安鄉伴隨著台中縣這樣一個地理階層出現，就可以十分確切的認為這是一個空間語句，再者，地址中的路、巷、弄與門號，我們可以把它視為是空間觸發詞，伴隨著這樣詞彙的出現，也能加深我們對於地理資訊的掌握，因此，文件中的地址資料，可以說是辨識地理位置的一個最好的資料來源，以民宿為例，民宿網頁必然包含有該民宿的地址，透過對於地址資料的解析，有較大的機會能掌握該文件直接的地理資訊，而且可適用於全國，而不必建置大量的地名資料與距離計算。

（三）空間位置資訊之檢索

本研究對於空間資料的解析，所不同的是以地址資料作為解析的對象，由於地址資料通常是屬於一種自由格式，以台灣地區而言，尚無一個通用的標準格式，但它卻又符合部分的模糊規律，隱身於網頁內容中。面對網頁這種格式鬆散的文件，我們使用正規

表示式（Regular Expression）以適當組合的過濾器對於地址資料進行模糊解析，當作是否為空間資料的推論依據。

本研究所使用的模糊搜尋字串pattern為：

〝\w{2}[縣|市](\d\d\d)?\w{2,3}[市|鎮|鄉|區]\w{1,20}\d*-?\d+號〞

在這個比對字串中，縣市及鄉鎮區可以視為是同時出現的地理階層結構，透過正規表示式取得後，可以再藉由地名詞庫確認其階層關係，是否為合法的住址，而門牌號碼則可以視為是空間觸發詞，不需要考慮樓層，只要取得住址門牌號碼即可找出位置。同時掌握了這幾項要素，透過模糊比對，即可擷取出台灣地區類似地址的資料，然後再藉由地名詞庫中的空間階層來判斷地址的合法性。

（四）文件與空間關係的對應

找出文件中的空間資訊後，若只以文字來描述這些資訊是不夠的，過去的研究大多僅止於此，既然是空間的資訊，當然最好的展示平台就是地圖，若能與地理資訊系統結合，一來可以很直觀的將文件與空間位置建立對應，二來透過地圖的展示，也可顯現出這些文件的空間分布，一個適當視覺化展現，會比更多的文字描述有價值，因為從圖像擷取有意義的資訊，是人類所擅長的。若能再以地理資訊系統進行空間分析，將可進一步的探討這些資訊的空間規律，對於未來應用於類似的商業行銷、商圈分析或適地性服務（LBS）的提供甚具潛力。

對於本研究而言，Google Map擔任類似於地理資訊系統的角色，它的地理資訊功能係透過網際網路叫用Google Map地圖服務所開放出來的應用程式介面（Google Map API）所達成。在本研究中，它提供了以下的功能：

1. 空間對位（Geocoding）：擷取出來的地址資料若未能將它與實際空間位置建立對應，則仍只是文字資訊。而空間對位係指將住址資料轉換為實際空間上的位置，為地理資訊系統的重要功能之一，也是本研究中建立文字資訊與空間資訊對應的重要關鍵。在輸入地址的文字資料後，該服務可以根據住址的行政區、道路與門牌號碼等空間資料，進行空間內插，計算出該地址可能的地理經緯度，接下來可用於民宿空間位置的展示，將網頁內容的空間資訊與實際空間位置進行對應。本研究除了將空間對位的結果展示於地圖，達到視覺化資料探勘之外，再藉由AJAX的程式將這些位置資料儲存於資料庫中，做為空間資料搜集的方法，以便進行後續商業上的地理分析。
2. 開發平台：Google Map是一個典型Web 2.0的產物，可以使用一般的混搭（mash-up）網頁撰寫技術開發，方便與本研究的AJAX技術整合，提供較一般商業網際網路地理資訊系統更為方便與快速的操作經驗，無須於第一次瀏覽時安裝一個地圖瀏覽程式，簡化了安裝與操作。
3. 操作與展示平台：地圖服務中已包含了全國的基本地圖資料與高解析度衛星影像，無須再另行購買地圖資料，且該服務會定期更新地圖的資料。可以直接將搜尋結果與背景地圖完全整合，或在直接在地圖上進行查詢操作。

三、資料的搜尋

（一）主題語意與空間搜尋

以所謂的聚焦式（focused search）搜尋將搜尋主題限制在特定的空間範圍與主題上，而不是全面廣泛性的搜尋，以提升搜尋的效率，並且將空間與主題語意的查詢進行結合。以主題語意查詢為而言，使用者以所挑選的主題搭配地點為條件，由網際網路上蒐集資料，它的搜尋格式為「地點 主題關鍵字」，例如「高雄縣 民宿」。並且從主題關鍵字資料庫中把該主題相關的關鍵字組找出，以這些關鍵字組再搭配由網際網路上查詢到的文件集合，進行文字探勘的相似性度量，篩選出與主題語意相近的文件。

就空間搜尋而言，以下列二種方式進行聚焦式的搜尋，

1. 以地點文字：以縣市名稱為篩選條件，透過第名詞庫的查閱，將該縣市的MBR找出，以便測試所篩選出之空間資料符合所限制的空間範圍，並且將該縣市名稱看為是另一種關鍵字，把該地名與主題關鍵字共組在網頁搜尋的關鍵字群中。
2. 以邊界範圍（region query）：此為本研究的特殊設計，當要搜尋的地點跨越好幾個縣市，這時如果只單以某縣市名稱作為搜尋的地點關鍵字，搜尋效果有限，例如，對玉山國家公園地區有興趣，但是玉山國家公園卻橫跨了南投、花蓮、嘉義、高雄等縣，此時就不再適宜使用地點文字做為搜尋的方式。如前例，可以將目前的地圖範圍調整到玉山國家公園的範圍或框選地圖上的某一地區，以該範圍（Extent）為查詢區域，也就是相當於該範圍的MBR，透過這樣的方式，可以得到此一矩形的座標，再將此一矩形座標與地名詞庫中各縣市的MBR進行比對，如果彼此的MBR有交集，則代表此查詢包含了這個縣市，若與多個縣市都有交集，則表示該次搜尋中，將會包含了多個縣市地名成為查詢時的空間關鍵字，而它的查詢條件關鍵字組合將可分為查詢條件，如：（主題X，縣市地名1）OR（主題X，縣市地名2）OR…

