Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 237-262

學術會議資訊之擷取及其應用

Information Extraction for Academic Conference and It's Application

陳光華*

Kuang-hua Chen

摘要

網際網路已成爲學術訊息傳播的主要管道，本研究關注擷取網際網路上學術研究人員關心的學術會議訊息，提供會議主題、時間、空間等訊息，企望減輕研究人員蒐集與管理會議資訊的負擔，進而提升學術研究出版的效率。本研究首先提出一套學術會議資訊檢索與擷取的自動程序，並藉由實驗確認其可行性，實驗結果顯示文件分類績效 F1 measure 超過 80%；具名實體擷取績效 Recall 超過86%，F1 measure超過70%。繼而實際開發學術會議檢索與擷取系統平台，提供文件檢索、資訊擷取、分類瀏覽、行事曆等功能，整合研究人員的學術活動與日常行程安排，展示前述學術會議資訊檢索與擷取程序的實用性。

關鍵詞：學術資訊、資訊擷取、資訊檢索、具名實體

Abstract

Internet has become a major channel for academic information dissemination in recent years. As a matter of fact, academic information, e.g., "call for papers", "call for proposals", "advances of research", etc., is crucial for researchers, since they have to publish research outputs and capture new research trends. This study focuses on extraction of academic conference information including topics, temporal information, spatial information, etc. Hope to reduce overhead of searching and managing conference information for researchers and improve

*國立臺灣大學圖書資訊學系
Department of Library and Information Science, National Taiwan University
E-mail: khchen@ntu.edu.tw

efficiency of publication of research outputs. An automatic procedure for conference information retrieval and extraction is proposed firstly. A sequence of experiments is carried out. The experimental results show the feasibility of the proposed procedure. The F1 measure for text classification is over 80%; F1 measure and Recall for extraction of named entities are over 86% and 70%, respectively. A system platform for academic conference information retrieval and extraction is implemented to demonstrate the practicality. This system features functionalities of document retrieval, named entities extraction, faceted browsing, and calendar with a fusion of academic activities and daily life for researchers.

Keywords: Academic Information, Information Extraction, Information Retrieval, Named Entities

1. 緒論

在全球化的趨勢之下，大學的學術評價更加受到前所未有的重視，有各式各樣以全球大學爲標的之學術評鑑報告陸續公告周知，如上海交通大學（ARWU, 2010）與英國 Quacquarelli Symonds（QS, 2010）所做的世界大學排名。此外，Thomson Reuters 公司的 SCI、SSCI、A&HCI 等資料庫，以及 Journal Citation Report（JCR），提供的統計數據，往往成爲各國評鑑國內大學學術成果的計量指標。在這種激烈的學術競爭環境之下，且學術競爭力被視爲國家競爭力的一環，大學教授莫不兢兢業業地、努力地從事學術研究。學術研究人員掌握學術會議資訊的即時性與確實性，對於其研究工作的進展與研究成果的發表，是非常重要的。本研究在這樣的背景下，研發學術會議資訊檢索與擷取系統，希望能夠有效地由充斥浮濫資訊的網際網路，擷取相關的學術會議資訊。

學術研究人員的學術活動是非常多元的，學術資源服務的類型眾多，本研究將著重於以資訊擷取爲基礎的學術會議資訊的檢索與擷取。研究人員的學術活動中很重要的一項便是「學術研究的出版」，學術的出版有兩個主要的方向，一個是學術會議，另一則是學術期刊。會議的 Call For Paper 有時間的期限，而期刊 Special Issue 的 Call For Submission 也有時間的期限，協助研究人員掌握這些重要的訊息，自動地由網路擷取學術會議的時間訊息、空間訊息、與主題訊息，協助研究人員管理時間與空間訊息，將有很大的助益。若能進一步搭配「行事曆（calendar）」的功能，對於研究人員而言更是事半功倍的。換言之，一般行事曆功能僅提供使用者新增資訊、更新資訊、刪除資訊，爲了搭配學術研究的出版，行事曆必須有更進階的功能，能夠依據使用者的 profile 搜尋 Call For Paper 與 Call For Submission，填入行事曆，並依據使用者的設定，提供警示（alert）的服務。

研討會通知或會議論文投稿須知，一般是透過既有的郵寄目錄發送，或是以網頁文件的形式發佈，也因此訊息傳播的目標通常局限於特定族群及研究機構。即使使用者自行利用網頁搜尋工具在網際網路上查找，所取得的資訊可能不完整，或是已錯過參與的時機。若要提供即時的且整合的研討會相關資訊，蒐集網際網路上與研討會通知相關網

頁的自動機制，是重要的一環。

一般在網路上大量蒐集網頁的方式，通常利用網頁擷取機器人（web crawler）到處拜訪網站並擷取所有網頁內容。由於 Web Crawler 的建置困難度較高，維護與效能控管也較爲複雜，不當的設計常會佔據網路頻寬資源，或導致被網站封鎖而無法擷取內容。因此另有一種方式，並不採用傳統的 web crawler 而是修改網頁擷取機制，以適當的關鍵字與網頁搜尋引擎的整合來蒐集網頁。

目標式網頁擷取（focused crawling）是一種蒐集研討會通知資訊的方式。有別於一般 Web Crawler 漫無目的地抓取所有的網頁，Focused Crawling 會先過濾與主題無關的內容，也就是會應用一組特定主題的關鍵詞，用以訓練並建立文件分類機制，再由此分類機制引導 crawler 擷取與主題相關的網頁。（Chakrabarti, van den Berg, & Dom, 1999）另外還可以將 Focused Crawling 稍加變化，依據一組系統已經記載的研討會議網站清單，反向地蒐集相關網頁文件，這種網頁資料蒐集的替代方案被稱爲反向式網頁擷取（backward crawling）。（Brennhaug, 2005）這種網頁蒐集機制首先以主題關鍵字，透過搜尋引擎取得相關網頁的網址及網頁內容，以建構候選相關文件集。再接續利用搜尋引擎的反向連結查詢功能（back link query），一併蒐集連結到候選文件的網頁。又考量到這種由反向連查詢所得的網頁也有可能再連結到其他研討會議網頁，所以再繼續以正向連結（forward crawling）擷取該網頁中的其他 URL，以發掘潛在的相關網頁。此程序將會一直重覆執行直到重覆的次數達到預設的門檻。

