天主教輔仁大學圖書資訊學碩士班論文

指導老師：吳政叡 博士

台灣地區中文網頁自動辨別日期之研究

研究生：邰文暉 撰

中華民國九十九年六月

謝　　辭

感謝父親與母親對我的栽培與耐心，讓我在小學四年級時就接觸個人電腦，當初完全不懂英文時便接觸了 DOS 作業系統，建立了我對電腦資訊最基礎的概念，他們給了我最好的學習環境與自由的學習空間，也感謝姐姐給了我最好的榜樣，讓我能夠依照自己的進度完成這本論文。

這本論文能夠完成，幸得吳政叡老師這一路來的細心指導，從大學時期程式語言的啟蒙，以及進入碩士班遇到瓶頸時的鼓勵，吳政叡老師一直都是最好的顧問，使我獲益良多。當然也要感謝輔大圖資所的陳舜德老師，在大學時期便不斷的用電腦人的角度來教導我如何思考問題並解決問題，給我許多寶貴的建議及鼓勵。在口試階段，感謝淡大資圖所的林信成老師，林老師對文暉的教導，文暉將永遠謹記在心。

在輔大的生活中，也感謝慶華助教及靜宜助教的照顧，幫助我解決許多煩人的問題，讓我可以順利完成碩士班的學業。

謹將這本論文獻給我最想念的爺爺

邰文暉

民國九十九年六月

摘　　要

研究生：邰文暉

系所名稱：輔仁大學圖書資訊學所

指導老師：吳政叡 博士

論文題目：台灣地區中文網頁自動辨別日期之研究

關鍵字：日期格式、網頁日期、自動日期辨別、元資料

論文總頁數：63 頁

摘要正文：

隨著網際網路的日益普及，線上資源也越來越豐富。要精準的為讀者找出有用的資訊，前提是必須能夠精準的分析網頁內容。日期是網頁元資料中的重要欄位，由於台灣在日期格式的書寫習慣，使得中文網頁的日期形式較為複雜，因而增加了在網路資料自動著錄日期時的困難。本研究的主要目的是針對網頁日期部份做深入的分析研究，以便能夠更精確的利用中文網頁中的日期欄位進行檢索利用。

本研究採用實驗研究法，使用亂數抓取繁體中文網頁，分析及統計樣本網頁中出現的日期格式，並使用正規表示式來嘗試自動抓取正確網頁日期，最後計算出正確率。實驗完成之後可以了解在進行中文網頁日期欄位自動辨識會遭遇到的困難，並評估自動擷取繁體中文網頁日期欄位的可行性。

在實驗結果方面，有日期網頁的正確率約為 61%，沒有日期網頁的部分則約為 62%。有日期網頁的平均誤差年為 0.62 年，且誤差年在一年內的網頁約佔

九成。雖然本研究的成果尚未能夠完全取代人工，但若應用得宜仍然可以提高網頁檢索時的效率。

Abstract

Title :
The Study of Auto Extraction of Dates from Chinese Web Pages in Taiwan Area

Keywords : date format, webpage date, auto date extraction, metadata

Abstract :

With the popularization of Internet services, the online resource from the Internet is more plentiful nowadays. ‘Date’ is one of the most important fields of metadata in web pages. Due to the special date displaying formats using in Taiwan, it has made the automatic cataloging on date for webpage more difficult. The major purpose of this research is to thoroughly analyze the different types of date displaying format using in Chinese webpage. These findings will be used to increase the precision on the date auto extraction of webpage.

The procedures of experiment are as follows. Firstly, the sample is randomly from Internet. Secondly, the statistic analysis on the date displaying format of each webpage is conducted. Lastly, Regular Expression is used to abstract the dates of each webpage and the accuracy ratio is calculated. The difficulties and feasibility of auto date extraction are discussed in the end of this work.

The results of the experiment suggest the accuracy ratio of web pages with date information is 61%. On the other hand, the accuracy ratio of web pages without date information is 62%. The average error of those web pages with date information is 0.62 year. The results of this research suggest that the auto date extraction mechanism can be used to improve the efficiency on webpage information retrieval.

目　　次

圖表目錄

第一章　緒論

第一節 研究動機及背景

隨著網際網路的日益普及，線上資源也越來越豐富，資訊的傳遞速度也更為快速，但資訊的快速流通也隨之產生了新的問題—資訊超載[1]。如何在網際網路這個無窮盡的大寶庫中，快速且有效率地找到符合自身需求的資訊，成為了最重要的問題[2]，搜尋引擎便在此時應運而生。

早期的 YAHOO 提供分類目錄服務，主要是一個樹狀主體結構的分類目錄，它讓使用者可以透過預先分類好的類別，過濾掉不需要的類別，再找出符合自身需求的資訊，但隨著資訊科技不斷的進步，這種建立在分類法上面，類似用辭典找字詞的搜尋方式，在效率方面已經不能滿足使用者，便產生了一種新的搜尋方式—全文檢索[3]。

所謂的網頁內容全文檢索服務，就是經由入口網站系統，去抓取 WWW 上的網頁，並建立自身特殊的資料庫供使用者查詢，已成為目前搜尋引擎的主流[4]。但由於網際網路的資訊量實在太過龐大，即使使用全文檢索的方式來搜尋資訊，使用者欲找出真正有用的資訊，仍必須依靠自身的判斷來過濾非相關的資訊[5]。

圖書館是用科學方法，採訪、整理、保存各種印刷與非印刷的資料，以便讀者利用的機構[6]。在資訊爆炸的時代下，隨著科技的進步，雖然透過電腦處理資料的速度比過去的人工處理更具效率，但有效率的檢索，仍是一個重要的問題亟待解決。因此某種形態的電子目錄有其必要性[7]。元資料便是在此一背景下產生並且受到重視而迅速發展。

在澳洲有一項研究，該研究是針對使用元資料進行資料檢索的 20 個澳洲政府及教育機構進行檢索精確率評估[8]，該研究中提出了一個非常重要的概念，元資料與搜尋引擎應該是相輔相成的兩個工具，因為若沒有元資料的輔助，搜尋引擎的精確率將非常難以再向上提高的。

要精準的為讀者找出有用的資訊，前提是必須能夠精準的分析網頁內容。元資料能夠針對電子資源的特性編製適當的電子目錄以增加檢索的精確性，甚至能夠提高資料之間的互換性[9]。日期是網頁元資料中的重要欄位之一，但是由於台灣及其他中文語系的國家，在日期格式的書寫習慣上，穿插中文文字做為年、月、日的區隔[10]。在台灣，更是以民國年做為紀年，使得中文網頁中的日期形式較為複雜，因而增加了在網路資料自動著錄日期時的困難。

本研究的主要目的是針對網頁日期部份做深入的分析研究，以便能夠更精確的利用中文網頁中的日期欄位進行檢索利用，從而提高圖書館針對網路資源檢索、擷取及自動著錄時的準確性，提高圖書館網路資源管理的品質，同時更希望能達到拋磚引玉的作用，期望未來能有更多中文網頁元資料欄位自動著錄的相關研究，來提高使用元資料進行資料檢索的成效。

第二節 研究目的

1、了解目前中文網頁資源的日期格式和存在比例

2、計算中文網頁日期欄位自動辨識的正確率

3、了解中文網頁日期欄位自動著錄可能遭遇的困難

4、評估使用正規表示式自動擷取日期的可行性

第三節 研究方法

本研究採用的研究方法為實驗研究法，由於經過一段時間的觀察，發現目前台灣地區繁體中文的網頁資源上，有將近 50%的網頁有日期欄位，但由於格式上並未統一，也都未依循國際標準組織所訂定的格式，且仍習慣穿插中文文字做為年、月、日的區隔，在初步的分析下，發現目前中文網頁的日期格式主要包含下列幾種：

1. 年/月/日，例如 2009/11/30 或是 98/11/30
2. 年–月–日，例如 2009–11–30 或是 98–11–30
3. 年．月．日，例如 2009．11．30 或是 98．11．30
4. 用中文表示的日期格式，例如 2009 年 11 月 30 日或是 98 年 11 月 30 日

在初步的觀察中，發現有很多中文網頁雖然有日期欄位，但卻沒有自動

著錄必要，這類型的網頁都存在著多筆的日期欄位，主要是討論區或是部落格的留言文章，在本研究中，也會將這類型的網頁先予以排除。

實驗流程如下：

1、從詞庫中隨機抽取 200 筆詞彙

2、再將此 200 筆詞彙設定為檢索詞並經由 Google 搜尋

3、將每個檢索詞搜尋後的結果擷取前 100 筆網頁

4、再從此 100 筆網頁中隨機抽取 10 筆資料，總數為 2000 筆網頁分析研究標的

5、將此 2000 筆網頁的 HTML 資料轉換成純文字檔

6、使用人工方式建立這 2000 筆網頁之正確日期並排除無意義之網頁

7、設定欲抓取日期格式的正規式表示法公式

8、將設定好的正規式表示法公式帶入並逐一比對

9、計算抓取正確率

第四節 辭彙定義

1、中文網頁

本研究所指之中文網頁台灣地區內容為繁體中文並且於網際網路中對大眾公開之網頁。

2、日期欄位

本研究中所指之日期欄位為網頁中顯示可代表此中文網頁之建立時間，或是網頁最後更新時間。

第五節 研究限制

本研究主要是研究台灣地區繁體中文網頁的日期格式,所以排除掉大陸地區簡體中文網頁的部份。由於本研究截取網頁採用電腦自動抓取的機制，所以必需排除一些特定類型網頁，例如：討論區網頁及部落格留言網頁，這些網頁內容通常包含有多組無法代表此網頁或文章的日期。

註釋

[1] 張菀菁，「以模糊理論建構之圖書推薦系統」(碩士論文，淡江大學資訊工程研究所，民國 90 年)，頁 1。

[2] 卜小蝶，「Internet 資源蒐尋系統的發展與應用」，大學圖書館 2 卷 1 期（民國 87 年 1 月)，頁 36-54。

[3] Michael Lesk. The Seven Ages of Information Retrieval.（1995）
http://community.bellcore.com/lesk/ages/ages.html

[4] 卜小蝶，「Internet 資源蒐尋系統與圖書館資訊服務 － 以 Gopher 為例」，中國圖書館學會會報 53 期（民國 83 年 12 月)，頁 83-109。

[5] 吳毅成，「WWW 全球資訊系統之介紹及其展望」，資訊與教育雜誌（民國 83 年 8 月)，頁 2-11。

[6] 胡述兆、吳祖善，圖書館學導論，台北市：漢美，頁 1。

[7] 吳政叡，「元資料實驗系統和都柏林核心集的發展趨勢」，國立中央圖書館臺灣分館館刊 4 卷 2 期（民國 86 年 12 月)，頁 11-25。

[8] Lloyd Sokvitne. An Evaluation of the Effectiveness of Current Dublin Core Metadata for Retrieval.（2000）