（二）超連結走訪（Hyperlink Traversal）

有許多網頁可能在主題上符合搜尋的條件，但是卻沒有明確的空間資訊存在，有時反而可以透過與它相連的超連結，來獲得較為明確的空間資訊，或是由現有包含明確空間資訊的網頁中，獲得更多與它相連結的其他網頁中的空間資訊（McCurley 2001）。基於這樣的觀念，本研究可設定要繼續走訪（traverse） 的超連結層數，以期找出更多的空間資訊。

四、執行流程

本研究在執行上述處理的運作原理如圖2的虛擬程式所示。首先，依據查詢選擇，當選擇的是範圍查詢，則透過Google Map 由目前地圖的查詢範圍（Extent） 獲取查詢的MBR，再將此MBR與地名詞庫中各筆地名資料的MBR進行比較，當彼此的MBR有交集時，則將該筆地名資料加入查詢的地名集合中（程式02～04行）。如果選擇的是由下

拉選單中挑選的地名，則直接將該地名加入查詢的地名集合中，並且以此地名到地名詞庫中，查詢符合該地名的MBR，並以此作為查詢的MBR（程式06～07行）。接下來，以所挑選的主題為條件，例如民宿，至查詢主題關鍵字資料庫中，查出代表該民宿的相關關鍵字組，並且建立一個查詢主題關鍵字集合（程式09行）。再同時以查詢的地名集合以及查詢主題關鍵字為查詢條件，透過呼叫Web Service的方式，叫用Google 所提供的Google Search Web Service，自動取得搜尋的結果，並且將這些符合的網頁，集結成HTML文件集合，這樣的效果相當於以手動方式在Google搜尋網頁中輸入搜尋關鍵字一般（程式第10行）。接著再將這些HTML文件的HTML tag去除，獲得純文字的文件集合（程式第11行），然後再以TCP/IP Socket 連線的方式，呼叫中央研究院的CKIP中文斷詞系統，為每篇純文字的文件進行斷詞的工作，取得該文件的字詞集合，據以建立代表該文件的字詞組集合與向量空間模型（程式第12行）。

再針對純文字文件集合中的每一篇文件，進行字詞集合與查詢主題關鍵字集合的BM25主題相似度計算，以得到該文件與查詢主題相符程度的得分（程式第15行）。如果相似度得分的排序名低於所設定的排序門檻值（如排名前300或前500高者），則透過正規表示式（Regular Expression） 以住址擷取的樣式（Pattern）字串作為模糊（fuzzy）比對依據，從文件中把地址資料擷取出來，建立成地址集合（程式第17行）。如果未能在該文件中找到任何地址資料，則可以根據該HTML文件中的超連結標籤進行走訪，以遞迴方式續探連結的網頁是否包含有符合的資料（程式18～20行）。

接著，再針對地址集合中的每一個地址字串資料，判斷該地址是否已經存在於標示過的集合中（程式22～23行），若未存在，才透過AJAX的呼叫方式，叫用Google Map API 所提供的空間對位（geocoding）功能，把地址字串資料轉換成該地址的經緯度座標（程式第24行）。本系統除了比對主題語意上的相似度外，還需進行空間鄰近（proximity）上的篩選，使查詢出來的位置必須坐落在查詢區域內，因此，針對地址的經緯度與查詢的範圍（MBR） 進行落點測試，判斷該地址經緯度是否坐落在查詢範圍內（程式第25行），如果地址經緯度坐落在查詢範圍內，則將該地址位置繪於 Google Map 的地圖上，以 Icon 表示，並且建立超連結的關係，將地圖與超連結的HTML文件建立關聯，以方便搜尋結果的檢視（程式26～27行）。