若以蒐集研討會議徵稿通告的相關資訊來檢視網頁自動擷取機制，無論是正向或反向擷取，都會面臨下列兩項議題：(1)網路上傳播的研討會會議資訊經常更新，例如投稿截止日期的延期、會議地點資訊的更新、或是新加入的 workshop 議程等等，而所蒐集的研討會會議資訊必需能夠即時反應各項更新資訊。(2)目前雖然將「研討會議通知資訊」定義爲與研討會議相關的訊息通知網頁，但網頁內容通常包含許多與研討會無關的各種式樣各種規格的其他資訊，例如文字或影音廣告，網站目錄選項，或其他網站連結等，這也造成在擷取網頁機制建置時，文件相關程度判斷的問題。

本研究基於前述的背景，運用網頁搜尋技術，以及資訊檢索與擷取技術，發展一套學術會議資訊檢索與擷取的自動程序，並實際建構系統平台，以服務學術研究人員。本文的結構如下：文獻探討一節將說明資訊擷取的技術，運用於學術會議檢索的情形，相關資訊服務系統的現況；學術會議資訊蒐集一節討論由網際網路蒐集學術會議資訊的方法，以及過濾不相關資訊與雜訊的作法；資訊擷取模型之訓練與建置一節探討學術會議資訊擷取模型的訓練與建立;系統實作與功能一節討論系統實作的方法,以及各項功能；最後則是簡短的結論。

2. 文獻探討

學術會議資訊之檢索屬於資訊檢索的應用研究，其中牽涉的研究議題眾多，至少有具名實體的辨識（named entities identification）、分群歸類（clustering and classification）、

文件檢索（text retrieval）。然而，若要建置完整的應用系統，則牽涉更多的技術，如時間與空間資訊的搭配，各種 API 應用元件的整合。本研究嘗試建構學術會議資訊檢索與擷取系統，首先探討資訊檢索與擷取技術的現況，以及現有檢索系統的發展。限於篇幅，本文並不嘗試進行全面而完整的相關文獻的探討。

學術會議資訊文件含有許多具名實體，包括會議名稱、會議時間、會議地點、會議主題、截稿日期等等，已有許多學術論文探討這個研究課題，訊息理解會議（Message Understanding Conference，簡稱 MUC）是第一個將具名實體的辨識視爲一項檢索研究的評量項目，企圖推動資訊檢索研究社群，投注研究能量，發展更新的技術，提昇具名實體辨識的績效。（MUC, 2001）訊息理解會議認爲不僅僅需要辨識重要的實體，還必須確認實體之間的關係（relationship），MUC-6 則明確地規範三個層次的資訊擷取的研究議題：具名實體之辨識、照應詞之解析、樣版資訊之建構。照應詞之解析是串連具名實體及其對應的照應詞（如代名詞）；腳本樣版則是依照預先訂定的樣版，由文件中擷取相關的資訊填入樣版的欄位。（Grishman & Sundheim, 1996）

雖然具名實體辨識的研究很早就開始了，但是學術會議資訊擷取的研究則是比較不受到許多研究者的關注。Lazarinis（1998）提出應該應用資訊擷取技術進行論文徵稿通告（call for paper，簡稱 CFP）的檢索，有別於傳統上僅以文件檢索技術檢索 CFP。Lazarinis 發現這種作法在固定 Recall 的情形下，可以提昇 45%-60%的 Precision，這項研究確認應將學術會議資訊的檢索，視爲資訊擷取的問題，而非單純的文件檢索的問題。

Schneider（2005）應用 Conditional Random Fields（CRF）模型，擷取 CFP 的重要訊息，Schneider 特別關注文件版面特徵（layout features）的貢獻，發現版面特徵可以提昇約 30%的 F1 分數（F1 measure）。因爲，Schneider 的研究關注於各項特徵的效益，使用的測試資料僅有 263 篇乾淨無雜訊的 CFP，而避開真實文件各種複雜的情況，因此很難建構一個實際可行的資訊服務系統。

目前亦有許多學術組織，建構了 Conference Calendar 的相關網頁，希望有利於會議資訊的流通，但是這種資訊彙整形式的網頁，僅提供瀏覽的功能，沒有進階檢索功能，使用者仍須耗費相當的精力，才能瀏覽相關的會議資訊。另外，尚有功能比較好的類似系統，例如 WikiCFP 與 EventSeer 等 CFP 資訊共享服務系統，但是提供的多爲電腦科學相關學術領域的學術會議資訊。WikiCFP（http://www.wikicfp.com/）是使用 Wiki 建構的 CFP 共享系統，資訊來源是依賴使用者提供相關會議資訊；EventSeer（http://eventseer.net/）是一個 Web 2.0 的網站，企圖建構一個電腦科學研究的社群網站，除了允許登錄使用者自由發佈學術資訊外，另外運用 Robot 主動搜集網際網路上的 CPF 資訊。

Takada（2008）建構的 ConfShare 資訊服務系統，透過瀏覽器提供學術會議資訊檢索的服務。Takada 認爲研究者爲了參加學術會議學習最新的研究成果，或發表本身的研究成果，都需要蒐集學術會議的相關資訊。蒐集資訊的工作是參加會議不可缺乏的，但也造成研究者不小的負擔。ConfShare 以使用者（亦即研究者）的角度，提供與學術會議相關資訊的各種服務，希望能夠減輕前述研究者的額外負擔。

Xin, Li, Tang, and Luo（2008）使用 Constrained Hierarchical CRF（CHCRF）標註學術會議官方網站的網頁以及屬性，企圖建構一個學術會議的行事曆系統。Xin 等人關注的是學術會議的官方網站而非 CFP，然而官方網站成立的時間通常都很晚，不像 CFP 的快速與即時，而且，官方網站的資料是透過下達會議名稱與時間，由 Google 檢索而得，這樣的假設並非很合理，因爲，類似的系統應該是藉由學術研究的主題取得學術會議資訊，而非藉由特定的會議名稱或是舉辦時間。