[9] 余顯強，「以資訊處理觀點論 Metadata 之本質與意涵」，教育資料與圖書館學 45 卷 2 期（民國 96 年冬)，頁 249-266。

[10] 沈靜，「基于 UCL 的網頁信息自動標引技術研究」，現代圖書情報技術（2008 年 8 月)，頁 58-62。

第二章　相關文獻

第一節　網路資源組織

圖書館的傳統功能在收集及保存文化資源，並透過各種管理及技術，協助整體社會共同使用這些資源以獲得必要的知識[11]。隨著科技的進步，知識儲存型式從傳統的紙本型式演進至電子型式，也促成了資訊的爆炸。圖書館所面臨的不再只是單純的圖書、報紙及期刊...等，充斥在網路上的各種電子資源的整理及保存，成了圖書館所面臨的最大挑戰。過多的資訊是資訊搜尋者在使用網路資源上所面臨的主要挑戰之一，因此如何有效組織網路資源一直是資訊整理者與使用者所關心的議題[12]。

在過去，透過圖書的編目分類，圖書館可以提供給讀者非常確準的書目資訊，以協助讀者精確的找到資料，由於網路上的資源是龐雜且無組織的，在網路上找到精確的資訊是非常困難的[13]，所以在面對大量存在於網際網路上的電子資源，圖書館要如何提供給讀者有效且精準的書目資訊呢？

為防止資訊過於泛濫，並協助使用者能及時獲取有效資訊及資源，對資訊品質的控制、篩選或過濾網路資源，的確有其必要性[14]。近年來許多

專家學者致力於網路資源的自動分類編目，時至今天這個領域的研究已有許多不同的成果問世。為組織整理為數日增的電子資源，除了圖書館傳統編目作業針對編目規則、機讀編目格式做了部份修訂外，另有學者進而另謀技術上更簡便、更適合 Web 界面、更便於網路上資源分享的方式以組織電子資源，於是便有都柏林核心集（Dublin Core，簡稱 DC）的產生，以及能夠處理結構化資訊的 XML 標示語言的發展，而 OCLC 所推動的網路資源合作編目計畫 Connexion(原名 CORC)則促進網路資源編目及資源共享[15]。

第二節　元資料

根據吳政叡的整理，元資料（metadata）一詞最先出現於 NASA 的 Directory Interchange Format (DIF)手冊中，最常見的英文定義是 "data about data"，可直譯為描述資料的資料，就其本義而言，跟目錄（Catalogue）所扮演的角色並無太大的差別。編製目錄的目的，也在描述收藏資料的內容或特色，進而達成協助資料檢索的目的[16]。

簡而言之，元資料是一種結構化的資料，用於描述、解釋及定位資料，亦或是使資料更易於被檢索，進而達到對資訊資源的管理[17]。元資料是描述資源屬性以及彼此關係特性的資料，使資源在電子環境中能有效的被檢索、管理及利用[18]。

余顯強則針對元資料所產生的歷史背景做了以下的歸納，在圖書館尚未自動化前，編目記錄是以書目卡片呈現，即便在圖書目錄由紙本轉變為數位化機讀編目格式時，圖書館界仍沿用舊有的稱呼。而然當編目人員面對網路上的電子資源時，卻面臨到原本熟悉使用的不論是 AACR 還是 MARC 的必要改變，以適用網路資源的特性。而網路資源的也同時牽涉了圖書館界、電腦界及其他資訊科學領域，因此元資料一詞便成為了各個學科領域共通能接受的字彙[19]。

根據維基百科（wikipedia）的定義，元數據（Metadata），又稱元資料、中介資料，為描述數據的數據（data about data），主要是描述數據屬性（property）的資訊，用來支持如指示儲存位置、歷史資料、資源尋找、文件紀錄等功能。元數據算是一種電子式目錄，為了達到編製目錄的目的，必須在描述並收藏數據的內容或特色，進而達成協助數據檢索的目的。[20]

自從元資料的概念出現後，許多不同的機構便開始制定適用於不同的類型資料的元資料，雖然目前有許多不同的元資料標準如美國聯邦地理資料委員會（Federal Geographic Data Committee，FDGC）的地理電子元資料（Digital Geospatial Metadata）標準及 1995 年 OCLC/ NCSA Metadata Workshop 共同制定的都柏林核心集（Dublin Core）...等。然而各種元資料的目的卻是一致的，都是為了能夠更有效和更精準同時也更廣泛的，將電子時代下以不同載體所儲存的資料做整理、分析及著錄，以便於更進一步的運用。

University of Wisconsin—Milwaukee 的 Steven J Miller 將元資料歸納成 5 大類型，並以圖 2.1 的樹狀圖說明了元資料的類型[21]：

- 元資料結構標準
- 元資料內容標準
- 元資料屬性值標準
- 元資料格式標準
- 元資料呈現標準

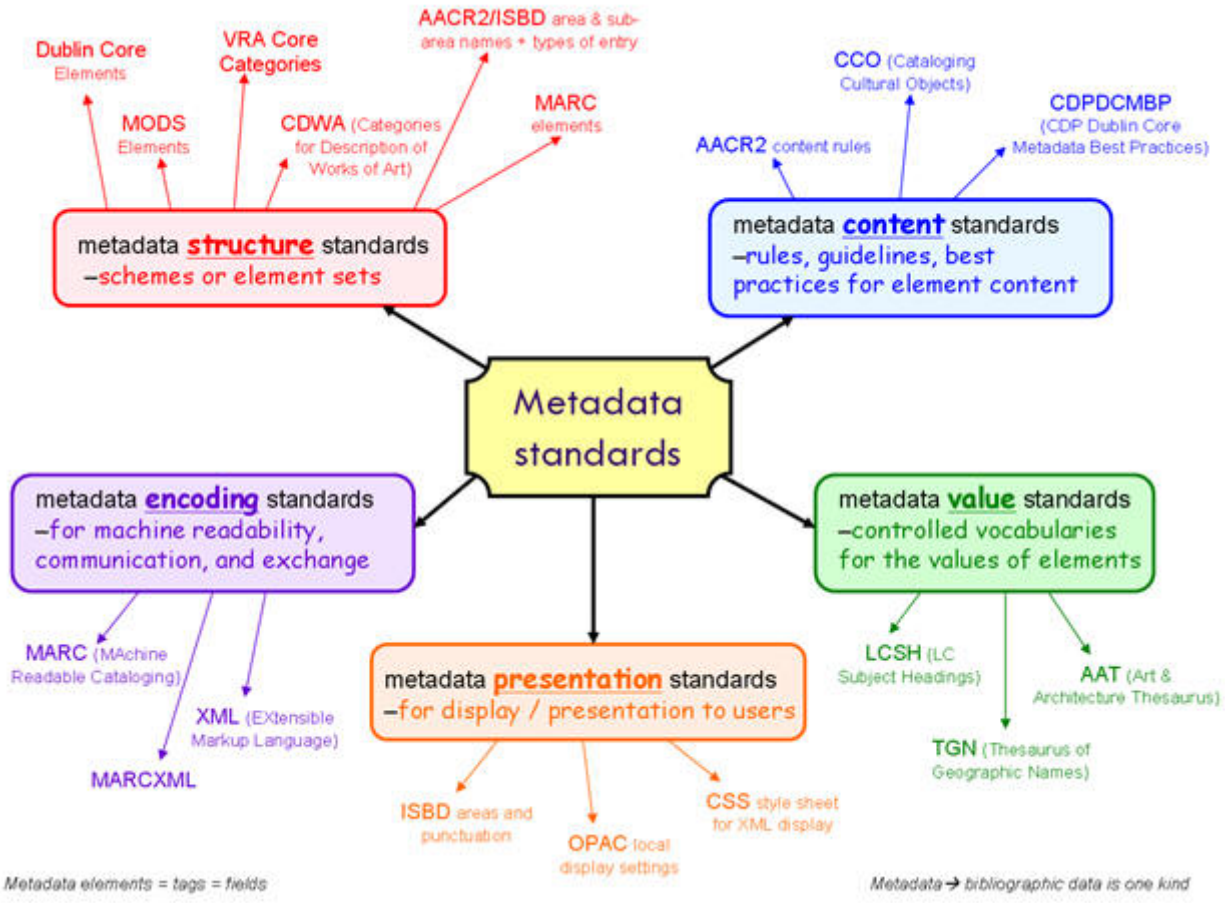


圖 2.1　Typology of Metadata and Cataloging Standards　http://www.uwm.edu/~mll/resource.html

同時根據 Steven J Miller 的結論，其中第一類型的元資料是目前最廣為被運用的[22]。都柏林核心集便是其中之一[23]。

都柏林核心集（Dublin Core Metadata Initiative，DCMI）是元資料的一種應用，是 1995 年 3 月由國際圖書館電腦中心（OCLC）和美國國家超級計算應用中心（National Center for Supercomputing Applications，NCSA）所聯合贊助的研討會，經由 52 位來自圖書館管理員和電腦專家共同制定的規格，建立一套描述網路上電子文件之特徵[24]。目的是希望建立一套描述網路上電子文件特色的方法，來協助資訊檢索。因此在研討會的報告中，將元資料定義為 "resource description"[25]。

隨著時代的進步及科技的日新月異，資訊的製造以及傳播也更加便利與快速，然而圖書館蒐集、整理、保存資料的使命數千年來卻不曾改變，圖書館的真正價值在於使圖書館的資料便於讀者利用，而元資料則是一個重要的工具，使圖書館在面對各類日新月異的資料時，能夠迅速並且有效的加以整理，並透過元資料可以顯著的提升讀者對正確資訊的獲取[26]，以達到服務讀者的最終目標。目前已有一些研究致力將元資料與自動編目著錄技術相結合，在可見的將來，圖書館能夠透過與科技的結合，提供給讀者既精確又即時的電子目錄，而讀者將能夠更有效率且精確的檢索到需要的資料，進而提升讀者資訊檢索滿意度，以及資訊檢索系統的可靠性[27]。

第三節　網頁自動擷取相關研究

元資料提供了一個能夠有效著錄網路資源的標準，而且面臨數以億萬計的網路資源，傳統的人工著錄技術已無法因應網路資源的增長速度。因此透過電腦技術來進行網頁相關書目資料的自動擷取是一個值得探討的解決方式。

在中文姓名自動擷取的方面，中國大陸已經有多篇相關的研究，吉林大學的計算機科學與技術學院曾經於 2006 年發表了一篇「基于神經網絡的中文姓名抽取技術」[28]，文中提到中文姓名的識別方式在目前技術下可以分為三種類型：規則法、統計法以及規則統計相結合的方法，這三種方式有一個共同的特性，都必須基於大規模的字詞資料庫，才能順利運作，而建立此字詞資料庫往往都必須要面對昂貴的代價及不易擴展的特性。正因如此該研究中採用類神經網路來進行漢語句子的分詞處理，解決了部份姓名的擷取問題，也提高了中文姓名自動擷取的精確度。