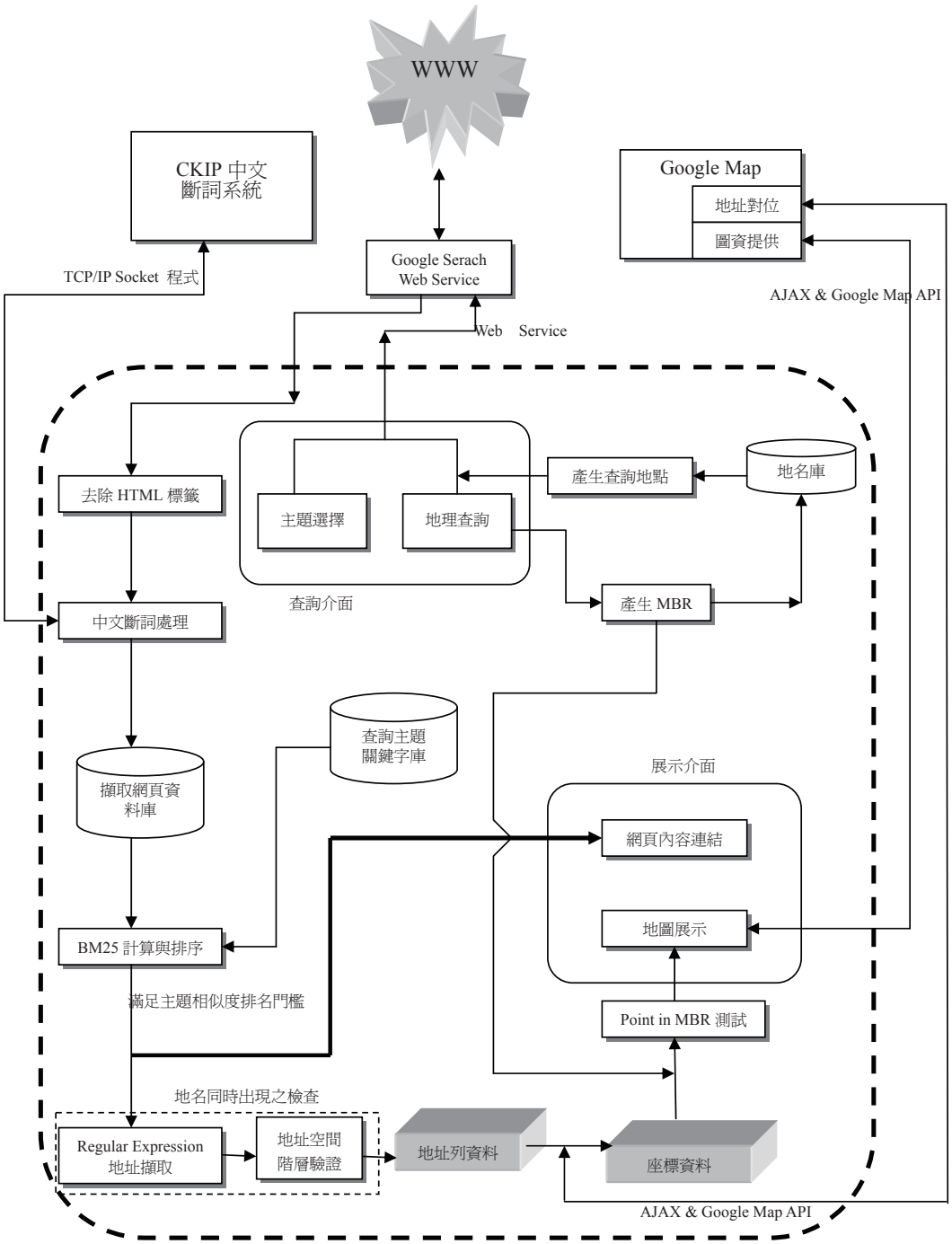

圖1：地理資訊檢索系統架構

```
01  //Geographic Range Query Selection
02  if Query By Map Extent then
03     QueryMBR = Current Map Extent ;    //MBR: Minimum Bounding Rectangle
04     QueryPlaceNameList = LookupGazetteerForPlaceNameList(QueryMBR);
05  else    //Query By Place Name
06     QueryPlaceNameList .Add(UserInputQueryPlaceName);
07     QueryMBR = LookupGazetteerForMBR(UserInputQueryPlaceName);
08  end if
09  QueryKeywordList = LookupThematicKeywordDB(UserInputQueryTheme);
10  HTMLDocumentList = CallGoogleWebSearch (QueryTheme, QueryPlaceNameList);
11  PureTextList = RemoveHtmlTag(HtmlDocumentList);
12  DocumentList = CallCKIPWordSegmentationAndBuildVSM(PureTextList);
13  for each Document in DocumentList do
14     AddressStringList = empty;
15     Score = BM25ThematicCalculation(DocumentList, Document, QueryKeywordList);
16     if Score > User SetupThreshold then
17        DocumentAddressList = RegularExpression(PureText, AddressPattern);
18        if DocumentAddressList is empty then
19           recursively trace the hyperlink in HtmlDocument;
20        end if
21        for each AddressString in DocumentAddressList do
22           if AddressString not exist in AddressStringList then
23              AddressStringList.Add(AddressString);
24              CoordinatePoint = CallGoogleMapAPIGeocoding(AddressString);
25              if    CoordinatePoint in QueryMBR then
26                 Get Google Map and Draw CoordinatePoint;
27                 Associate CoordinatePoint Icon with HtmlDocument;
28              end if
29           end if
30        end for
31     end if
32  end for
```

圖2：地理資訊檢索系統程式流程虛擬碼

肆、實驗設計與結果評估

本研究以ASP.Net為實驗的建置工具，相關模組包含語意檢索模組、地理資訊檢索模組、中研院斷詞系統之TCP/IP Socket連線模組、Google Web Service 、Google Map API程式等幾個大模組。以地圖為整個系統的操作與查詢界面的中心，文件內容與其內容中地

理資訊的空間定位，透過地圖上的超連結，將彼此連結起來，也就是 Scharl（2007）所稱的空間網（Geospatial Web）的概念，所有的資訊與操作均與地圖相連結（圖6）。實驗的設計是以Google搜尋引擎與Google的Local Search做為語意與空間部分測試比較的平台，設定多個搜尋關鍵字串交由Google Search 與 Local Search 進行搜尋與實驗，並比較其結果。

一、BM25參數的設定

本研究關於公式2中k_1與b的設定，由於網頁文件在內容上相對較為簡短，大部份查詢的詞頻（TF）數均在7以內，經以固定的文章數目（10000篇）與包含個別查詢關鍵字的文章數（20篇）為基準，且假設受測文章為一般的平均長度來進行實驗（如圖3），再參考了相關的實作（Hawking et al. 2004；Lin et al. 2005；Andrade & Silva 2006；Perez-Iglesias 2008）後，發現k_1與b的設定對於個別關鍵字的BM25得分值最後均趨近于收斂（當$k_1=0$時），在低詞頻的情況下對於BM25的得分並沒有太大的差異，因此設定$k_1=2.0$、$b=0.75$。未來對於具有較長詞頻且不同類型或主題的文件來源，例如非網頁文件之文獻或報告，則有需要做進一步的探討。