本研究企圖建構的學術會議資訊檢索與擷取系統（Academic Conference Information Retrieval and Extraction System，ACIRES），較接近於 Takada（2008）的 ConfShare 系統，但是在功能面仍有差異，使用的技術亦不相同，涵蓋的學科主題範疇亦有很大的差異。下文將說明本研究的資訊的蒐集、處理、模型的訓練、以及系統的實作 。

3. 學術會議資訊蒐集

學術會議資訊的檢索與擷取，當然需要被檢索的標的物，必須有一套機制蒐集網路上的論文徵稿通告，作爲系統開發前，資訊擷取模型訓練之用；系統開發完成，正式運轉時，亦需要這套機制持續蒐集論文徵稿通告，以服務學術研究人員以及一般的使用者。

爲了有效地蒐集相關的學術論文徵稿通告，本研究採用目標式網頁擷取（focused crawling）的概念，先以學門分類表做爲各學科主題的查詢關鍵字，利用網頁搜尋引擎蒐集所需之論文徵稿通告。我們採用澳洲與紐西蘭標準研究分類表（Australian and New Zealand Standard Research Classification，簡稱 ANZSRC）爲主（Pink & Bascand, 2008），再整合 Wikipedia 提供的學術領域列表以補充新興學科。由於論文徵稿通告不一定會標示所屬學科領域，以學門分類名稱爲查詢關鍵詞所蒐集的論文徵稿通告，可能無法涵蓋各學科領域所有重要的研討會資訊。因此，可再進一步分析第一批搜集的論文徵稿通告的研究議題相關詞彙，整合到學科主題關鍵詞列表，形成所謂的 bootstrapped crawling，讓學術會議資訊的蒐集更爲廣泛且完整。表 1 依字母順序，簡要列出部分之主題關鍵詞。

利用前述的主題關鍵詞，透過 Google 搜尋引擎，分別取得查詢結果前五十筆最相關的網頁，再接續依相關網頁的內容執行一次正向連結查詢（forward link query），一併收錄該五十筆網頁中超連結所指到的網頁。透過網頁搜尋引擎，可一次性地蒐集大量的相關網頁，但無法掌控網頁提供的會議資訊是否已過期。再考量研討會資訊的提供，必須符合即時性與時效性，因此再進一步利用網頁快訊服務（Google Alert），補充最新的研討會資訊。

網頁快訊服務就是當新的網頁發佈於網際網路時，網頁搜尋引擎比較該新網頁與使用者預設的 profile 的相關度，若是在搜尋結果的前 20 名內，就會立即以電子郵件通知快訊訂閱客戶。利用此服務特性，將前述的學科主題關鍵詞，做爲取得快訊的搜尋詞彙，即時取得最新發佈的網頁文件。對於以網頁快訊服務取得的相關網頁，本研究也會進一步執行一次正向連結查詢。

無論是從網頁搜尋引擎或是網頁快訊服務蒐集而得的網路資訊，必定會有重覆的情形，因此在蒐集網頁時，必須初步過濾重覆的網頁。以網頁搜尋引擎取得的相關網頁，由於是同一時間取得的網頁內容，因此不需考量網頁更新的因素，直接比對網址過濾重覆者。以網頁快訊服務取得的新網頁，若網址與現有文件相同，則必須考量網頁更新因素，先比對兩筆網頁的上次更新時間，再保留更新時間較近的網頁。若無法取得網頁的上次更新時間，則保留由網頁快訊服務取得的網頁。

由於從網頁搜尋引擎及網頁快訊服務廣泛蒐集的網頁數量龐大，大量的文件中可能包含與研討會論文徵稿通告無關的網頁，爲了提升學術會議資訊自動標註的準確度，必須篩選無關的網頁文件。本研究運用文件自動分類技術，可以迅速處理大量文件，避免繁瑣且冗長的人工分類作業，我們採用開放程式碼 Rainbow Classifier 自動過濾非會議徵稿通告的網頁文件。（McCallum, 1996）由於 Rainbow Classifier 需要一組已分類的文件做爲分類模型所需的訓練文件，此訓練文件將利用人工分類的方式產生，該人工分類的作業一併整合至人工標註輔助系統，讓標註人員可同時並行訓練文件分類與文件內容標註工作。

表 1. 部分主題關鍵詞

abnormal psychology accompanying accounting scholarship acoustic engineering acoustics acting actuarial science adapted physical education admiralty law advertising aerobiology aeronautical engineering aerospace engineering aesthetics affine geometry african studies agricultural economics agricultural education agricultural engineering agrology agronomy air force studies algebraic computation algebraic geometry algebraic number theory algebraic topology american history american politics american studies analytical chemistry ancient egyptian religion ancient history animal communications animal science animation anthropology of technology apiculture appalachian studies applied psychology approximation theory aquaculture architectural engineering archival science art education art history artillery arts administration asian american studies asian studies associative algebra astrobiology astronomy astrophysics atheism and humanism atomic, molecular, and optical physics australian literature automotive systems engineering beekeeping behavioral geography behavioural economics behavioural science bilingual education biochemistry bioeconomics biogeography bioinformatics biological psychology biology biomechanical engineering biomedical engineering biophysics black studies or african american studies botany business administration business english business ethics calligraphy campaigning canadian literature canadian studies canon law cardiology cardiothoracic surgery cartography category theory cell biology celtic studies chamber music chemical engineering cheminformatics chemistry education chicano studies child welfare children geographies chinese history chinese studies or sinology choreography christianity chronobiology church music civics civil procedure classical archaeology classics climatology coastal geography cognitive behavioral therapy cognitive psychology cognitive science collective behavior combat engineering communication design communication engineering

4. 資訊擷取模型之訓練與建置

學術會議的論文徵稿通告主要包含會議名稱、會議地點、會議時間、會議主題、會議官方網站、以及各項截止日期或公佈日期等。論文徵稿通告與一般文件最大的差異在於其重要資訊不一定是以完整的語意文句組成，可能利用內容配置及排版以突顯各項資訊。例如，一份論文徵稿通告的會議名稱通常單行置中且前後各有空行，研討會議題以項目符號逐項表列，各項重要期限或公佈日期通常利用表格呈現。除了排版上的特色之外，還可利用特定詞彙判斷是否爲重要通知資訊，例如會議名稱通常會出現 conference、international、annual 等詞彙，submission、notification、deadline 等詞彙則經常伴隨日期出現，另外也可以利用完整的地名詞典擷取會議舉行地點。雖然可利用排版及詞彙兩種特性設計論文徵稿通告的資訊自動擷取機制，但是網路上或電子郵件提供的論文徵稿通告，並沒有一致的文件格式，通知項目也沒有統一的名稱，這都增加資訊判斷的困難度。