除了姓名的自動擷取外，針對網頁日期的部份，目前已有一些相關的研究，美商國際商業機器公司(IBM)擁有一項名為「用於從網站提取標注日期的內容的方法和系統」[29]的專利，根據此項專利的說明書，其中「標

注日期的內容」是指 URL 中含有日期的任何網頁資源，主要是在自動擷取動態產生之 URL 中的日期資料並將日期格式做辨識，以便後續其他程式做進一步加工利用。此項專利申請於 2005 年 7 月 12 日並於 2006 年 4 月 12 日公開。然而此一專利雖然能自動擷取 URL 中的日期並做辨識，但在日期辨識時所依據日期的格式需經特別指定，否則便直接採用瀏覽器上的預設日期格式，如：電腦地區設定為美國的預設日期格式為 mmddyyyy，而電腦地區設定為歐洲的預設日期格式為 ddmmyyyy。此專利雖為技術上的一大進展，但同時其精確度也受限於地區的限制。另外，此項專利的研究並不包含中文的日期格式，更遑論台灣地區所慣用的民國年。

在抓取網頁中已有的標籤(Tag)資料方面，2008 年中國西南科技大學信息工程學院發表了一篇名為「基于 UCL 的網頁信息自動標引技術研究」[30]，文中也認為在目前網際網路資訊超載的狀況下，要精確的找到所需要的資料的確非常困難，因此以都柏林核心集為基礎，建構了一套 UCL（Uniform Content Locator）統一內容定位技術，主要是應用在有線電視網路及網際網路。其目的主要是訂定網頁元資料的統一標籤格式，來增進網頁自動擷取的正確率。主要缺點是並不能適用於一般非使用此 UCL 架構建立的網頁。

另一項針對 wiki 線上編目系統的研究，則是藉由不同圖書館書目資料的比對、驗證以及加入對照表的方式，來證明該分類機制的可行性[31]。該研究的主要目的在透過一簡易、開放的協同編輯系統，藉由大家的協同合作，完成繁重的資料蒐集、整理工作[32]。此研究先利用圖書館的書目資料，來制定一套自動分類的機制，此後再藉由 Wiki 共筆系統的特性，來提供一個開放性平台，讓網路上的使用者能共同來進行修正、分享與討論，以讓圖書館的書目資料能更有效的運用。

以現有的資料處理技術，要在短期達成全自動的網頁自動著錄、分類是很困難的。因此目前較實際的作法，是發展自動擷取技術再加上人為的介入，共同提供盡可能完善、正確的資料給讀者，並且必需投入更多的研究朝向長遠的目標努力，期望未來能找出個最佳化的解決方案，能夠同時增加圖書館處理資料的效率，同時又能提供精確的書目資訊給讀者加以利用。

網路資源的更新的速度比傳統的圖書、期刊及報紙紙本型式更快速且無固定周期，因此，「出版日期」這個欄位的意義，在網路資源中又比傳統的紙本式圖書又來得更為重要。因此本論文特別針對自動著錄的日期

格式問題做更深入的研究，期望能夠研發出一套針對中文日期格式的擷取系統，為未來網頁的全自動著錄系統做出些許貢獻。

第四節　日期格式

國際上，不同的國家有不同的日期格式呈現方式。在台灣，年份有民國及西元兩種，而格式則有中式及西式。然而日期格式的問題，不止存在於中文語系國家，在英語系國家中，日期格式也有美式及英式二個主要大類。

日期格式呈現習慣是區域性的，但在國際上使用容易造成混淆及誤解[33]，ISO(International Organization for Standardization)便制定了一套以數字呈現的日期格式標準 ISO8601，ISO8601 標準的完整名稱為"Data elements and interchange formats – Information interchange – Representation of dates and times"，為國際日期格式之標準[34]。

ISO8601 源起於 1988 年的 ISO8601:1988 歷經多次修定，至今最新版本為 ISO8601:2004[35]。ISO8601 規範了日期格式、時間格式、國際標準時間、當地時間...等的統一呈現方式。而其優點則為易於系統讀寫、比較及排序以及無語言問題[36]。

而著名網際網路標準制定組織 W3C 則認為，使用 ISO8601 的數位日期系統的確有一些易讀性方面的問題。儘管這種方法並不完美，但 ISO

日期格式仍然是全球訪客都能理解（並且精確的）日期記法的最佳選擇[37]。

美商國際商業機器公司(IBM)擁有的一項名為「用於搜索電子文檔的日期的系統和方法」[38]的專利，根據此項專利的說明書，此專利主要是可擷取 free text 的電子文件中內含的所有日期資料，並可將所有擷取出的日期資料做一分析，進而歸納出此份文件所使用的日期格式，來辨識出正確日期。此項專利申請於 2006 年 10 月 17 日並於 2007 年 5 月 2 日公開。

此專利研發的技術背景是定位於全球資訊網上的內容，在非結構化文章中的日期可能以各種不同形式出現，如：“2004 年 10 月 11 日”，在文件中可能為：

11th of October 2004

11-10-2004

11 October, ‘04

Oct 11 th 04

11/10/04

10.11.2004

2004 Oct 11

甚至更多形式。這是非結構化網頁內容的特性，因此要有效的執行日期查詢是具有挑戰性的。

另外此文也提到針對日期的查詢時會遇到一個含糊格式的辨識困難問題，如 11.10.2004 可被解讀成 2004 年 10 月 11 日或 2004 年 11 月 10 日。而年份要是只有兩位數時（如：11/10/04）則又有更多的可能性。因此該專利的理論是擷取同一網頁內的所有日期資料再分析出其所使用的日期格式來辨識出日期。

然而在台灣及其他中文語系的國家，在日期格式書寫的習慣上，仍未完全依循 ISO8601，且仍習慣穿插中文文字做為年、月、日的區隔。同時國際上也尚無任何針對中文日期格式書寫的規範。雖然無上述西式日期格式容易產生如此眾多的歧異，但在台灣，卻另有以民國年做為紀年的習慣，因而增加了在網路資料自動著錄日期欄位時的困難。

本研究正是特別針對台灣的現況，期望能夠撰寫一個具有辨識各種台灣地區慣用的日期格式的系統，來增加於圖書館在針對網頁做日期欄位自

動著錄時的效率及準確性。進而提升圖書館各類資訊檢索的精確性。

註釋

[11] 宋瓊玲,「網路資源過濾技術在圖書館資訊服務的應用」,圖書館通訊 35 期(民國 88 年 2 月),頁 2-4。
[12] 羅思嘉,「網路資源組織：議題探討與相關計劃」,國立成功大學圖書館館刊 6 期(民國 89 年 10 月),頁 14-34。
[13] 李美幸,「一個編目館員對網路資源編目的看法」,清華大學圖書館館訊 26 期(民國 85 年 6 月),頁 37。
[14] 同註 11。
[15] 施孟雅,「資訊組織專題班—電子資源研習班研習紀要」,清華大學圖書館館訊 51 期(民國 91 年 9 月),頁 25。
[16] 吳政叡,「三個元資料格式的比較分析」,中華民國圖書館學會會報 57 卷(民國 85 年 12 月),頁 35-45。
[17] National Information Standards Organization. Understanding Metadata. (2004) http://www.niso.org/publications/press/UnderstandingMetadata.pdf
[18] Sherry L Vellucci, "Metadata and authority control" Library Resources & technical Services. 44 no. 1(2000) : 33-43.
[19] 余顯強,「以資訊處理觀點論 Metadata 之本質與意涵」,教育資料與圖書館學 45 卷 2 期(民國 96 年冬),頁 249-266。
[20] Wikipedia. 元數據. http://zh.wikipedia.org/wiki/Metadata
[21] Steven J. Miller. Metadata and Cataloging Online Resources. (July 2009) http://www.uwm.edu/~mll/resource.html
[22] Steven J. Miller. Metadata and Cataloging Online Resources. (July 2009) http://www.uwm.edu/~mll/resource.html
[23] Leif Andresen. Dublin Core as a tool for interoperability: Common presentation of data from archives, libraries and museums. http://dcpapers.dublincore.org/ojs/pubs/article/view/844
[24] Wikipedia. 元數據. http://zh.wikipedia.org/wiki/Metadata
[25] 同註 16。
[26] Baca, Murtha (editor). Introduction to Metadata. Second Edition.(2008) http://www.getty.edu/research/conducting_research/standards/intrometadata/
[27] A guide to metadata by the Metadata Advisory Group of the MIT Libraries. http://libraries.mit.edu/guides/subjects/metadata/index.html
[28] 吳芬芬,「基于神經網絡的中文姓名抽取技術」,吉林大學學報 44 卷 3 期(2006 年 5 月),頁 411-414。
[29] 國際商業機器公司(IBM),「用於從網站提取標注日期的內容的方法和系統」,中國大陸專利(2006 年 4 月 12 日)。
[30] 沈靜,「基于 UCL 的網頁信息自動標引技術研究」,現代圖書情報技術(2008 年 8 月),頁 58-62。
[31] 傅屹璽,「Wiki 線上編目應用系統建置之研究」(碩士論文,玄奘大學資訊傳播研究所,民 96 年),頁 1。
[32] 同註 31。
[33] Numeric representation of Dates and Time. http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm

[34] W3C. Date and Time Formates.
http://www.w3.org/TR/NOTE-datetime
[35] ISO 8601:2004.
http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=40874
[36] Numeric representation of Dates and Time.
http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm
[37] 張杜一維. 使用國際日期格式.
http://zdyx.org/w3c/QATips/cht/iso-date
[38] 國際商業機器公司（IBM），「用於搜索電子文檔的日期的系統和方法」，中國大陸專利（2007 年 5 月 2 日）。

第三章　系統實作方法與流程

第一節 Regular Expressions 演算法理論

RE 演算法的英文全名是 Regular Expressions，中文譯名為正規表示式，可簡稱為 regexp、regex 或 RE，最初的正規表示式出現於電腦科學的自動控制理論和形式化語言理論中。主要用途在於利用字元字串的特徵（Pattern）對文件內文做比對搜尋，甚至是替換，因此正規表示式是一套規則，一種可以與其他語言或工具相結合發揮組合相乘功效的表達方法。

1940 年代，Warren McCulloch 與 Walter Pitts 將神經系統中的神經元描述成小而簡單的自動控制元。在 1950 年代，數學家 Stephen Kleene 利用稱之為「正則集合」的數學符號來描述此模型。Ken Thompson 將此符號系統引入編輯器 QED，然後是 Unix 上的編輯器 ed，並最終引入 grep。自此，正規表示式被廣泛地使用於各種 Unix 或者類似 Unix 的工具，例如 Perl。

Perl 正規表示式源自於 Henry Spencer 寫的 regex[39]，它已經演化成了 pcre[40]（Perl 相容正規表示式，Perl Compatible Regular Expressions），一