二、語意檢索評估

在語意檢索評估上，本研究藉由專家的建議，由全省北、中、南、東四個地區挑選出9個縣市（如表2），分別與「民宿」、「餐廳」、「旅館」三個主題進行27種組合實驗，得到像「縣市 主題」等不同區域、不同主題做為搜尋的關鍵字串，與Google search進行比較，並將所搜尋到的網頁予以存檔，以便進行實驗的評估。由於組合眾多，而本研究主要的關注點在於南部地區的民宿發展，特別是高雄縣的部分，因此以高雄縣作為說明的示範案例。我們以 Rijsbergen（1979）對資訊檢索所用的評估指標—精確率（Precision-P）、召回率（Recall-R）與評估方法做為評估的依據，由於精確率與召回率的變化與網頁的BM25相似度評分排序後，所擷取用以進行判斷的排名前面 k 個數目的大小有關，也就是說當 k 值較少時，會有較高的精確率但是較低的召回率，當 k 值較高時，會有較低的精確率但是較高的召回率，k 值可以由使用者自行設定。限於人力及時間，我們以人工進行語意相符上的判斷，先以Google Search 所得到的500篇網頁進行實驗，因此將 k 設定在 1～500 之間，並且使用P-R曲線來分別表達實驗的結果，如表3～表8以及圖4、圖5。評估指標計算的方式如下：

$$Precision=\frac{N_P \cap N_S}{N_S} \quad (3)$$

$$Recall=\frac{N_P \cap N_S}{N_P} \quad (4)$$

N_S：系統排序後前面 k 篇數，k 在 1～500 之間。

N_P：500篇文章中經人工判斷符合題意的篇數。

$N_P \cap N_S$：共同被系統與人工同時判定符合題意的篇數。

由於目前旅遊風氣的盛行，因此不論本研究或是Google所搜尋出的結果，大多與主題的關聯性很高，進一步分析比較搜尋的結果，得到下列心得：

應用：

1. Google所搜尋的頁面中往往包含了太多的資訊，例如，把非相關主題或非相關的地理資訊也收納在其中，不利於使用者的資訊吸收，相較而言，本研究的搜尋結果更貼近使用者的需求。
2. 本研究所搜尋到的結果，會具有更詳細的文字說明，且帶有明確的空間資訊，對於使用者而言，可以獲更多寶貴的資訊。雖然僅係以地址資料進行空間對位，所得難免有誤，但經過語意方面的內容比對後，可以降低錯誤的發生。
3. Google因缺少空間資訊處理的功能，對於跨區域或是概念區域（如「大台北地區」）的範圍查詢，有時很難以關鍵字或地名進行表達，本研究可以藉助查詢區的MBR或地名詞典達到這樣的效果。
4. Google搜尋所得到的結果，僅是文件內容的超連結，並未具有地理資訊的另外處理與展示。本系統則除了提供文件超連結之外，還具網頁內容空間資訊的地圖展示與位置座標儲存，透過這樣的方式，可以視覺化的進行空間分佈趨勢之知識發現（knowledge discovery）的參考依據。
5. 由表3～表8所示，前100排序的精確率在區域上差異不是很大，但當以前200排序的精確率來看，可以發現我們認知中觀光較熱門的縣份，如北部、東部，明顯的較中、南部縣份排序高，確實符合Tezuka 與 Tanaka（2005）所謂認知地理學中人們對於空間的認知意像。

限制：

1. 本研究對於主題關鍵字的挑選具有較高的敏感性，例如在「縣市 旅館」這個主題的查詢下，不論精確率、召回率均表現較不理想（如表4、表7與圖5），原因是「民宿」與「旅館」具有類似的關鍵字串，因此很容易在搜尋的網頁中雜有類似民宿的資料在其中，但關鍵字串的訂定往往帶有相當程度的主觀性，因此未來對於關鍵字串的定義應更具鑑別性與客觀，此外，也可以考慮針對於個別特具代表性的關鍵詞給予較高的權重，以增加該文的排名，提升鑑別度。
2. 部份在Google 搜尋中具有很高排名的網頁，卻沒有出現在本研究的搜尋結果中或是排名較後，因為本系統結合了語意的相似度的評估，因此，對於僅做總體民宿列表，或是僅提供查詢介面的網頁，得分較低，但這通常也是很重要的資訊來源，此為本研究的一大限制。
3. 對於僅包含Flash或是圖片的首頁，因文字內容較少，故排名得分也較低會，特別是現在有相當多的網站為了達到多媒體包裝的效果，其文字內容或住址資訊已成為圖片的一部分，而不再是純文字，如此將影響搜尋的成效。
4. 在Google通泛搜尋所找到的網頁中，有相當多是來自於部落格文章，這方面的文章大多是寫些出遊歷者的親身經驗與感想，內容較中肯，少商業宣傳，更重要的是，每一篇文章都會搭配有時間的標示，對於某些具有季節性的景點而言，更具參考價值。然而這些文件中，有相當大的比例未包含較明確（explicit）的空間資

訊，而是內隱（implicit）的空間資訊，本研究在搜尋條件限制下，反而容易忽略了這樣的文章。因此，對於內隱空間資訊與時間的擷取，將是日後發展應考慮的重要因素。同時，也應提供查詢者在對於主題語意與空間敘述二條件下，可以微調權重比例的機制，期以篩選出多樣化的結果。

三、空間資料檢索評估

另外，再針對空間資料的搜尋與展示部分進行評估，本研究以Google Map 的 Local Search 做為本實驗在空間資訊搜尋上效能評估的對象，由於Google Local Search具有中文化的資料，因此以它做為比較的平台。而交通部觀光局所建置的觀光旅館業管理資訊系統（http://hscc.tbroc.gov.tw/）與旅館業及民宿管理資訊系統（http://hscc.tbroc.gov.tw）網站內詳列了全國各縣市合法登記之觀光旅館、旅館以及民宿的資料，可以做為評比數據的依據，因此我們以「民宿」及「旅館」二主題做為空間資訊檢索效能測試的主題，查詢的條件約略高雄縣的地圖顯示範圍，進行「民宿」及「旅館」這二個關鍵字的查詢。