本研究應用 Conditional Random Field（CRF）建立自動擷取會議資訊的模組，從會議通告網頁文件，擷取重要的會議資訊欄位（如會議名稱，會議日期，會議地點等）。CRF 爲機器學習式（machine learning-based）演算法，需設定數種資料特徵以訓練模型，因此以學術會議徵稿通告必備的重要資訊項目，作爲資料特徵欄位（如表 2 所示），再使用一部分學術研討會徵稿通告，做爲訓練文件集，先以人工的方式標註特徵欄位，並利用特殊詞典或地名資料庫標示特定詞彙（例如地名、會議專有名詞等），建立 CRF 學習樣版，再經由 CRF 自動學習與測試，調整資訊辨識的準確度，以建置資訊擷取的自動機制。

CRF 是在機率演算的架構之下，針對某種結構組成的文字資料進行分段（segment）或是標註（label）的工作，其文字資料結構包含序列式或是矩陣式等。某些機器學習的演算法必須假設每一個序列資訊都是相互獨立，例如 Hidden Markov Model（HMM），但是真實世界的序列資料並不是由一連串獨立的資訊組成的。CRF 不同於其他機器學習演算法，會考量隨機序列資訊的關聯性，以求整體序列的聯合條件機率，以避免詞彙標註的偏置（bias）問題（Wallach, 2004）。本文並不試圖詳細描述 CRF 的理論與技術，相關說明請參考（Sutton, Rohanimanesh, & McCallum, 2004; Lafferty, McCallum, & Pereira, 2001）。

表 2. 徵稿通告之特徵及對應之標籤

中文名稱	英文名稱	HTML 標籤	標籤範例
會議全名	Conference Name	confname	<confname> Multimedia in Ubiquitous Computing and Security Services</confname>
會議名稱縮寫	Abbreviation of Conference Name	confabbr	<confabbr> MUCASS 2008 </confabbr>
會議地點	Conference Location	confloc	<confloc> Hobart, Australia </confloc>
會議日期	Conference Date	confdate	<confdate> October 14-16, 2008 </confdate>
會議網址	Conference Website	confwebsite	<confwebsite> http://www.sersc.org/MUCASS2008 </confwebsite>
會議主題	Conference Topic	conftopic	<conftopic> Real-time and interactive multimedia applications </conftopic>
報名截止日期	Registration Deadline	registdue	<registdue> Registration - 15th October, 2007 </registdue>
摘要提交截止日期	Abstract Submission Due	abstractdue	<abstractdue> Deadline for abstract 11 June 2008 </abstractdue>
摘要錄取通知日期	Abstract Notification	abstractnotify	<abstractnotify> Acceptance of papers - August 30, 2009 </abstractnotify>
論文提交截止日期	Paper Submission Deadline	submissiondue	<submissiondue>February 15 23, 2009 - Paper submission</submissiondue>
論文錄取通知日期	Author Notification	authornotify	<authornotify> March 23, 2009 - Author notification </authornotify>
論文定稿截止日期	Final Paper Due	finalpaperdue	<finalpaperdue> Camera-ready copies: April 7, 2009 </finalpaperdue>
海報論文截止日期	Poster Paper Due	posterdue	<posterdue> Poster Paper Submission Deadline May 15, 2008 </posterdue>
專題提案截止日期	Workshop Proposals Due	workshopdue	<workshopdue> workshop submissions due : Sunday, 2 Mar 2008 </workshopdue>
教學提案截止日期	Tutorial Proposals Due	tutorialdue	<tutorialdue> Tutorial Proposals: June 30, 2003 </tutorialdue>
博士生論壇投稿截止日期	Doctoral Consortium Due	doctoraldue	<doctoraldue> Doctoral consortium submissions due: 6 Apr 2008 </doctoraldue>

整體工作流程如圖 1 所示，包含文件前置處理、分類模型的訓練、CRF 模型的訓練三項工作。文件前置處理包含去除文件雜訊、標註學術會議資訊、Tokenization 與詞彙特性標示。