個由 Philip Hazel 開發的，為很多現代工具所使用的函式庫。

正規表示式在歷經幾個時期的發展過後，現在的標準已經被 ISO（國際標準組織）批准和被 Open Group 組織認定（POSIX 1003.2）[41]。

正規表示式的主要用途是利用 RE 指定搜尋字串樣版，然後從檔案中找出符合該樣板的字串，並加以處理，常見用途如下：

- 將特定檔案中類似的字串取代掉
- 檢查使用者輸入字串是否符合指定樣版
- 更改日期格式的顯式模式
- 搜尋指定目錄下是否有符合字串樣版的檔案
- 語法剖析

正規表示式內容的組成主要有兩種，分別是：

1. 一般字元（Characters）：其代表意義與欲找尋字元之字面意義相同。
2. 運算操作元（Operators）：用以表示某一規則的意義，主要用以作為找尋比對時的特殊控制字元。

以下是幾個常用的正規表示式語法規則：

1. 單一字元（single character）：例如—Regular Expression "A"，如此一來任何包含大寫 A 字元的字串，都會被比對出來，如"Apple"、"ABC"但不會找出"any"。
2. . any （character）：．可用以代表任意的字元，例如—Regular Expression ".t.m"，那麼"atum"、"item"跟"strm"都會被比對找出。
3. +（kleene plus）：用以代表一個或一個以上的字元，例如可以 Regular Expression "50+" 比對找出"50"、"500"跟"500000"，但不會比對找出"5"。
4. * （kleene star）：用以代表零個或零個以上的字元，例如可以 Regular Expression "50*" 比對找出"5"、"500"跟"500000"。
5. ?（question symbol）：用以表示當句子中如果存在某一字串或某一字元的話，此一字串或此一字元恰出現一次的情況，例如可以使用 Regular Expression "?erd"，比對找出"berd"、"herd"與"erd"。
6. ^（caret）：用以比對一句子的開頭，例如一 Regular Expression "^THERE"，可以比對找出句子"THERE is my school"，而不會比對找出"I am still THERE!"。
7. $ （dollar symbol）：用以比對一句子的結尾，例如—Regular

Expression "here$"，可以比對找出"I am here."，而不會比對找出"here is my home."。

8. \（escape）：用以比對句子中代表某種意義的字元時，需在此字元前加上一反斜線，如：尋找文章中含有*的句子，可以Regular Expression " *"，作為尋找比對的特徵。

9. [ ]（ranges）：用以比對中括弧內任何的字元或字串，例如：Regular Expression "1[456]512"，會比對出"14512"、"15512" 與"16512"。

10. （）（group）：Regular Expression 可將字串或字元集組合成一個個基本的元素（element、atom），以（小）括弧包起來，然後以此元素作為比對的特徵；例如：Regular Expression "（eLand）（）（is）"，可比對找出句子"eLand is excellent!"。

11. |（alternatives）：此一符號相當於 OR 運算，在文章句子當中找出符合其中的選項，例如：Regular Expression "(cat|dog|bird)"，可比對找出句子"The pet store sold cats, dogs, and birds."中的"cat"、"dog"與"bird"。

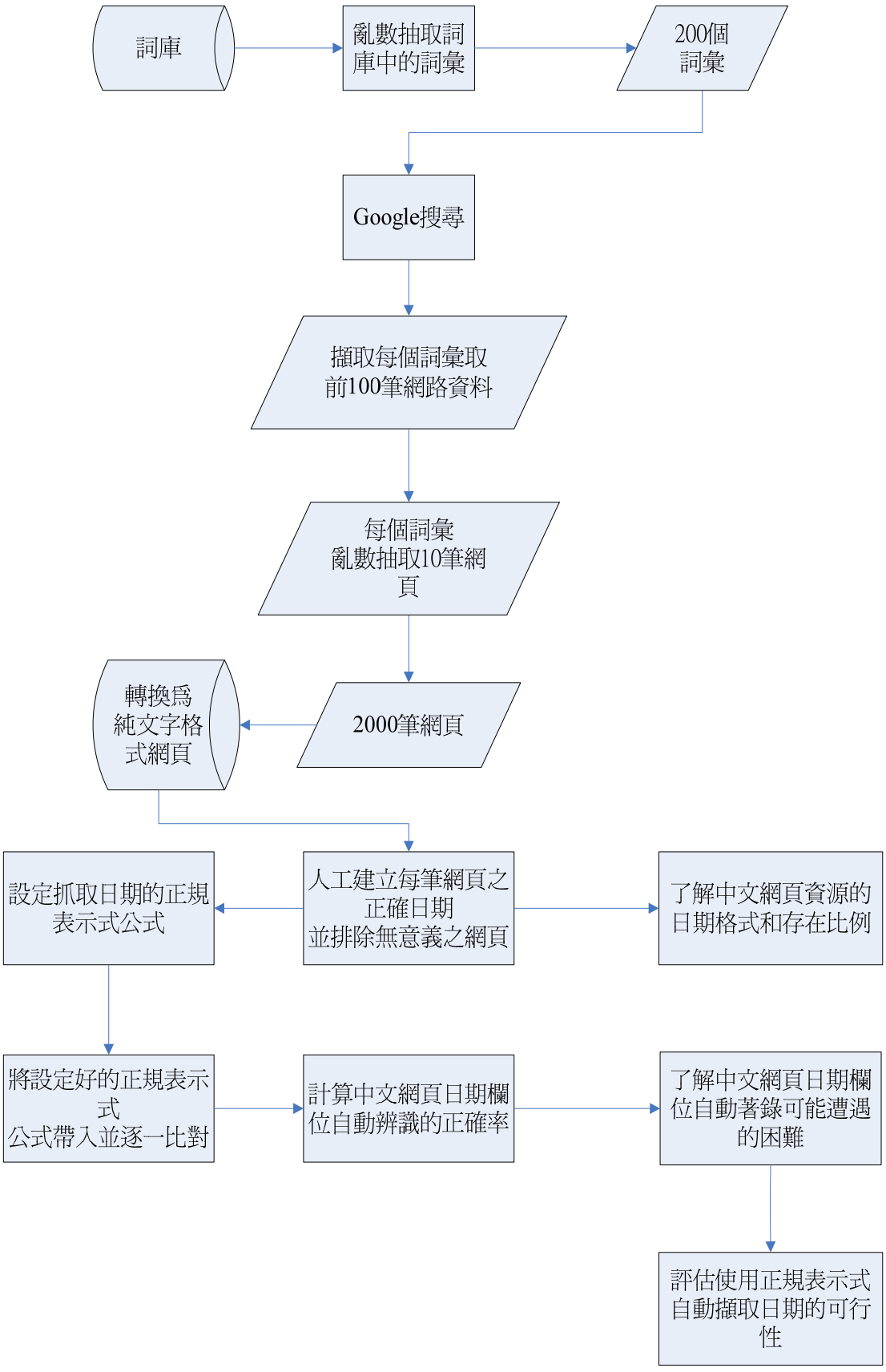
12. {n}、{n, m}（repetition）：使用 Regular Expression 時可以指定欲尋找的任意字元或字元集出現的次數，例如："a{5} b{,6} c{4,8}"，找出符合的結果會有，"aaaaa"、"bbbbb"與"ccccc"。

13. ^（complement、negation）：在 Regular Expression ^ 的另一個意義是作為"否"－not 的意義，例如：Regular Expression "[^a-z]a"，在句子"Mary had a little lamb. And everywhere that Mary went, the lamb was sure to go."中，只會找出部分。

正規表示式除了以上常用的規則外，以下簡式是經常被使用到的幾個例子：

1. \d：代表[0-9]數字。
2. \D：代表[^0-9] 非數字。
3. \s：代表[\t\n\x0B\f\r] 空白字元。
4. \S：代表[^ \t\n\x0B\f\r] 非空白字元。
5. \w：代表[a-zA-Z_0-9] 數字或英文字。
6. \W：代表[^a-zA-Z_0-9] 非數字或英文字。

第二節　研究流程圖


詞庫
亂數抽取詞庫中的詞彙
200個詞彙
Google搜尋
擷取每個詞彙取前100筆網路資料
每個詞彙亂數抽取10筆網頁
2000筆網頁
轉換爲純文字格式網頁
設定抓取日期的正規表示式公式
人工建立每筆網頁之正確日期並排除無意義之網頁
了解中文網頁資源的日期格式和存在比例
將設定好的正規表示式公式帶入並逐一比對
計算中文網頁日期欄位自動辨識的正確率
了解中文網頁日期欄位自動著錄可能遭遇的困難
評估使用正規表示式自動擷取日期的可行性

第三節 抽樣方式統計說明

本研究為求實驗精準，所採用的方式是使用亂數來抽取樣本，因網頁含有為數不少部落格等類的文章，預期將有 50%經篩選出的資料會被剔除，其餘的部份經重新整理後可做為實驗之用。

本研究所採用的詞庫是由飛資得資訊股份有限公司提供，詞庫總數量為 136961 個詞彙，先經由亂數取得 200 筆詞彙，再經由 Google 搜尋，每個詞彙取前 100 筆的網頁資料，在這 100 筆網頁資料中，經由亂數再抽取 10 筆出來，因此尚未前置處理的網頁資料樣本數量預計為 2000 筆。在將此 2000 筆的樣本經由人工排除一些不必要的網頁資料後，再採用人工建立每筆網頁資料樣本的正確日期格式，最後經由程式比對，計算正確率。

註釋

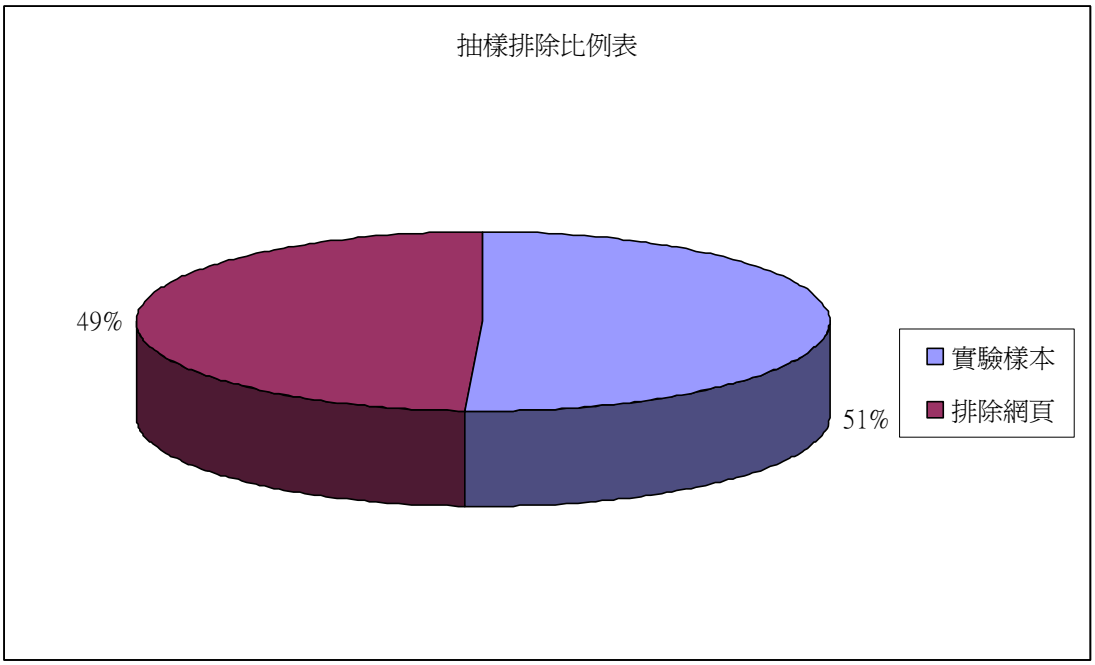
[39] An Introduction to Perl Regular Expressions.
http://www.perlfect.com/articles/regex.shtml
[40] PCRE - Perl-compatible regular expressions.
http://www.pcre.org/pcre.txt
[41] POSIX 1003.2 regular expressions.
http://www.unusualresearch.com/regex/regexmanpage.htm