結果正如本文第二節所述，Google Local Search與提供城市商業資料庫的網站或廣告黃頁結合，因此在高雄縣的約略地圖範圍內，只找出28家民宿，遠低於Google Search所搜尋出來數目，若再與交通部觀光局旅館業及民宿管理資訊系統網站的所提供的官方正式資料相比，高雄縣合法登記有案的民宿就已達到51家，因此若僅依Google Local Search想要獲得完整的資料，效果將會相當有限，它主要仍是著眼於商業的結合。而本研究在進行了網頁探勘與地理資訊檢索後，則可以找出了包含未登記在案的民宿資料至少60家以上。

而關於「旅館」這個主題，若依交通部觀光局所載的資料顯示（http://hscc.tbroc.gov.tw），高雄縣共有旅館業者113家（1家觀光旅館業者及112家一般旅館業者），Google Local Search所找出來的家數為104家（其中包含有民宿），若扣掉民宿業者後則為83家，而本研究只找到了71家，召回率僅63%，效果遠不如Google Local Search以及「民宿」的這個主題，經分析後發現，部分規模較小的一般旅館業者並未設有網站，因此無法分析到該旅館的資料，而民宿則因拜旅遊風氣盛行之賜，有許多的民宿網與個人部落格推薦或引用，反而更容易找到相關的資料，因此，這樣的主題有助於本研究的評估。

伍、結論與建議

本研究以網頁中的民宿主題為例，透過文字探勘（Text Mining）的技術，對於搜尋的網頁，進行空間敘述（Spatial Context）與語意內容的探勘、檢索與排序，找出與所查詢主題在內容與地理資訊上最相關的網頁。並透過地理資訊檢索與正規表示式，由篩選過的網頁內容中，檢索出有用的地理資訊，如地址，再以空間對位的技術，將檢索出來的地理資訊與文件內容結合顯示於地圖上。透過這樣的方式，發現在加入了地理資訊的檢索處理之後，能夠有效改善對於空間主題相關搜尋的效果，也降低了開發的成本，同

時達到文字資料探勘、視覺化資料探勘與空間資料探勘的效果。

惟在本研究中所使用的地名詞典與主題關鍵字庫，目前仍然相當有限且不夠標準，若定義不良將會造成效能的降低，未來應結合專家建構更廣泛與標準化的空間知識本體論（ontology）與關鍵字串，除了可以找出明顯（explicit）的地理資訊外，還可以把內隱（implicit）的空間資訊也找出來，以提升及擴大搜尋的效果。此外，受限於資料來源的取得，本研究目前只能針對網頁資料進行實驗，對於越來越流行的多媒體化包裝的商業網站的搜尋，效果將會大打折扣，對於這樣的趨勢，如何做更進一步的網頁探勘，將是未來的一大挑戰。再者，有關空間關係的探討，目前本研究主要是以落點測試（contained in）作為主要的空間關係測試，未來可做更多空間關係的擴展，如方向、位相、鄰近等空間關係。由於Web資料的日漸累積增加，其實已可視為人類社會意見表達的縮影，包含了相當多人們對於空間以及主題認知的描述，因此，從認知地理學（Cognitive Geography）的角度來進行資訊的檢索與探勘，將是一個值得努力的方向。

儘管如此，但以目前的方法與成效，針對於非網頁的文字資料仍然具有相當大的潛力。未來可應用於大量空間資訊的自動蒐集，以便進行後續的空間分析，提供有別於傳統地理資訊系統外，在空間資料探勘（Spatial Data Mining）與知識發現（Knowledge Discovery）上新的研究方法與應用，藉以找出文件集合中的空間Pattern或法則規律以及統計趨勢。透過與文字探勘的結合，將可為與空間主題相關的文件內容做檢索、分析、摘要或分類，進行與空間主題相關的知識管理。另外，隨著U化（ubiquitous）行動通訊與行動上網技術的普及，使用者透過行動電話基地台或GPS等裝置的定位，主動傳送使用者目前位置，若文件具有空間上的意義與定位，則使用者將可以及時獲得在地的資訊與服務，有助於所謂的適地性服務（Location-Based Service-LBS） 或在地搜尋（Local Search）服務的推動。