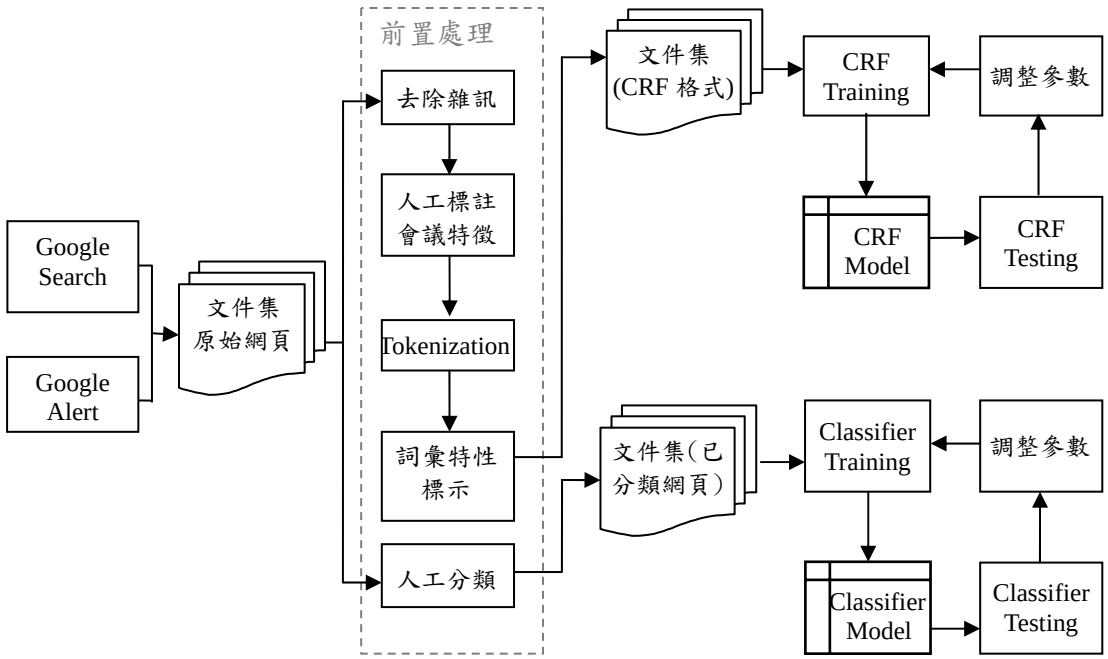

圖 1. 學術會議資訊檢索與擷取自動模型之建置流程

4.1 文件前置處理

4.1.1 去除文件雜訊

由於由網際網路蒐集的文件，通常爲 html 的網頁，包含許多各式各樣的資訊，除了該網頁的主要內容之外，尚有網頁相互連結的資訊，以及網站外部的延伸資訊。有些網頁的作者爲讓網頁更吸引使用者瀏覽，採用了動態網頁或是多媒體的呈現模式，增加處理網頁內容工作的複雜度。無論在資訊擷取的訓練階段或是正式的應用上，過多與會議資料無關的雜訊將會影響資訊欄位判斷的精確度，因此必須先去除與網頁內容主體無關的雜訊，包含廣告，圖片，網站目錄，視覺特效相關程式段落等等。

4.1.2 標註學術會議資訊

建構自動文件分類機制以及自動資訊擷取模型，需要大量的訓練資料，本研究另外建置類別標註系統（Genre Annotating System，GAS），整合內容標註與文件分類二大功能，以求內容特徵標註與文件分類標註的一致性與效率。GAS 以瀏覽器爲系統平台，爲典型的 Web-Based Application，主要功能分成三部分：候選文件瀏覽、文件分類標註，以及內容特徵標註。圖 2 爲本研究建構之類別標註系統的操作畫面。

1. 候選文件瀏覽區

圖 2 右上方的功能區塊爲候選文件瀏覽區。如前文所述，候選文件是以學門分類表的學科名稱爲關鍵字，經由 Google Search 及 Google Alert 於網路上蒐集與會議論文徵稿通告相關的網頁文件集合，經由去除雜訊處理之後，自動載入 GAS 系統。標註人員登入 GAS 後，系統會於候選文件瀏覽區展示由該人員負責標註之文件清單，標註人員也可以利用左方的查詢功能篩選網頁文件，清單上同時標示每份候選文件的標註狀態及記錄。

圖 2. GAS -功能畫面

2. 文件分類標註區

文件分類標註區位於圖 2 系統功能畫面中間的狹長矩形區塊。候選網頁文件主要分成相關與不相關兩類，所謂的相關與不相關，是以該網頁文件是否與會議論文徵稿通告相關與否，作爲判斷的依據。但是，考量有些網頁文件內容資訊太複雜而無法斷定，也可以暫時不將該網頁歸類，且可以註記無法歸類的原因，作爲後續文件分類例外處理的參考，如圖 3 所示。標註人員從內容特徵標註區可檢視網頁文件，判斷該文件內容是否是會議論文徵稿通告，若確定是會議論文徵稿通告，才需要進一步針對文件內容標註各項會議資訊。

3. 內容特徵標註區

內容特徵標註區位於圖 2 的 GAS 系統功能畫面的下方功能區塊。選取候選文件瀏覽區的任一筆資料，系統會將該網頁文件全文載入內容特徵標註區，內容特徵標註區係以 HTML 模式呈現網頁文件內容。內容特徵標註區上方的功能列，除了提供「復原動作」、「重覆動作」、「去除 HTML 標籤」、及「字串查詢」等功能按鈕之外，最重要的功能是「樣式」的下拉式選單，此樣式選單列出所有本研究採用的會議資訊特徵，標註人員於網頁內容中框選特徵資訊後，再選取對應的會議資訊特徵樣式，標註之後，所選取的特徵資訊會以特定的 HTML 標籤標示。例如會議名稱在 HTML 原始碼中標示爲 <confname>會議名稱</confname>，本研究考量的會議資訊特徵與對應的 HTML 標籤請再次參見表 2。

圖 3. GAS -文件分類標註區

4. Tokenization 與詞彙特性標示

CRF 需切割序列性資料爲一連串 Token 後，並賦予各 Token 適當的詞性標示，再依每個 Token 的特徵向量，計算各 Token 之間的條件機率，以做爲建構詞彙辨識模型的依據。因此去除雜訊後的網頁內容，要再抽取非 HTML 標籤的字串，將字串以單一詞彙或標點符號爲單位，切割成更小的片段爲 Token，針對每一個 Token，進一步做一般詞性標示及專門詞性標示。一般詞性標示包含標點符號，大小寫，數字，日期型態等識別。專門詞性則包括地名，會議資訊經常使用專門詞彙，例如 conference、congress、

association、annual、national 等，本研究採用 GeoNames 地名資料庫爲地名辨視依據，並整理會議資訊經常使用的專門詞彙，用以比對並標示相關詞彙，如表 3 所示。

表 3. 會議資訊使用之專門詞彙列表

專門詞彙類別	詞彙項目
機構名稱	Center, centre, college, department, institute, school, univ., university
組織名稱	Association, consortium, council, group, society
事件名稱	Colloquium, conf., conference, congress, convention, forum, meeting, round, roundtable, seminar, summit, symposium, table, track, workshop
時間屬性名稱	Annual, autumn, biannual, biennial, European, fall, int., interdisciplinary, international, joint, national, special, spring, summer, winter