第四章　研究結果與分析

第一節　中文網頁資源樣本統計與分析

本研究的目的是在使用正規表示式來自動辨識台灣地區中文網頁的日期，但前置作業仍必須依靠人工，來建立實驗樣本資料庫，在建立此資料庫時，必須準確的排除掉在研究限制中所排除掉的網頁，並建立出每筆網頁資料中正確的日期格式。在此前置作業完成後，也可以同時歸納出目前中文網頁資源所使用的日期格式種類、數量及存在比例。

本研究抽樣數量總數為 2000 筆網頁，經過人工建立正確日期答案，並剔除在研究限制中排除的討論區及部落格的網頁類型後，剩下 1018 筆網頁資料，刪除率約為 49%(圖一)，由此可見，目前網際網路網頁類型中，討論區及部落格這些網頁類型所佔的比例非常高，也是目前網頁的主要類型之一。



圖一 抽樣排除比例圖

在所有的有效樣本中，又可將這些網頁分成三大類型：新聞類型網頁、學術類型網頁以及一般類型網頁。新聞類型網頁顧名思義主要是包含新聞相關網頁(如圖二)，此網頁所描述內容往往包含人、事、時、地、物這幾項元素，也由於這些特性，所以新聞類型網頁存在日期欄位的比例相當高，在後面也會有詳細的統計分析。

圖二 新聞類型網頁

資料來源：台灣商務網 http://www.taiwanpage.com.tw/new_view.cfm?id=24980

第二類是學術類型網頁，由於本研究是使用亂數抓取詞庫中的詞彙進行檢索，所以常會抓取到一些論文、期刊或是網路書店的書目資料，這些網頁在本研究中都將視為學術類型網頁(如圖三)。

圖三 學術類型網頁

資料來源：國立清華大學機構典藏 http://nthur.lib.nthu.edu.tw/handle/987654321/27712

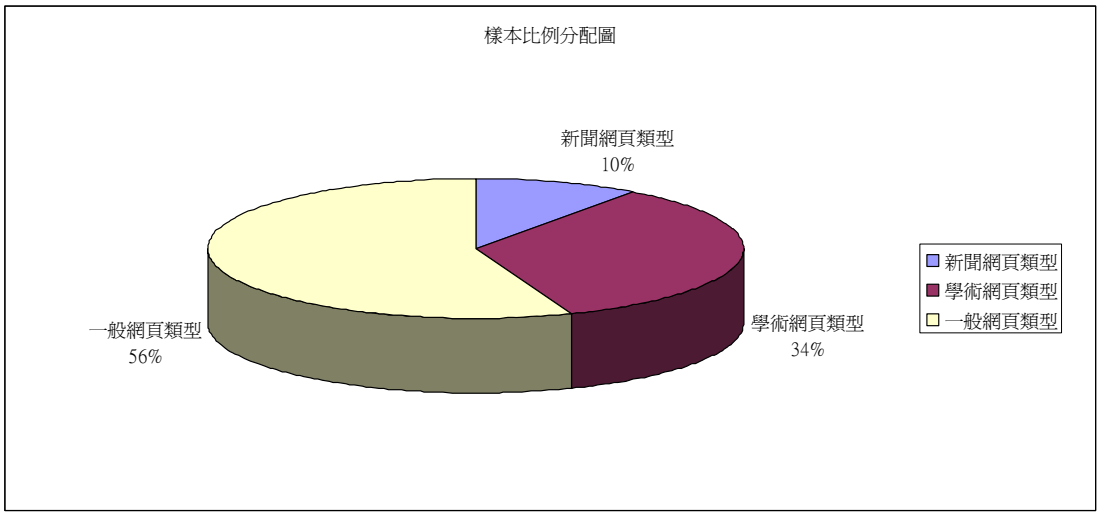
最後所謂的一般類型網頁就是泛指其餘類型的網頁，可能包含政府機關網頁、學校網頁、中小企業網頁或是一般的個人網頁，在本研究中都歸為一般類型網頁(如圖四)。

圖四 一般類型網頁

資料來源：內政部營建署全球資訊網

http://www.cpami.gov.tw/chinese/index.php?option=com_efsearch&view=efsearch&Itemid=138

圖五顯示，1018 筆有效樣本中，第一類的新聞類型網頁包含了 102 筆，佔總比例的 10%。第二類的學術類型網頁共有 348 筆，佔總比例的 34%。第三類的一般類型網頁則有 568 筆，佔總比例的 56%。

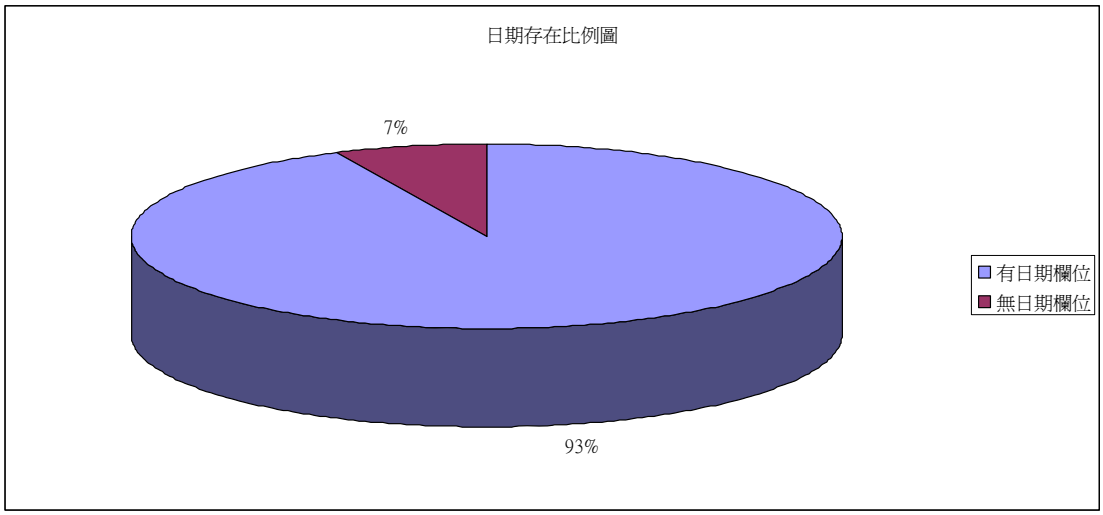


圖五 有效樣本比例分配圖

圖六顯示，在新聞類型的 102 筆網頁中，有日期欄位的網頁共 95 筆，佔這類型網頁的比例有 93%，其中的日期格式則包括表一中的幾種不同形式：

表一 新聞類型網頁之日期格式統計表

使用的日期格式	出現次數	範例	比例
西元年 “/” 月 “/” 日	66	2010/4/26	69.47%
西元年 “年月日”做爲區隔	19	2010 年 4 月 26 日	20.00%
西元年 “.” 月 “.” 日	3	2010.04.26	3.16%
民國年 “.” 月 “.” 日	3	99.04.26	3.16%
使用英文月份縮寫	3	26-Apr-10	3.16%
後置西元年份	1	26/04/2010	1.05%
總數	**95**		**100.00%**



圖六 新聞類型網頁之日期存在比例圖

在此類型的網頁中沒有日期的網頁包含了下列兩種，其中有三筆是抓取到新聞資料庫的檢索頁面(如圖七)；而另外四筆則是抓取到轉貼的新聞訊息，重點是在傳達新聞內容，但卻只註明出處，並未註明日期(如圖八)。在格式方面，用斜線區隔的日期格式佔了近七成，例如：2010/4/26。次多的則是用西元年但卻用中文文字年、月、日做為區隔，例如：2010 年 4 月 26 日，此種格式佔兩成的比例(見表一)。

圖七 新聞資料庫檢索頁面

資料來源：PChome 新聞

http://news.pchome.com.tw/search.php?q=%E6%B0%B8%E7%BA%8C%E7%99%BC%E5%B1%95

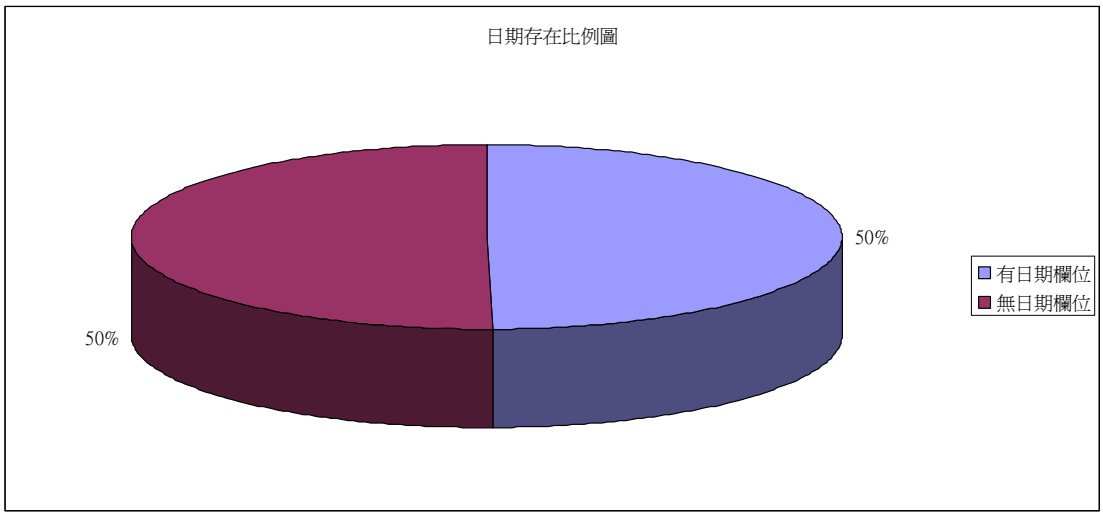
圖八 轉貼新聞訊息頁面

資料來源：吳舜文新聞獎助基金會 http://www.vivianwu.org.tw/02-23.php

圖九顯示，在學術類型網頁中，共有 348 筆網頁，其中有日期欄位的網頁為 173 筆，佔這類型的網頁比例 50%，其中的日期格式則包括表二中的幾種不同形式：

表二 學術類型網頁之日期格式統計表

使用的日期格式	出現次數	範例	比例
西元年 "/" 月 "/" 日	137	2010/4/26	79.19%
西元年 "年月日"做爲區隔	12	2010 年 4 月 26 日	6.94%
西元年 "." 月 "." 日	1	2010.04.26	0.58%
民國年 "." 月 "." 日	0	99.04.26	0.00%
使用英文月份縮寫	9	26-Apr-10	5.20%
後置西元年份	1	26/04/2010	0.58%
使用中文月份	8	26-四月-2010	4.62%
民國年 "年月日" 做爲區隔	5	民國 99 年 4 月 26 日	2.89%
總數	173		100.00%