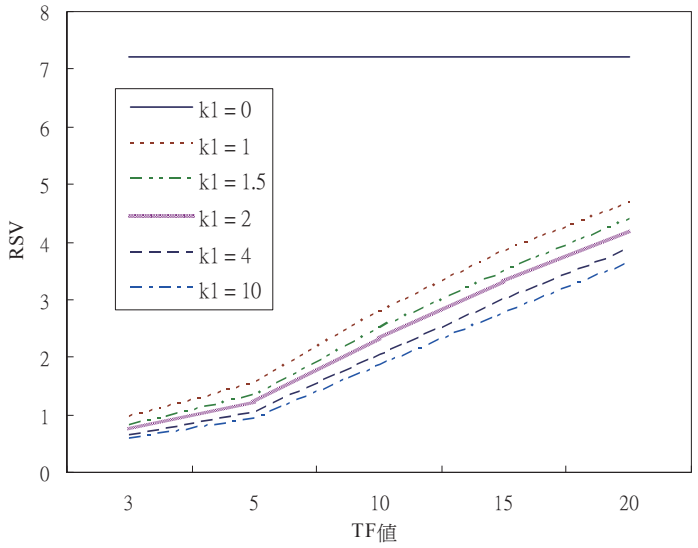

圖3：查詢關鍵字針對不同k1值在不同TF值下所對應之BM25得分

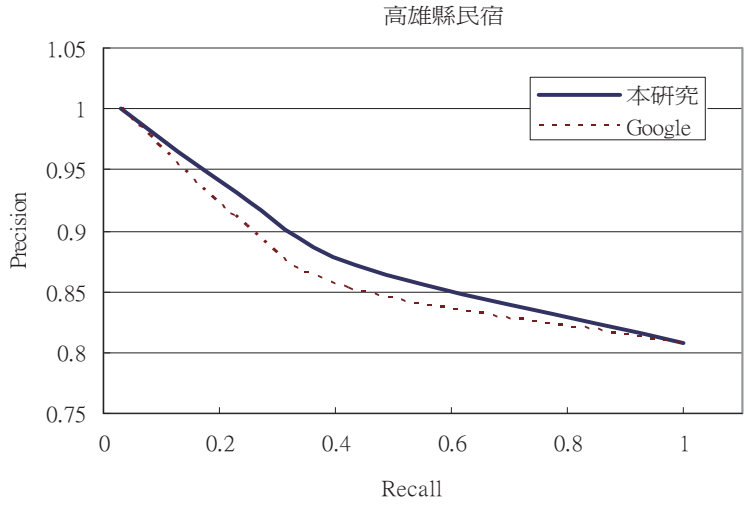

圖4：以「高雄縣民宿」為查詢主題之實驗結果的P-R曲線

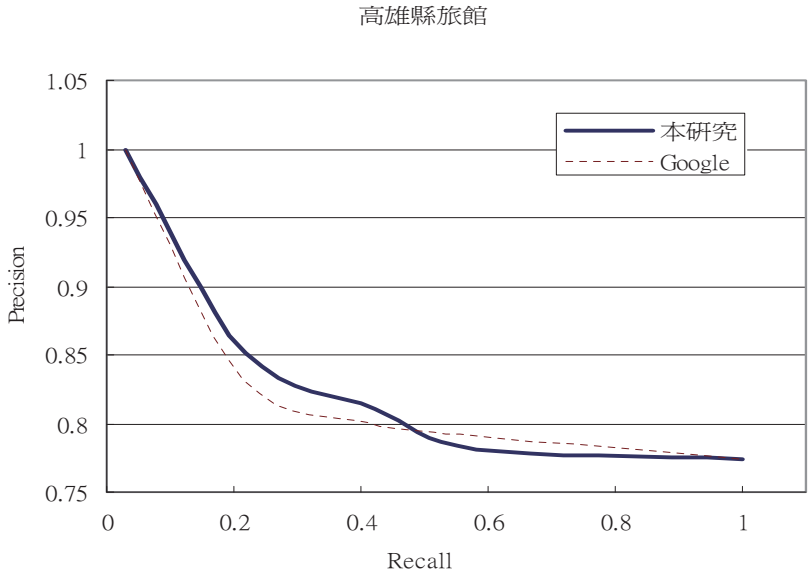

圖5：以「高雄縣旅館」為查詢主題之實驗結果的P-R曲線

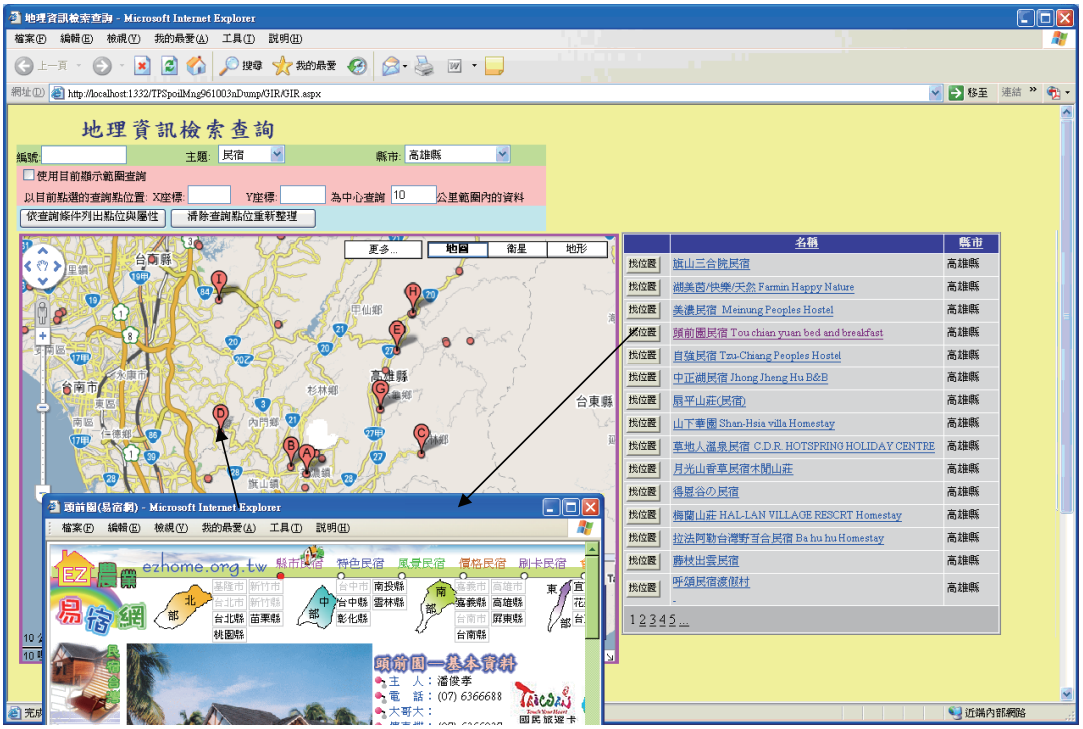

圖6：實驗測試平台

表2：測試縣市

地區	縣市
北部	台北縣
	桃園縣
中部	台中縣
	南投縣
南部	嘉義縣
	高雄縣
	屏東縣
東部	宜蘭縣
	花蓮縣

表3：BM25相似度排序前100之民宿網頁的評估結果
A：本研究，G: Google

縣市		Precision	Recall
宜蘭縣	A	93.5%	24.8%
	G	89.1%	22.3%
南投縣	A	92.8%	25.1%
	G	88.6%	23.2%
台中縣	A	92.6%	24.3%
	G	86.2%	22.1%
花蓮縣	A	92.5%	23.0%
	G	87.4%	20.8%
高雄縣	A	92.3%	23.5%
	G	86.9%	21.1%
台北縣	A	91.5%	24.1%
	G	85.7%	22.2%
屏東縣	A	91.3%	23.6%
	G	85.9%	22.9%
桃園縣	A	90.4%	24.1%
	G	84.8%	21.6%
嘉義縣	A	89.5%	22.8%
	G	84.7%	21.8%

