大學圖書館
15卷1期（民國100年3月） 頁132-143

臺灣地區中文網頁自動辨別日期之研究

A Study of Auto Extraction of Dates from Chinese Web Pages in Taiwan Area

邰文暉 Wen-Hui Tai
輔仁大學圖書資訊研究所研究生
Graduate Student, Department of Library and Information Science,
Fu-Jen University
E-mail: mars.tai@gmail.com

吳政叡 Cheng-Juei Wu
輔仁大學圖書資訊系專任教授
Professor, Department of Library and Information Science,
Fu-Jen University
E-mail: lins1022@mail.fju.edu.tw

【摘 要】

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Metadata中的重要欄位，由於臺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，因而增加了自動著錄網頁創造（或修改）日期時的困難。本研究的主要目的是針對網頁日期部分做深入的分析研究，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。本研究以隨機抽樣方式來抓取繁體中文網頁，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自動抓取正確的網頁日期，最後計算出正確率。透過此研究可以了解在進行中文網頁日期欄位自動辨識時可能會遭遇到的困難，並評估自動擷取繁體中文網頁日期欄位的可行性。實驗結果顯示，有日期資料網頁的正確率約為61%，沒有日期資料網頁的部分約為62%。有日期資料網頁的平均誤差年約為0.62年，且83.4%的網頁能精準預測其年份（即誤差年為0），因此雖然本研究的成果尚未能完全取代人工，但若應用得宜仍然可以提高網頁檢索時的效率。

收稿日期：民國99年9月23日；接受日期：民國99年12月24日

【Abstract】

Online resources have become more plentiful nowadays, thanks to the popularization of Internet services. In order to achieve accurate search results for the users, it is necessary to analyze web pages precisely. '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 different types of date displaying formats applied to Chinese web pages. These findings will be used to increase the precision on the date auto extraction of web pages. The procedures of experiment are as follows. Firstly, samples were randomly selected from Internet. Secondly, the statistic analysis on the date displaying format of each web pages was conducted. Lastly, Regular Expression was used to abstract the dates of each web page, while the accuracy ratio was also 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.

關 鍵 字：日期格式；網頁日期；自動日期辨別；元資料；詮釋資料；後設資料
Keywords：Date format; Webpage date; Auto date extraction; Metadata

壹、前言

隨著網際網路的日益普及與線上資源越來越豐富，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（註1），搜尋引擎（例如Google）便在此時應運而生，但由於網際網路的資訊量實在太過龐大，資訊超載的問題也日益嚴重，使用者欲找出真正有用的資訊，有時仍必須依靠自身的判斷來過濾非相關的資訊，因此雖然電腦處理資料的速度已經非常快，但是有效率的檢索和資訊超載，仍是亟待解決的重要問題（註2）。

Metadata（元資料、詮釋資料、後設資料或元數據）便是在此一背景下產生並且受到重視而迅速發展，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。Metadata能夠揭示電子資源的特性以增加檢索的精確性，甚至能夠提高資

料之間的互換性（註3）。澳洲有一項研究是針對使用Metadata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（註4），該研究發現Metadata與搜尋引擎應該是相輔相成的兩個工具，因為若沒有Metadata的輔助，搜尋引擎的精確率將非常難以再向上提高的。

日期是網頁中的一個重要資訊，國際上不同國家有不同的日期格式呈現方式。在臺灣，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。因為日期格式呈現習慣是有區域性的，在國際上使用容易造成混淆及誤解（註5），因此ISO（International Organization for Standardization）便制訂了一套以數字呈現的日期格式標準ISO 8601，為國際網頁日期格式之標準（註6）。ISO 8601源起於1988年的ISO 8601:1988，歷經多次修訂，至今最新版本為ISO 8601:2004（註7）。ISO 8601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。ISO 8601的優點為易於系統讀寫、比較及排序，同時也無語言辨識問題，雖然如此，現今大多數網頁中的日期並未使用ISO 8601的格式，這造成在處理網頁日期時很大的困難。