4.2 分類模型的訓練

文件分類的目的是爲了預先過濾並非論文徵稿通告的文件，以降低內容自動標註時的負擔。當系統運轉後，大量的網路文件進入系統時，必須先判斷是否爲論文徵稿通告的相關文件，然後再透過內容特徵擷取功能，擷取所需要的會議資訊。由於目前有許多的開放程式碼可供使用，以開發文件分類的功能模組，本研究使用 McCallum（1996）的 Bow Library，開發統計學習爲本的文件自動分類功能模組，用以過濾由網路取得的會議通告文件，Rainbow 則是基於 Bow 的應用程式，可由 http://www.cs.cmu.edu/~mccallum/bow/rainbow/取得。基本上，Rainbow 是利用已知類別的文件，統計分析各文件特徵並建立分類模型，再依此分類模型對新文件進行自動分類。在人工標註輔助系統所產生的相關文件集與不相關文件集，是收錄原始網頁文件，而不是已被人工標註特徵項目的新網頁內容，因爲本研究的會議資訊自動擷取系統，是先過濾非會議通告網頁，才進行資訊擷取程序，因此文件自動分類功能模組，是以原始網頁做爲訓練文件。我們進行大量的訓練與測試，使用 k-Nearest Neighbor（kNN）、Naive Bayes（NB）、Support Vector Machine（SVM）三種分類模式，隨機抽取文件進行 20 次的實驗，使用訓練文件與測試文件比例分別爲（7:3）、（5:5）、（3:7），觀察分類績效的變動情形，以決定系統使用的分類模型。分類結果的優劣是以 Recall（求全率）與 Precision（求準率）評量，可以進一步將兩項指標結合爲單一的 F1 指標，計算方式說明如下。每一篇文件皆已有正確的分類標記，在每一次的分類實驗，分類模型會爲每一篇自動賦予其分類標記，可能與正確的分類標記一樣，或是不一樣，因此有四種可能性，如表 3 所示。

依據表 4 可以計算 Recall (R)、Precision (P)、以及 F1 Measure。

$$P_i = \frac{TP_i}{TP_i + FP_i}, \qquad R_i = \frac{TP_i}{TP_i + FN_i}, \qquad F1 = \frac{2P_iR_i}{P_i + R_i}$$

表 4. 分類結果列聯表

Category *i*		Expert Assignment	
		TRUE	FALSE
System Judgment	TRUE	TP_i	FP_i
	FALSE	FN_i	TN_i

因爲進行了 20 次實驗，可以計算 Micro Recall、Micro Precision、Marco Recall、Macro Precision，以及對應的 Micro F1 Measure 與 Macro F1 Measure，以觀察每次實驗的變異情形，計算方式如下所示，其中 n 代表實驗次數。

$$P_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)}, \qquad R_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)}$$

$$P_{macro} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i}, \qquad R_{macro} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i}$$

實驗結果如表 5 所示，Outside Test 意指測試資料與訓練資料不同，Inside Test 意指測試資料與訓練資料相同。Inside Test 的結果一定會比 Outside Test 的結果好，如果 Outside Test 的結果很接近於 Inside Test，代表分類模型的適應性很好；訓練資料越多，涵蓋面越廣，分類結果也越好。

實驗結果顯示，SVM 模型的表現最好，Naive Bayes 次之，而 kNN 最差。SVM 在 Inside Test 與 Outside Test 的表現差異最小，而 Naive Bayes 變動的幅度很大，代表 SVM 模型對於未知資料的解釋性很強。除此之外，無論是何種模型，F_{micro} 與 F_{macro} 的表現相當，代表每一次實驗結果的變異性很小。值得注意的是，本研究是採用 Recall-Oriented 的作法，調整系統參數，進行文件的自動分類，原因是希望能夠儘量取得會議相關的文件，因此較著重於 Recall。依據前述實驗的結果，本研究發展的系統將採用 SVM 模型，自動分類大量的網路文件，判定是否爲 CFP 文件後，再進一步擷取文件中的會議資訊。

4.3 CRF模型的訓練

本研究使用 CRF 模型建構會議資訊擷取的自動程序，由於目前也已有許多現成的開放程式碼可供使用，決定採用 Kudo（2010）開發的 CRF++套件，以擷取會議論文徵稿通告的特徵資訊，CRF++可由 http://crfpp.sourceforge.net/取得。吾人可以使用 CRF++開發文件自動分詞（segmenting）或內容特徵標註（labeling）等序列性資料的應用系統。CRF++宣稱使用者可以自訂資料特徵，而且計算速度快，僅使用少量的記憶體。由於 CRF++使用特定文件格式，必須將文件內容切割成一連串的 Token，以表格的形式陳列每一個 Token 的詞彙特性、版面特性以及會議資訊等特徵，無論訓練文件或是測試文件，都必須依循此特定格式編排。

表 5. 分類結果績效比較

方法	訓練：測試	Inside/ Outside	P_{micro}	P_{macro}	R_{micro}	R_{macro}	$F1_{micro}$	$F1_{macro}$
SVM	**70**%：30%	Outside Test	75.30	75.34	92.07	92.07	82.84	82.87
		Inside Test	77.94	78.31	92.70	92.70	84.68	84.90
	50%：50%	Outside Test	74.19	74.21	90.36	90.36	81.48	81.49
		Inside Test	76.07	77.09	92.14	92.14	83.34	83.94
	30%：70%	Outside Test	72.90	72.93	89.10	89.10	80.19	80.21
		Inside Test	74.83	76.08	92.85	92.85	82.87	83.63
Naive Bayes	**70**%：30%	Outside Test	78.00	78.07	62.63	62.63	69.48	69.50
		Inside Test	75.29	75.50	95.30	95.30	84.12	84.25
	50%：50%	Outside Test	76.31	76.40	63.02	63.02	69.03	69.07
		Inside Test	75.28	75.59	94.18	94.18	83.68	83.87
	30%：70%	Outside Test	69.76	69.85	95.37	95.37	80.58	80.64
		Inside Test	74.84	75.51	96.33	96.33	84.23	84.66
kNN	**70**%：30%	Outside Test	66.97	69.32	58.67	58.67	62.54	63.55
		Inside Test	56.88	57.39	94.73	94.73	71.08	71.48
	50%：50%	Outside Test	65.74	67.77	61.82	61.82	63.72	64.66
		Inside Test	56.14	56.54	95.70	95.70	70.77	71.09
	30%：70%	Outside Test	63.51	67.03	58.67	58.67	60.99	62.57
		Inside Test	57.98	59.23	91.42	91.42	70.96	71.89

完成人工標註的網頁文件轉換成此特定格式後，將其中四分之三的文件做爲訓練文件集，四分之一做爲測試文件集。透過 CRF 以訓練文件的 Token 特性，演算並建構自動標註模型，再使用測試文件測試自動標註之效果，並依測試結果調校運算參數或調整會議資訊特徵人工標註規則，以提升自動標註模型的績效。CRF 的實驗結果如表 6 所示，由於希望加強 Recall，以儘可能地擷取相關的 Entities，以避免遺漏會議資訊，因此表 6 顯示 Recall 相對較高。對於可能造成的誤判，再應用許多 Heuristic Rules 過濾不適當或是錯誤的訊息，這些 Heuristic Rules 可分爲下列五種型式：