表4：BM25相似度排序前100之旅館網頁的評估結果
A：本研究，G: Google

縣市		Precision	Recall
台北縣	A	87.2%	19.6%
	G	85.1%	20.1%
宜蘭縣	A	86.6%	18.6%
	G	84.5%	19.1%
台中縣	A	86.5%	20.1%
	G	82.7%	18.8%
花蓮縣	A	86.4%	21.7%
	G	83.6%	19.6%
屏東縣	A	86.2%	20.6%
	G	82.3%	21.2%
南投縣	A	85.7%	21.4%
	G	83.1%	19.3%
高雄縣	A	85.3%	21.3%
	G	82.8%	20.7%
桃園縣	A	85.3%	19.3%
	G	82.5%	20.1%
嘉義縣	A	85.1%	20.2%
	G	81.8%	19.2%

表5：BM25相似度排序前100之餐廳網頁的評估結果
A：本研究，G: Google

縣市		Precision	Recall
桃園縣	A	90.2%	20.7%
	G	88.5%	21.7%
高雄縣	A	89.7%	21.8%
	G	88.8%	22.3%
台北縣	A	89.3%	20.6%
	G	90.4%	19.6%
台中縣	A	89.1%	22.3%
	G	89.5%	23.4%
嘉義縣	A	88.7%	21.2%
	G	86.1%	22.1%
屏東縣	A	88.5%	23.3%
	G	85.6%	19.7%
南投縣	A	84.6%	22.6%
	G	85.3%	21.5%
花蓮縣	A	84.2%	24.2%
	G	84.9%	22.7%
宜蘭縣	A	83.2%	22.7%
	G	84.2%	24.4%

表6：BM25相似度排序前200之民宿網頁的評估結果
A：本研究，G: Google

縣市		Precision	Recall
台北縣	A	89.7%	42.7%
	G	88.8%	43.2%
宜蘭縣	A	89.5%	44.4%
	G	87.5%	42.2%
台中縣	A	88.7%	42.8%
	G	88.3%	40.3%
花蓮縣	A	88.4%	44.1%
	G	86.2%	43.6%
南投縣	A	87.5%	43.5%
	G	85.2%	41.2%
高雄縣	A	86.6%	43.8%
	G	84.8%	42.7%
屏東縣	A	86.2%	42.5%
	G	83.2%	40.7%
桃園縣	A	85.6%	43.7%
	G	84.5%	44.8%
嘉義縣	A	84.8%	44.3%
	G	83.5%	41.4%

表7：BM25相似度排序前200之旅館網頁的評估結果
A：本研究，G: Google

縣市		Precision	Recall
台北縣	A	88.5%	43.4%
	G	89.3%	44.8%
桃園縣	A	88.1%	43.1%
	G	87.2%	42.6%
台中縣	A	87.9%	42.3%
	G	88.2%	40.4%
高雄縣	A	87.8%	42.6%
	G	85.7%	41.3%
嘉義縣	A	87.2%	43.1%
	G	85.3%	41.2%
宜蘭縣	A	86.8%	41.8%
	G	82.1%	39.5%
南投縣	A	86.3%	40.2%
	G	83.4%	38.2%
花蓮縣	A	85.8%	41.2%
	G	82.7%	39.9%
屏東縣	A	85.6%	41.9%
	G	83.6%	40.5%

表8：BM25相似度排序前200之餐廳網頁的評估結果
A：本研究，G: Google

縣市		Precision	Recall
台北縣	A	84.3%	43.5%
	G	82.6%	41.5%
台中縣	A	84.1%	42.5%
	G	80.1%	40.2%
宜蘭縣	A	81.7%	42.6%
	G	81.3%	39.9%
花蓮縣	A	83.2%	41.8%
	G	79.2%	39.7%
南投縣	A	82.6%	43.1%
	G	83.7%	40.2%
高雄縣	A	81.5%	41.7%
	G	78.3%	40.1%
桃園縣	A	81.3%	40.6%
	G	77.6%	40.1%
嘉義縣	A	80.6%	41.8%
	G	80.6%	42.0%
屏東縣	A	80.4%	40.2%
	G	77.5%	38.6%

參考文獻

1. 李俐槿、李祐陞、林金龍、黃國倫，民96，『自動旅遊行程空間對位』，2007台灣地理資訊學會年會暨學術研討會，台灣地理資訊學會主辦。
2. Amitay, E., Har'E,l N., Sivan, R., and Soffer, A.“Web-a-Where: Geotagging Web content,”*in Proceedings of the 27th annual international ACM SIGIR Conference on research and development in information retrieval*, 2004, pp. 273-280.
3. Andrade, L. and Silva, M.“Relevance Ranking for Geographic IR,”*in Proceedings of the workshop on Geographic Information Retrieval*, 2006.
4. Boguraev, B. and Neff, M. S.“Discourse segmentation in aid of document summarization,”*in Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000, pp. 10-17.
5. Buyukkokten, O., Cho, J., Garcia-molina, H., Gravano, L. and Shivakumar, N.“Exploiting geographical location information of web pages,”*in Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB'99)*, 1999, pp. 91-96.
6. Byrd, R. and Ravin, Y.“Identifying and extracting relations in text,”*in Proceedings 5th Jt Conference on Information Sciences, JCIS2000*, 1999, pp. 149-154.
7. Cai, G.“GeoVSM: An integrated Retrieval model for geographic information,”*GIScience*,

Egenhofer M.J. and Marks D.M. (eds.), Berlin: Springer-Verlag, 2002, pp. 65-79.