在網頁日期的處理方面，目前已有一些相關的研究，其中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」的專利（註8），其「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。雖然此一專利能自動擷取URL中的日期並做辨識，但在日期辨識時所依據的日期格式需經特別指定，否則便直接採用瀏覽器上的預設日期格式，如：美國地區的預設日期格式為mmddyyyy，而歐洲地區的預設日期格式為ddmmyyyy。不能針對網頁內容來處理，是此專利技術的另外一大缺憾。另外，此項專利的研究並不包含中文的日期格式，更遑論臺灣地區所慣用的民國年。

在2006年10月IBM公司另外一項“System and method for searching dates in electronic documents”的專利中，使用正規表示式（Regular Expressions，簡稱RE）從網頁內容中擷取各種格式的日期（註9）。然而根據此項專利的說明書，其處理的日期格式為英文的格式，此外文件中也未有任何實驗數據或成效的展示。最重要的一點，此項專利仍然是以全文檢索方式來擷取，換言之，只將所有日期取出並放入資料庫，並未嘗試替使用者辨識那個日期是此網頁的創造（或修改）日期，因此從有效率檢索和資訊超載的

角度來看，是有很大的缺失。

日期是網頁Metadata中的重要欄位之一，但是由於在使用華文的地區，日期格式的書寫習慣上會穿插中文文字做為年、月、日的區隔（註10），在臺灣更有以民國年做為紀年，使得中文網頁中的日期形式較為複雜，因而增加了自動著錄日期時判斷的複雜度。本研究先針對臺灣地區中文網頁中的日期格式（或形式）做抽樣統計分析，然後利用分析整理的結果，設計一套正規表示式來自動擷取和辨識網頁的創造（或修改）日期，希望能達到拋磚引玉的作用，期望未來能有更多中文網頁Metadata欄位自動著錄的相關研究，來提高Metadata在資料檢索上的利用與成效。

貳、實作方法與流程

如上所述，本研究是採用正規表示式來自動擷取和辨識網頁的創造（或修改）日期。正規表示式的主要用途是利用所指定搜尋的字串樣版，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：

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

為使本研究所抽取的網頁樣本能盡量符合目前臺灣地區繁體中文網頁的整體現況，以下詳述網頁樣本的抽樣步驟：

一、本研究所採用的詞庫是由飛資得資訊股份有限公司所提供，詞庫共有136,961個使用者曾經使用的檢索詞彙。首先經由亂數從該詞庫中取得200個檢索詞彙。
二、將每個檢索詞彙利用Google搜尋，並提取Google所回覆的前100筆網頁資料。
三、再經由亂數從前項的100筆網頁資料中隨機抽取10筆網頁。
四、每個檢索詞彙取得10筆隨機抽樣的網頁，全部200個檢索詞彙，共取得2,000筆網頁樣本。
五、由於討論區以及部落格類型的網頁通常會記錄下所有參與者發表言論的日期，這些日期一般而言較無參考價值，因此予以排除。所以在排除這二類型的網頁後，剩下1,018筆有效的網頁樣本。

接下來網頁樣本的自動擷取和辨識處理流程如下：

一、將每個樣本網頁轉換為純文字檔。
二、人工建立每個樣本網頁的創造（或修改）日期，做為自動擷取和辨識的正確率計算之用。
三、分析了解中文網頁資源的日期格式（或形式）和存在比例，並利用整理分析的結果來設定抓取日期的正規表示式。

四、將設定好的正規表示式公式帶入（純文字檔網頁）並擷取和辨識網頁中的日期。

五、計算網頁創造（或修改）日期自動辨識的正確率。

參、研究結果與分析

首先是1,018筆有效網頁樣本日期格式（或形式）的分析整理結果，本研究根據網頁內容將其分成三大類型：

一、新聞類型網頁（數量有102筆，約為10%）：主要是包含新聞相關網頁，此類型網頁內容往往包含人、事、時、地、物這幾項元素，由於這些特性，所以新聞類型網頁存在日期資料的比例相當高。

二、學術類型網頁（數量有348筆，約為34%）：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。

三、一般類型網頁（數量有568筆，約為56%）：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。

在新聞類型的102筆網頁中，有日期資料的網頁共95筆，占這類型網頁的比例為93%，其中的日期格式則包括如表一所示的幾種不同形式。

在學術類型網頁中，共有348筆網頁，其中有日期資料的網頁為173筆，占這類型網頁的比例約為50%，其中的日期格式則包括如表二所示的幾種不同形式。

在一般類型的568筆網頁中，有日期資料的網頁共164筆，占這類型的網頁比例約為29%，其中的日期格式則包括如表三所示的幾種不同形式。

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”做為區隔	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%