圖九 學術類型網頁之日期存在比例圖

在此類型中，有日期的網頁佔了一半的比例，其中大部分都是期刊論文的摘要網頁。而沒有日期的網頁大部份是程式直接抓取到各圖書館博碩

士論文的檢索網頁，所以並沒有顯示出正確的日期(如圖十)。在此類型的網頁中，使用西元年並用斜線區隔年、月、日來表示日期的網頁為多數，佔了將近八成，顯示出此種日期格式為大部分網頁習慣使用的日期格式。統計表中出現另外兩種新聞類型網頁中並未出現的日期格式，第一種是在博碩士論文封面頁廣泛被使用的日期格式，以民國年表示並用年、月、日做為區隔的日期格式，例：民國 99 年 4 月 26 日；另一種則是少見的使用中文表示月份的格式，例：26-四月-2010。

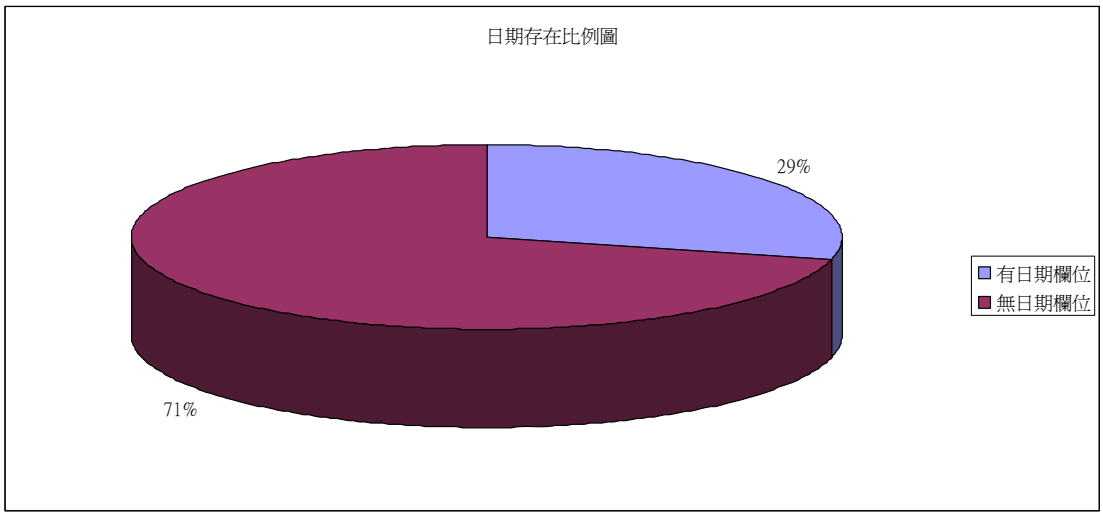
圖十 博碩士論文檢索頁面

資料來源：全國博碩士論文資訊網 http://etds.ncl.edu.tw/theabs/site/sh/search_result.jsp

圖十一顯示，在一般類型的 568 筆網頁中，有日期欄位的網頁共 164 筆，佔這類型的網頁比例 29%，其中的日期格式則包括表三中的幾種不同形式：

表三 一般類型網頁之日期格式統計表

使用的日期格式	出現次數	範例	比例
西元年 "/" 月 "/" 日	99	2010/4/26	60.37%
西元年 "年月日"做爲區隔	28	2010 年 4 月 26 日	17.07%
西元年 "." 月 "." 日	8	2010.04.26	4.88%
民國年 "." 月 "." 日	3	99.04.26	1.83%
使用英文月份縮寫	6	26-Apr-10	3.66%
後置西元年份	16	26/04/2010	9.76%
使用中文月份	2	26-四月-2010	1.22%
民國年 "年月日" 做爲區隔	2	民國 99 年 4 月 26 日	1.22%
總數	164		100.00%



圖十一 學術類型網頁之日期存在比例圖

在這類型的網頁中，有日期的網頁只佔了 29%，原因是有很多網頁是抓取到各人網頁或是中小企業的網站，這些網頁大部分都不存在日期欄位

(如圖十二)。而在日期格式的比例中，同樣是以使用斜線區隔的日期格式佔多數，約六成，例如：2010/4/26。以西元年並以年、月、日做為區隔的日期格式為次多，佔 17%，例如：2010 年 4 月 26 日。而第三多的格式則是使用後置的西元年份之格式，佔近 10%，例如： 26/04/2010。

圖十二 中小企業網站

資料來源：有泉行有限公司 http://www.nedtex.com.tw/product_detail.php?Pro_Seq=26

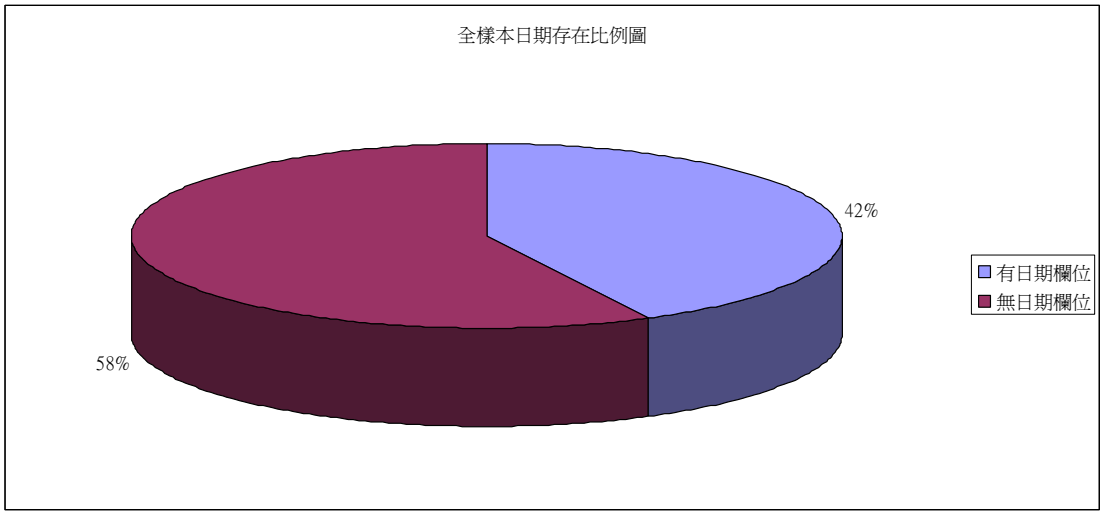
經過分析與統計，此次實驗中發現台灣地區網際網路之日期格式存在著至少如表四中表示的八種不同的日期格式，圖十三則顯示此次分析的 1018 筆可用網頁中，有 432 筆的網頁含有日期欄位，相當於總比例的 42%，而有 586 筆網頁沒有日期欄位，相當於總比例的 58%。

在日期格式方面，使用西元年並用斜線做為區隔的日期格式佔了將近七成，顯示此日期格式在台灣地區網頁被使用的最為廣泛。次多的則是使用西元年並用中文文字年、月、日做為區隔的日期格式，佔 14%。值得一提的是，在目前日期使用習慣上面，會使用民國年做為日期的年份單位之使用者，約佔總比例的 3%，關於這方面國人使用習慣的轉變，應該還有深入探討的價值。

最後，在本次統計中，有極小部分的網頁，在日期欄位部分只標示了年份以及月份，在此次的統計分析中並沒有將之視為獨立分類，而是依據其使用格式來併入一般分類中，但是這些只有年份及月份的日期欄位，在未來的自動辨識上，可能會成為雜訊的產生來源。

表四 台灣地區網頁之日期格式統計表

使用的日期格式	出現次數	範例	比例
西元年 "/" 月 "/" 日	302	2010/4/26	69.91%
西元年 "年月日"做爲區隔	59	2010 年 4 月 26 日	13.66%
西元年 "." 月 "." 日	12	2010.04.26	2.78%
民國年 "." 月 "." 日	6	99.04.26	1.39%
使用英文月份縮寫	18	26-Apr-10	4.17%
後置西元年份	18	26/04/2010	4.17%
使用中文月份	10	26-四月-2010	2.31%
民國年 "年月日" 做爲區隔	7	民國 99 年 4 月 26 日	1.62%
總數	432		100.00%



圖十三 台灣地區網頁之日期存在比例圖

第二節　中文網頁日期欄位自動辨識正規表示式公式設計與成效

在本次研究中，是使用亂數抽取詞庫中的詞彙，並使用 Google 搜尋來抓取本研究的研究樣本，在樣本尚未排除研究限制的網頁之前，共抓取 2000 筆網頁資料，經由人工逐筆過濾並建立正確日期答案後，真正成為實驗樣本的網頁資料共有 1018 筆，其中包含有日期及無日期的網頁，在處理樣本前置工作時，也將網頁進行分類及統計，並依照期日期格式設計出正規表示式的比對公式，接下來要進行的就是使用針對本次研究設計出的系統來進行日期辨識的工作。

根據前一節的台灣的區網頁之日期格式統計表(見表四)，已經知道台灣地區網頁的日期格式有哪些種類，因此依照所有出現過的日期格式，來設計其正規表示式的比對公式，並依其所佔比例，來給予不同權重值，使程式在辨識網頁日期時，能分辨其優先順序。

表五是依據本次的統計數據，設計出來的正規表示式公式表，以及各公式之權重值：

表五 正規表示式公式及權重表

使用的日期格式	正規表示式公式	比例	權重
西元年 "/" 月 "/" 日	[12]\d\d\d/\d{1,2}/\d{1,2}	69.91%	200
西元年 "年月日"做爲區隔	[12]\d\d\d 年\d{1,2}月\d{1,2}日	13.66%	150
後置西元年份	\d{1,2}/\d{1,2}/[12]\d\d\d	4.17%	100
後置西元年份	\d{1,2}-\d{1,2}-[12]\d\d\d	4.17%	100
使用英文月份縮寫	d{1,2}-\D\D\D-[12]\d\d\d	4.17%	100
西元年 "." 月 "." 日	[12]\d\d\d[..]\d{1,2}[..]\d{1,2}	2.78%	80
使用中文月份	d{1,2}-\十?[一二三四五六七八九]月-[12]\d\d\d	2.31%	70
民國年 "年月日" 做爲區隔	\d\d 年\d{1,2}月\d{1,2}日	1.62%	60
民國年 "." 月 "." 日	\d\d[..]\d{1,2}[..]\d{1,2}	1.39%	60