8. Gey, F., Larson, R., Sanderso, M., Joho, H., and Clough, P. "GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track," Working Notes for the CLEF 2005 Workshop, 2006, pp. 908-919.
9. Hawking, D., Upstill, T. and Craswell, N. "Toward Better Weighting of Anchors," *in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 512-53.
10. Jones, C. B., Alani, H. and Tudhope, D. "Geographical Information Retrieval with Ontologies of Place," *in Proceedings of the International Conference on Spatial Information Theory*, 2001, pp. 322-325.
11. Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Kreveld, M. van and Weibel, R. "Spatial information retrieval and geographical ontologies: An overview of the SPIRIT project," *in Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp. 387-388.
12. Kanada, Y. "A Method of geographical name extraction from Japanese text for thematic geographical search," *in Proceedings of the 8th international conference on Information and knowledge management*, 1999, pp. 46-54.
13. Kornai, A. and Sundheim, B. *in Proceedings of the NAACL-HLT Workshop on the Analysis of Geographic References*, 2003.
14. Larson, R. R. "Geographic Information Retrieval and Spatial Browsing," *Geographic Information Systems Patrons Maps and Spatial Information*, Smith L. C. and Gluck, M. (eds.), 1995, pp. 81-123.
15. Lin, K. H.-Y, Hou, W.-J. and Chen, H.-H. "Retrieval of Biomedical Documents by Prioritizing Key Phrases," *in Proceedings of the Fourteenth Text REtrieval Conference* (TREC 2005), 2005, Gaithersburg, Maryland.
16. Martins, B. and Silva, M. J. "A graph-ranking algorithm for geo-referencing documents," *in Proceedings of the 5th IEEE International Conference on Data Mining*, 2005, pp. 741-744.
17. May, W.-Y. and Chang, K.-J. "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff," *in Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 168-171.
18. McCurley, K. S. "Geospatial mapping and navigation of the Web," *in Proceedings of the 10th International conference on World Wide Web*, 2001, pp. 221-229.
19. Mitra, S. and Acharya, T. *Data Mining: Multimedia, Soft Computing and Bioinformatics*, John Wiely & Sons, Inc. , 2003.
20. Perez-Iglesias, J. "Integrating BM25 & BM25F into Lucene," June 2008 (available online at http://nlp.uned.es/~jperezi/Lucene-BM25/)
21. Periakaruppan, R. and Nemeth, E. "GTrace-A Graphical Traceroute Tool," *in Proceedings*

of the 13th USENIX conference on System administration, 1999, pp. 69-78.

22. Purves, R. R., Sanderson, A., Sester, M. M., Kreveld, M. V. and Weibel, R. "Spatial information retrieval and geographical ontologies an overview of the SPIRIT project," *in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 387-388.
23. Reid, J. A. "geoXwalk-A Gazetteer Server and Service for UK Academia," *Research and Advanced Technology for Digital Libraries*, Koch/Solvberg (eds.), Berlin: Springer, 2003, pp. 387-392.
24. Rijsbergen, V. C. J. *Information Retrieval* (2nd ed.), Butterworth, London, 1979.
25. Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M. "Okapi at TREC-3," *in Proceedings of the 3rd Text REtrieval Conference (TREC 1994)*, 1994, pp. 109-126.
26. Robertson, S. and Zaragoza, H. "The Probabilistic Relevance Method: BM25 and beyond (SIGIR 2007 Tutorial 2D)," June 2007 (available online at http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=prm).
27. Salton, G. *Automatic Information Organization and Retrieval*, McGraw-Hill, New York,1968.
28. Salton, G. and Buckley, C. "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management* (24:5), 1988, pp. 513-523.
29. Scharl, A. "Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories," *The Geospatial Web - How Geo-Browsers, Social Software and the Web 2.0 Shaping the Network Society*, Scharl, A. and Tochtermann, K. (eds.), London: Springer, 2007, pp. 3-14.
30. Souza, L., Davis, C. J., Borges, K., Delboni, T., and Laender, A. "The role of gazetteers in geographic knowledge discovery on the web," *in Proceedings of the 3rd Latin American Web Congress*, 2005, pp. 157.
31. Tezuka, T. and Tanaka, K. "Landmark Extraction: A Web Mining Approach," *in Proceedings of COSIT'2005*, 2005, pp. 379-396.
32. Tezuka, T., Kurashima, T. and Tanaka, K. "Toward tighter integration of web search with a geographic information system," *in Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 277-286.
33. Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. "Spatio-textual indexing for geographical search on the web," *in Proceedings of SSTD-05, the 9th Symposium on Spatial and Temporal Databases*, 2005, pp: 218-235.
34. Vestavik, O. "Geographic Information Retrieval: An Overview," June 2008 (available online at http://www.idi.ntnu.no/~oyvindve/article.pdf)
35. Vogel, D., Bickel, S., Haider, P., Schimpfky, R., Siemen, P., Bridges, S. and Scheffer, T. "Classifying search engine queries using the Web as background knowledge," *SIGKDD*

Explorations Newsletter (7:2), 2005, pp. 117-122.

36. Woodruff, A.G. and Plaunt, C. "GIPSY: Geo-referenced Information Processing System," *Journal of the American Society for Information Science* (45:9), 1994, pp. 645-655.