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”做為區隔	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”做為區隔	5	民國99年4月26日	2.89%
總數	173		100.00%

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”做為區隔	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”做為區隔	2	民國99年4月26日	1.22%
總數	164		100.00%

表四　三種類型網頁之日期格式綜合比較統計表

使用的日期格式	比　例		
	新聞類型	學術類型	一般類型
西元年“/”月“/”日	69.47%	79.19%	60.37%
西元年“年月日”做為區隔	20.00%	6.94%	17.07%
西元年“.”月“.”日	3.16%	0.58%	4.88%
民國年“.”月“.”日	3.16%	0.00%	1.83%
使用英文月份縮寫	3.16%	5.20%	3.66%
後置西元年份	1.05%	0.58%	9.76%
使用中文月份	0%	4.62%	1.22%
民國年“年月日”做為區隔	0%	2.89%	1.22%
總數	100.00%	100.00%	100.00%

經過分析與統計，此次抽樣中發現臺灣地區網頁之日期格式存在著至少如表五所示的八種不同日期格式，同時在此次分析的1018筆網頁中，有432筆（約42%）的網頁含有日期資料，而有586筆（約58%）網頁沒有日期資料。

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占了將近七成，顯示此日期格式在臺灣地區網頁中被使用的最為廣泛。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，占14%。最後，在本次統計中，有極小部分的網頁，在日期資料部分只有年份以及月份，在此次的統計分析中並沒有將其獨立分類，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。

根據表五的臺灣地區網頁之日期格式統計表，已經知道臺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，並依其所占比例，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。

表六是依據表五的統計數據，所設計出來的正規表示式公式表，以及各公式之權重值：

在計算正確率方面，本次研究總共使用了兩種計算方式：

一、有日期部分網頁的正確率：是指

經由程式比對所判斷產生的第一

表五　臺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”做為區隔	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”做為區隔	7	民國99年4月26日	1.62%
總數	432		100.00%

表六　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	優先順序
西元年“/”月“/”日	[12]\d\d\d/\d{1,2}/\d{1,2}	69.91%	1
西元年“年月日”做為區隔	[12]\d\d\d年\d{1,2}月\d{1,2}日	13.66%	2
後置西元年份	\d{1,2}/\d{1,2}/[12]\d\d\d	4.17%	3
使用英文月份縮寫	d{1,2}-\D\D\D-[12]\d\d\d	4.17%	4
西元年“.”月“.”日	[12]\d\d\d[..]\d{1,2}[..]\d{1,2}	2.78%	5
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%	6
民國年“年月日”做為區隔	\d\d年\d{1,2}月\d{1,2}日	1.62%	7
民國年“.”月“.”日	\d\d[..]\d{1,2}[..]\d{1,2}	1.39%	8

順位日期，完成吻合人工所給予的（正確）日期的網頁比例。

二、無日期資料網頁的正確率：經由程式比對能正確的辨別出該網頁沒有日期的網頁比例。

如表七所示，在正確率方面，有日期網頁的成功筆數共有267筆，正確率為61.81%（267 / 432）；在無日期網頁部分，程式成功辨別無日期的筆數共有368筆，正確率為62.80%（368 / 586）。

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，本次研究中也將表六中的權重值倒置，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。

若依據三種網頁類型（有日期資料者）來分別計算其正確率，結果如表八所示，新聞類型網頁正確筆數共有59筆，正確率為62.11%（59/95）；學術類型網頁正確筆數共有121筆，正確率為69.94%（121/173）；一般類型網頁正確筆數共有87筆，正確率為53.05%（87/164）。

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和辨識日期方式來著錄日期欄位的可能誤差狀況，本研究也設計了一個稱為「誤差年」的計算方法，主要是計算（人工方式建立之）網頁正確日期欄位的年份，與經

表七　正確率統計表

	正確筆數	正確率
有日期網頁正確率	267	61.81%
無日期網頁正確率	368	62.80%

表八　各類型（有日期資料）網頁正確率統計表

網頁分類	正確筆數	總筆數	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%

由程式辨別出之第一順位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年份為2008年，其誤差年的值-2，依此類推。