- 序列規則（Sequence Rule）：考量時間資訊的序列性。
- 詞彙規則（Term Rule）：考量特定的詞彙。
- 位置規則（Location Rule）：考量具名實體的相對位置。
- 格式規則（Format Rule）：考量時間資訊的格式。
- 相似規則（Similarity Rule）：考量具名實體的相似性。

表 6. 具名實體的擷取

Documents / System	True Entities	False Entities
Positive Entities	1632	1079
Negative Entities	261	2785

Recall (R) = 1632/(1632+261)=86.21%; Precision (P) = 1632/(1632+1079)= 60.20%
F1 measure (F1) = (2*P*R)/(P+R)=70.89%

5. 系統實作與功能

爲了實作本研究提出的學術資訊自動擷取的機制，並提供學術會議資訊之應用服務，我們建構學術會議資訊檢索與擷取系統平台（Academic Conference Information Retrieval & Extraction System，簡稱 ACIRES）。ACIRES 由後端資訊處理系統與前端使用者系統構成，兩者皆爲自動化與即時性之服務，系統架構如圖 4 所示。後端系統蒐集網路上的學術會議資訊網頁、過濾非相關網頁、擷取會議資訊、並進而建立文件索引，前端系統是與使用者互動的入口，使用後端系統建構之索引資料，提供使用者各項服務，並與 Google Calendar 聯繫，建構個人行事曆。以下分別介紹後端資訊處理系統以及前端使用者系統的各項功能。

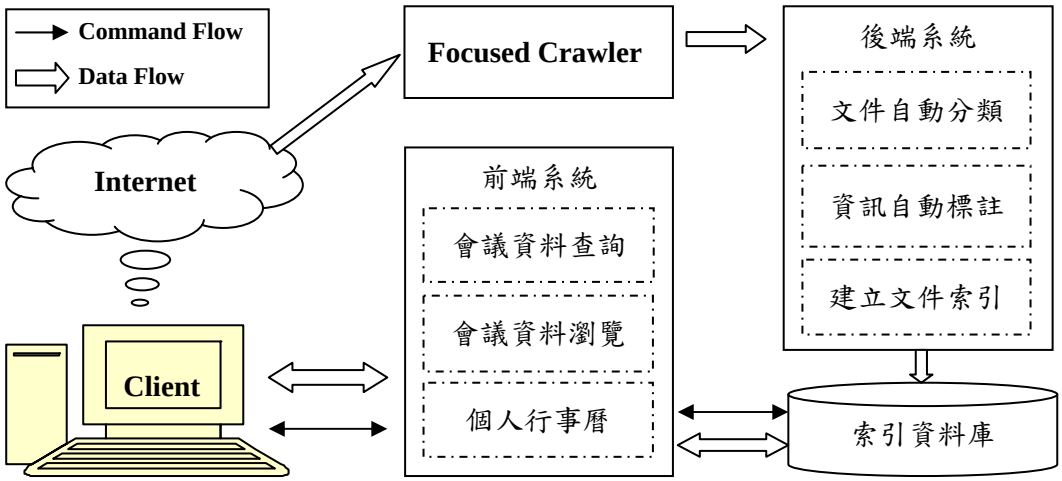

圖 4. ACIRES 整體系統架構

5.1 後端資訊處理系統

後端資訊處理系統主要的工作爲文件自動分類、資訊自動標註、以及建立文件索引，請參考圖5。後端系統使用Google Alert蒐集網路上可能的學術會議資訊、過濾無關的內容、擷取會議各項時間與地點資訊、建置文件索引資料，分別說明如下。

5.1.1 文件自動分類

ACIRES 持續以 Google Alert 快訊服務，以本研究整理的學科主題關鍵字，訂閱各主題相關網頁通知，取得最新的學術會議資訊，保持資料的即時性與時效性。由 Google Alert 蒐集而得的網頁，先經由 Rainbow Classifier 的文件分類模型，自動過濾非相關網頁。再經過去除雜訊的程序，刪除廣告，動態網頁程式等與會議資訊無關的內容。

5.1.2 資訊自動標註

已去除雜訊的網頁，進一步轉製成特定格式，以本研究建置的 CRF 資訊擷取模型，自動標註網頁中的會議資訊特徵。系統解析完成標註的文件，一一擷取各項特徵項目，再針對不同資料格式進一步處理，例如統一日期格式、轉換 HTML 特殊字元等。另外，有些網頁可能包含一個以上的學術會議資訊，因此同一份文件所擷取的項目會有重覆出現的狀況，例如有兩個會議時間、有三個會議地點等。系統則依文件排版的先後順序關係，將特徵項目分組爲多筆會議資料。

5.1.3 建立文件索引

透過自動資訊擷取所取得的各項會議資訊，以及研討會通知網頁中未被擷取的其他相關資訊，都需進一步整合爲容易查找的資料集合，以提供快速且簡便的檢索及瀏覽服務。

ACIRES 採用 Lucene 檢索系統整合所蒐集與整理的會議資料。（Apache Software Foundation, 2010）Lucene 爲完整的資訊檢索系統，提供全文資料及欄位資料的索引建立與資料查詢功能。ACIRES 取用已去除雜訊的網頁內容建立全文索引。每一筆會議資料是由一份網頁全文及多個自動擷取的特徵項目所組成，這些特徵項目也是建立索引資料庫時，各學術會議資料的欄位索引項目。

5.2 前端使用者系統

如前文所述，前端系統乃是支援使用者各項功能的入口，其架構如圖 6 所示，各項功能可分爲兩大模組：1) 會議資料搜尋；2) 個人行事曆。會議資料搜尋爲了滿足使用者檢視資料的不同需求，實際提供了包括基本檢索、進階檢索、分類瀏覽、時間瀏覽、地點瀏覽等功能；個人行事曆則是提供行事曆的管理功能。圖 7 爲前端使用者系統的入口首頁，分爲時間資訊畫面、檢索功能畫面、分類瀏覽畫面、檢索結果畫面，下文簡要說明各項功能。