在計算正確率方面，本次研究總共使用了兩種計算方式：

(一)有日期部份的完全命中率，完全命中率是指經由程式比對第一次就正確辨別出日期的網頁。

(二)第二種則是針對沒有日期的網頁，主要是計算在沒有日期欄位的網頁中，經由程式比對能正確的辨別出該網頁沒有日期欄位的正確率。

在完全命中率方面，有日期且經由程式比對第一次就命中的網頁共有 267 筆，正確率為 61.81% (267/432)；而在沒有日期欄位的網頁部分，程式能正確指出沒有日期的網頁部分共有 368 筆，正確率為 62.80% (368/586)。(見表六)

為了證明依據日期格式的出現比例來設計權重會出現最高的正確率，本次實驗中也將表五中的權重值倒置，並重新計算一次正確率，得到的正

確率為 13.89%。而若只倒置更改前三高的權重值順序，得到的正確率則為 23.40%。由此可見，依照日期格式出現比例高低來給予權重值的方式，才能取得本次實驗中最高的正確率。

表六 正確率表

	筆數	正確率
有日期完全命中率	267	**61.81%**
沒有日期命中率	368	**62.80%**

表七 分類正確率表

網頁分類	完全命中筆數	有日期網頁筆數	正確率
新聞類型網頁	59	95	**62.11%**
學術類型網頁	121	173	**69.94%**
一般類型網頁	87	164	**53.05%**
總數	267	432	**61.81%**

在網頁分類的完全命中率方面，新聞類型網頁完全命中筆數共有 59 筆，正確率為 62.11% (59/95)；學術類型網頁完全命中筆數共有 121 筆，正確率為 69.94% (121/173)；一般類型網頁完全命中筆數共有 87 筆，正確率為 53.05% (87/164)。(見表七)

由於網頁日期自動擷取很難達到百分之百的正確率，為了要瞭解直接使用網頁自動擷取日期欄位來著錄日期可能的誤差狀況，本研究也設計了

一個稱為「誤差年」的計算方法，主要是計算網頁正確日期欄位的年份與經由程式辨別出第一筆日期欄位的年份差異，如果是完全命中的網頁，其誤差年的值就是 0，若正確日期年份為 2010 年但程式猜測出的日期為 2008 年，其誤差年的欄位值則給-2，以此類推。

表八 絕對誤差年敘述統計表

敘述統計

	個數	最小值	最大值	平均數	標準差
絕對誤差年	403	0	33	.62	2.322
有效的 N (完全排除)	403				

表八是針對誤差年取絕對值後所做的統計敘述，有日期的網頁共有 432 筆，其中有 29 筆網頁雖然有日期，但是透過正規表示式並沒有抓取到日期，因此無法計算誤差年，所以在敘述統計中有效的樣本數量為 403 筆。在本次敘述統計中最大誤差值是 33 年，最小誤差值為 0 年，而平均誤差年為 0.62 年，標準差為 2.322。由此可見在本次實驗中，人工判斷出的正確日期與程式判斷出的正確日期平均誤差為 0.62 年，這代表著雖說有日期網頁方面的完全命中率約六成多，但在誤差年份方面相差不遠，誤差平均都約在 0.6 年左右。

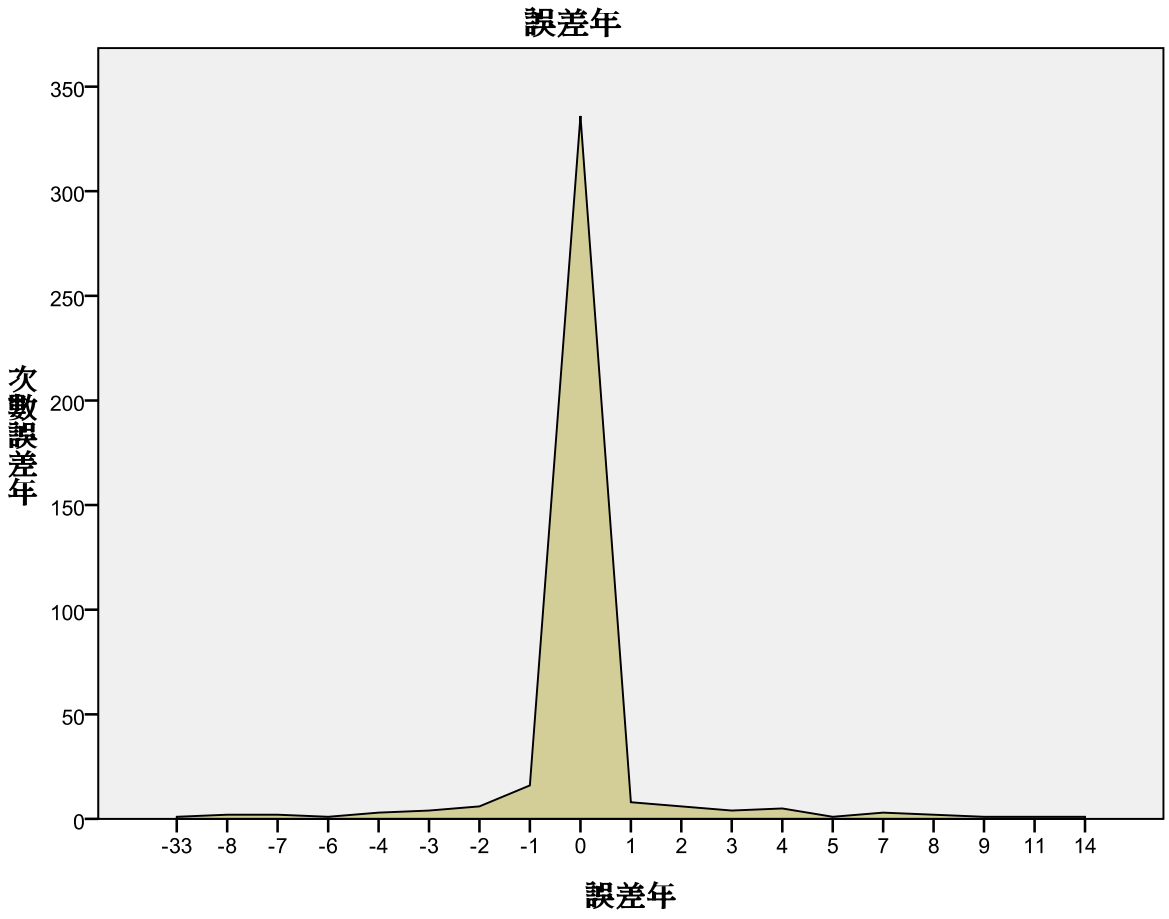
表九是針對誤差年所做的次數分配表，誤差年為 0 的網頁共有 336 筆，佔有效樣本中的 83.4%，而誤差年±1 的網頁共有 24 筆，佔有效樣本中

的 6%，合計誤差年小於一年的網頁佔有效樣本的比例共 89.4%。在圖十四中，可以清楚見到本次實驗的誤差年次數分配狀況，有近九成的網頁集中在 0～±1 的區間，更可顯示出本次實驗在辨識日期上的誤差程度。

表九 誤差年次數分配表

誤差年

		次數	百分比	有效百分比	累積百分比
有效的	-33	1	.2	.2	.2
	-8	2	.5	.5	.7
	-7	2	.5	.5	1.2
	-6	1	.2	.2	1.5
	-4	3	.7	.7	2.2
	-3	4	1.0	1.0	3.2
	-2	6	1.5	1.5	4.7
	-1	16	4.0	4.0	8.7
	0	336	83.4	83.4	92.1
	1	8	2.0	2.0	94.0
	2	6	1.5	1.5	95.5
	3	4	1.0	1.0	96.5
	4	5	1.2	1.2	97.8
	5	1	.2	.2	98.0
	7	3	.7	.7	98.8
	8	2	.5	.5	99.3
	9	1	.2	.2	99.5
	11	1	.2	.2	99.8
	14	1	.2	.2	100.0
	總和	403	100.0	100.0	



圖十四 誤差年次數分配圖

第三節　中文網頁資源日期欄位自動辨識之困難與可行性評估

在本次研究中，以亂數抽樣的方式取得台灣地區中文網頁的樣本，並詳細分析網頁類型、日期存在比例以及存在的各種日期格式，清楚的了解到目前網際網路上日期格式並未完全依循 ISO8601，且國人在使用習慣上會使用中華民國為紀年的也是少數，所以在做自動辨識日期欄位的動作時，需要比對及過濾多種不同的日期格式。相對的，設定越多種不同的日期欄位比對公式，就會造成許多無法預期的雜訊，而這些雜訊在處理上亦非常棘手。

在剛開始設計比對公式之時，筆者曾針對 19 筆日期格式只有年份跟月份的網頁來設計比對公式，但收到的成效不佳，造成許多無法預期的雜訊，有日期方面的完全命中率亦下降到四成左右，並且經過詳細分析後，這些多餘的雜訊並無法透過修正公式來排除，唯一的辦法只有先放棄只有年份跟月份的日期格式，才能去除這些多餘的雜訊，因此在本次實驗中有作用的正規表示式公式均有年、月、日。

在整個實驗流程進行完成之後，針對有日期欄位但程式比對不出的網頁進行分析，發現造成問題的原因包含下列幾種：

1、網頁日期欄位只有年份及月份

有些特定的網頁基於其要表達的資訊內容，上面只會有年份及月份，例如圖書館每月的新書目錄或期刊的書目資料(如圖十五)，都只會註明年份及月份，像是 98 年 06 月或是 12/2004，這些類型的日期格式共有 13 筆，在本次實驗中都無法將日期辨識出來。

圖十五 圖書館新書目錄

資料來源：中原大學圖書館 http://www3.lib.cycu.edu.tw/newlycat/history/9806/chinbk_4.html

2、網頁日期欄位年、月、日中穿插空白字元

例如：九 十 八 年 十 二 月

2009 年 9 月 28 日

2005 / 08 / 01

這個會使程式失誤的原因在實驗一開始是筆者無法預料到的，在實驗完成後，才發現有 11 筆網頁，也許是為了讓網頁排版較為美觀，

在日期欄位的每個字元中會多穿插一個空白字元(如圖十六)，這樣使得正規表示式在判斷上會失去其功用，而若從修改公式的方法要來抓取這些日期欄位，只會徒增更多不必要的雜訊，唯一的解決之道就是在前置作業時增加一項功能，把網頁轉換成純文字格式的同時，將空白字元去除，這樣應該可以解決此種問題。

圖十六 日期欄位穿差空白字元

資料來源：C-Science http://www.c-science.com/txt/tc/en/981207en.htm

3、網頁日期欄位是嵌入在圖片裡面的

在本次實驗中，使用這種方式表達日期的網頁雖然只有 2 筆(如圖十七)，但當遇到此種網頁日期類型，使用正規表示式來辨別日期的方式就失去其效用，而此種表達日期的方式必須使用人工或其它能解讀圖片內容的方法，才能辨別出正確日期。