表九是針對誤差年取絕對值後所做的統計，由於沒有日期資料的網頁無法計算其誤差年，所以只有計算有日期資料的432筆網頁，但是其中有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，因此也無法計算誤差年，所以在表九統計中的有效樣本數量為403筆。

表九　誤差年絕對值統計表（403筆網頁）

	最小值	最大值	平均數	標準差
誤差年絕對值	0	33	0.62	2.32

表十　誤差年次數分配表

誤差年	次數	百分比
-33	1	.2
-8	2	.5
-7	2	.5
-6	1	.2
-4	3	.7
-3	4	1.0
-2	6	1.5
-1	16	4.0
0	336	83.4
1	8	2.0
2	6	1.5
3	4	1.0
4	5	1.2
5	1	.2
7	3	.7
8	2	.5
9	1	.2
11	1	.2
14	1	.2
總和	403	100.0

表十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（267筆正確網頁＋69筆年份正確但月或日有錯誤的網頁），占有效樣本403筆的83.4%（336 / 403）。由於使用者在檢索網頁的日期時，絕大多數只使用年份來檢索，因此在一定程度上已顯示自動擷取和辨識網頁創造（或修改）日期欄位的可行性。

肆、結語

在本次研究中，以隨機抽樣的方式取得臺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，分析結果顯示網頁日期格式並未完全依循

ISO 8601標準（至少有八種不同日期格式），所以在做自動辨識日期資料時，需要比對及過濾多種不同的日期格式。

在剛開始設計比對公式時，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，但是成效不佳，且造成許多無法預期的雜訊，有日期資料網頁的正確率下降到四成左右。經過詳細分析，發現這些多餘的雜訊很難透過修正公式來排除，唯一的辦法似乎只能先放棄只有年份跟月份的正規表示式公式，才能去除這些多餘的雜訊，因此在本次實驗所使用的正規表示式公式均有年、月、日。

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，再加上國內網頁在日期表達上有多種形式，所以要針對日期資料來自動辨識，進而自動著錄日期欄位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來做，的確還有一段路要走。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來執行，但是由於計算出的平均誤差年約為0.6年，且83.4%的網頁能精準預測其年份（即誤差年為0），因此在某種程度上已具有實用與參考價值，可以讓使用者根據自動辨別出的年份來篩選資料，成為一種輔助工具。只要將程式自動著錄的日期欄位加註特別標記，來跟人工著錄者區別，以免使用者混淆即可。換言之，只要應用得當，應該可以增進在尋找和過濾網路資料時的效率。

由於目前文獻上關於網頁日期欄位自動著錄的相關研究非常少見，所以本研究抱著投石問路的態度來進行實驗，雖然實驗結果顯示（有日期資料網頁及無日期資料網頁）的正確率都只有約六成，因此以目前的狀況還是未能完全取代人工，但是若將來配合其他更精確的技術，仍有可能進一步來提升正確率，進而提昇檢索效率。

註釋

註 1：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（1998年1月）：頁36-54。

註 2：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（1997年12月）：頁11-25。

註 3：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（2007年冬）：頁249-266。

註 4：Lloyd Sokvitne, “An Evaluation of the Effectiveness of Current Dublin Core Metadata for Retrieval”, paper presented at the Proceedings of VALA2000 (Melbourne,

Australia, February 16-18, 2000), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.2757& rep=rep1&type=pdf (accessed August 15, 2010).

註 5：International Organization for Standardization (ISO), “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 (accessed August 15, 2010).

註 6：World Wide Web Consortium (W3C), “Date and Time Formates”, http://www.w3.org/TR/NOTE-datetime (accessed August 15, 2010).

註 7：International Organization for Standardization (ISO), “ISO 8601:2004”, http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=40874 (accessed August 15, 2010).

註 8：International Business Machines Corporation (IBM), “Retrieving dated content from a website”, http://ip.com/patapp/CN1758243 (accessed September 16, 2010).

註 9：International Business Machines Corporation (IBM), “System and method for searching dates in electronic documents”, http://www.chemyq.com/patentfmen/pt95/948875_F08A6.htm (accessed September 16, 2010).

註 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* 24卷8期（2008年8月）：頁58-62。