圖十七 網頁日期嵌入圖片

資料來源：第九屆研究所博覽會 http://edu-show.tkb.com.tw/web/index.jsp#

4、抓取到的日期欄位在月份及日期上無法判定

在本次實驗中，也發現了一個較為棘手的現象，程式辨別出的日期與人工辨別出的日期在月份上面無法判讀，當月份及日期都小於數字 12 時，在判定正確日期上會出現問題，此種狀況在本次實驗中發生 18 次(如圖十八)。不過關於此種情況，就算用人工判斷都無法正確解讀，因此這方面的問題尚無方法可解決，

例如：1991 年 5 月 6 日或 1991 年 6 月 5 日 以及 06/05/1991

圖十八 月份以及日期無法判定的網頁

資料來源：美國國會法律圖書館 http://www.glin.gov/view.action?glinID=198016

在可行性評估方面，由於本次實驗只排除掉討論區及部落格類型的網頁，因此要面對多種不同類型的網頁來進行自動比對工作，再加上國內在日期格式的表達上面並沒有訂定標準，所以要針對日期欄位進行自動著錄或是自動辨識會存在著一定的難度。以目前的成果來看，要將辨別日期的工作完全交由程式自動辨別的確還有一段很長的路要走。但以目前約六成的正確率來說，雖然無法完全將辨別日期的工作交由程式執行，但若適當運用本次實驗採用的方法，應該可以提高使用者在尋找網路資源的效率，例如可以將辨別出的日期欄位註明是參考值，供使用者在檢索時參考，由於本實驗的誤差年約為 0.6 年，且誤差年在一年內的比例約佔九成，使用者可以根據自動辨別出的日期來篩選資料，成為一種輔助工具，所以只要應用得當應可以增進在尋找網路資料以及處理日

期欄位時的效率。

第五章　結論與建議

第一節　日期格式的制定與建議

在國內，日期書寫的格式上並未完全依循 ISO8601，仍習慣民國年或是西元年擇一使用，雖然此種現象已經持續非常多年，國人也早已習慣在民國年與西元年間的轉換，但對數位科技的發展來說，日期格式未統一的確造成了許多不必要的問題。回首網路科技發展的路程，發展時期不過二十年左右，但電子資料增加的速度早已超越紙本資料的發展歷程，而在目前環保意識抬頭的觀念下，發展無紙化生活必定是未來的趨勢，因此保存數位歷史發展的數位資料變成非常重要。時間對歷史的意義是無庸置疑的，而表達時間的方法，就是要依靠日期格式來達成，如果沒有一套大家依循的標準，日期便可能成為一個無意義的欄位。

在元資料的發展中，其最重要的概念就是要詳細的描述資料，好讓電子資料能夠被應用的更廣泛且更方便，如此一來可以加強資料間的互換性，又能提高資料的被使用性。以筆者的角度來看，目前元資料在學術上的應用已經非常廣泛，但在幾乎人人每天都會接觸到的網際網路上，應用度似乎成效不彰，這點從日期格式這個欄位便可看出，缺乏統一格式的日期欄位對搜尋引擎來說，是無法下手提升其搜尋精確度的，否則

一篇無法辨識日期的網路資料，儘管記載著許多有意義的資訊，但失去了時間軸的相對關係，仍舊是無法被使用的。

由於國內目前尚未有針對日期格式的統一標準規範，所以目前在台灣地區的網頁，其使用的日期格式也非常煩雜，以學術論文來說，在電子檔的封面頁會使用以中華民國為紀年的日期格式，但在博碩士論文的檢索頁面，則會使用以西元記年的日期格式來記載通過口試的日期，雖然在人工判讀上不會有任何問題，但若用電腦資訊的角度來看，在解讀上便成為一個棘手的狀況，畢竟以目前的資訊技術，資訊系統的人工智慧尚未有人類的邏輯與判斷能力，所以在解讀這兩種同義卻格式不同的日期上，會造成一定程度的問題，而間接降低資料間的互換性與共通性。

針對此次的實驗結果來看，國人在建置網頁時，大部分的使用習慣還是比較接近 ISO8601 之規範，而非使用所謂後置年份的英式日期格式，所以筆者認為，由於我國國情不同，政府相關單位可以盡速規範出兩種標準的日期格式，分別是以民國年與西元年為紀年的標準日期格式，如此一來不僅可以解決許多在日期自動辨識上的困難，在兩種日期格式僅紀年單位不同的狀況下，轉換上也不會造成過多煩人的問題，更可以大大提升網路資訊的可用性。

第二節 結語

本次研究的最初目的，主要是希望能夠藉由使用最簡單的方式，來進行網際網路日期格式的自動辨識，由於目前都尚未有相關研究，所以本研究抱著投石問路的態度來進行實驗，雖然實驗結果的最終正確率在有日期網頁及無日期網頁都各只有約六成的正確率，此種數據資料表示以目前的狀況還是尚未能夠完全取代人工，但由於此次實驗的誤差年離散程度有近九成集中在一年，所以相信若應用得宜，使用此方法來搜尋及處理網頁日期欄位仍然可以成為一種輔助力量，進而提升檢索效率。

在本次實驗的過程中，筆者也發現了許多在自動擷取日期欄位上，產生雜訊的原因，當然最主要的原因不外乎如前一節所談到的日期格式煩雜，也希望有政府的相關單位能重視此問題，畢竟網路資訊與人類歷史至今已經緊密結合，現在看起來不重要的網頁，在多年後可能都變成紀錄人類發展史的極重要資料，但一旦喪失了相對日期的解讀，就會成為無意義的垃圾資源。正因如此，唯有盡早規範出日期格式使用規則，才能真正提高網路資源的可用度。

第三節 未來研究方向

在此次實驗中，是針對台灣地區中文網頁的日期欄位來做辨別，在有效樣本中只剔除了討論區以及部落格的網頁類型，正因如此，本實驗要處理的網頁類型非常多種及複雜，雖然正確率約六成左右，但也證明了使用正規表示式自動辨別日期欄位的可行性，目前雖然無法完全將辨別日期的工作交由電腦自動執行，但若將樣本範圍縮小，甚至針對特定類型的電子檔案，應該可以獲得更高的正確率。

本研究主要是在自動辨別網頁元資料的日期欄位，但在電子資源的元資料中，和日期欄位同等重要的還有作者欄位，除了針對日期欄位的自動辨別，也可以嘗試針對作者欄位的自動辨別來做研究，若這兩個欄位在自動辨別的技術上都有所進步，相信不僅可以減少圖書館在著錄電子資源的人力虛耗，提升工作效率，更可以進一步來提升搜尋引擎的精確率，此時受惠的不在只侷限於圖書館，而是所有網際網路的使用者。

參考書目

中文部份

卜小蝶。「Internet 資源蒐尋系統的發展與應用」。大學圖書館 2：1（民國 87 年 1 月），頁 36-54。

卜小蝶。「Internet 資源蒐尋系統與圖書館資訊服務 － 以 Gopher 為例」。中國圖書館學會會報 53（民國 83 年 12 月），頁 83-109。

李美幸。「一個編目館員對網路資源編目的看法」。清華大學圖書館館訊 26（民國 85 年 6 月），頁 37。

沈靜。「基于 UCL 的網頁信息自動標引技術研究」。現代圖書情報技術（2008 年 8 月），頁 58-62。

宋瓊玲。「網路資源過濾技術在圖書館資訊服務的應用」。圖書館通訊 35（民國 88 年 2 月），頁 2-4。

吳芬芬。「基于神經網絡的中文姓名抽取技術」。吉林大學學報 44：3（2006 年 5 月），頁 411-414。

吳政叡。「三個元資料格式的比較分析」。中華民國圖書館學會會報 57（民國 85 年 12 月），頁 35-45。

吳政叡。「元資料實驗系統和都柏林核心集的發展趨勢」。國立中央圖書館臺灣分館館刊 4：2（民國 86 年 12 月），頁 11-25。

吳毅成。「WWW 全球資訊系統之介紹及其展望」。資訊與教育雜誌（民國 83 年 8 月），頁 2-11。

余顯強。「以資訊處理觀點論 Metadata 之本質與意涵」。教育資料與圖書館學 45：2（民國 96 年冬），頁 249-266。

胡述兆、吳祖善。圖書館學導論。台北市：漢美，民 80。

施孟雅。「資訊組織專題班—電子資源研習班研習紀要」。清華大學圖書館館訊 51（民國 91 年 9 月），頁 25。

國際商業機器公司。「用於從網站提取標注日期的內容的方法和系統」。中國大陸專利，2006 年 4 月。

國際商業機器公司。「用於搜索電子文檔的日期的系統和方法」。中國大陸專利，2007 年 5 月。

張杜一維。使用國際日期格式
＜http://zdyx.org/w3c/QATips/cht/iso-date＞（2009 年 12 月 01 日）

張菀菁。「以模糊理論建構之圖書推薦系統」。淡江大學資訊工程研究所，碩士論文，民國 90 年。

傅屹璽。「Wiki 線上編目應用系統建置之研究」。玄奘大學資訊傳播研究所，碩士論文，民國 96 年。

Wikipedia。元數據
＜http://zh.wikipedia.org/wiki/Metadata＞（2009 年 12 月 15 日）

西文部份

A guide to metadata by the Metadata Advisory Group of the MIT Libraries.
http://libraries.mit.edu/guides/subjects/metadata/index.html

An Introduction to Perl Regular Expressions.
http://www.perlfect.com/articles/regex.shtml

Andresen, Leif. Dublin Core as a tool for interoperability: Common presentation of data from archives, libraries and museums.
http://dcpapers.dublincore.org/ojs/pubs/article/view/844

Baca, Murtha (editor). Introduction to Metadata. Second Edition.(2008)
http://www.getty.edu/research/conducting_research/standards/intrometadata/

ISO 8601:2004.
http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=40874

Lesk, Michael. The Seven Ages of Information Retrieval.（1995）
http://community.bellcore.com/lesk/ages/ages.html

Miller, Steven J. Metadata and Cataloging Online Resources. (July 2009)
http://www.uwm.edu/~mll/resource.html

National Information Standards Organization. Understanding Metadata. (2004)

http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Numeric representation of Dates and Time.
http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm

PCRE - Perl-compatible regular expressions.
http://www.pcre.org/pcre.txt

POSIX 1003.2 regular expressions.
http://www.unusualresearch.com/regex/regexmanpage.htm

Sokvitne, Lloyd. An Evaluation of the Effectiveness of Current Dublin Core Metadata for Retrieval. （2000）

Vellucci, Sherry L. “Metadata and authority control” Library Resources & technical Services. 44 no. 1(2000) : 33-43.

W3C. Date and Time Formates.
http://www.w3.org/TR/NOTE-datetime