5.2.1 查詢學術會議資訊

系統提供基本的全文檢索功能，以及可指定欄位的進階檢索功能。當使用者進行關鍵字檢索時，系統查找研討會通告中含有查詢關鍵字的文件，依序列出查詢結果。使用者亦可進一步利用不同欄位間的布林邏輯進行進階檢索，查找更精確的會議資料。使用者點選進階檢索的鏈結，系統展現進階檢索的功能畫面，使用者可使用"AND"、"OR"、"NOT"組合不同欄位，進階檢索提供的檢索欄位，包含所有會議資訊特徵項目，請參見圖 8。

5.2.2 檢視詳細會議資訊

查詢結果清單的每筆會議資訊包含會議名稱、會議日期、會議地點以及查詢關鍵字在文件中出現的片段。使用者可點選每筆會議資訊的[Detail]按鈕，檢視更詳細的資料。[Detail]視窗分爲二部分，上方是本系統摘錄的會議基本訊息，下方式系統儲存的會議通告文件，使用者也可以進一步在詳細資料視窗點選原始網頁位址，進入該學術會議官方網站取得進一步資訊，請參見圖 9。

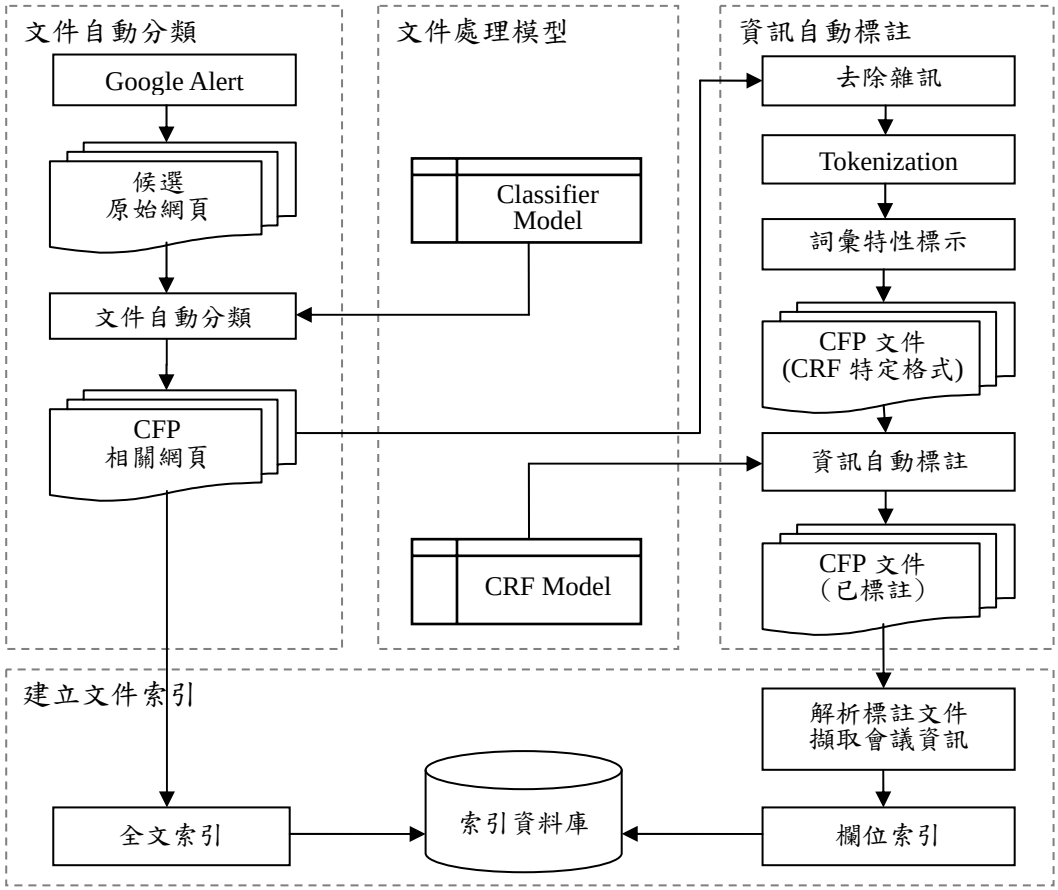

圖 5. ACIRES 系統架構：後端資訊處理系統

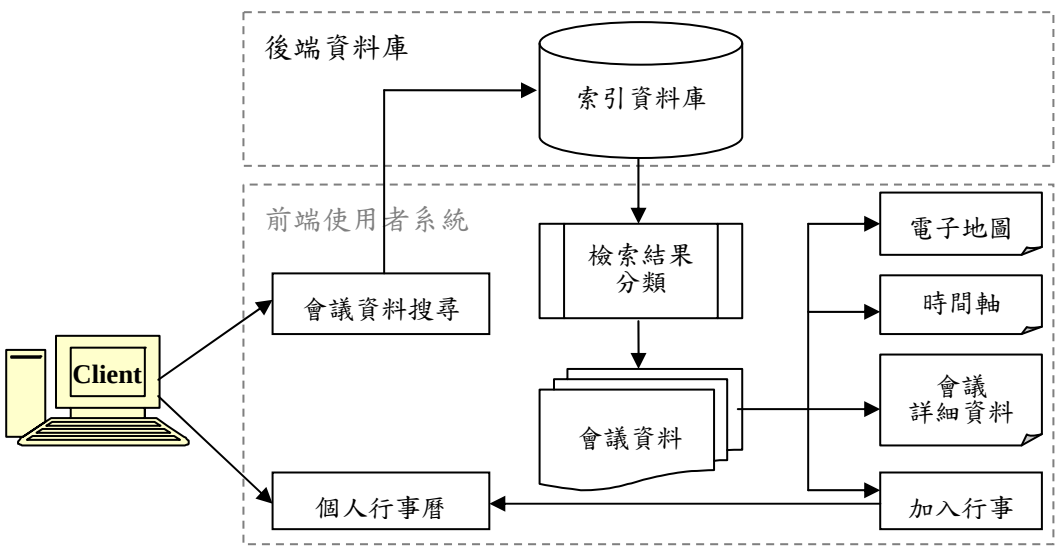

圖 6. ACIRES 系統架構：前端使用者系統

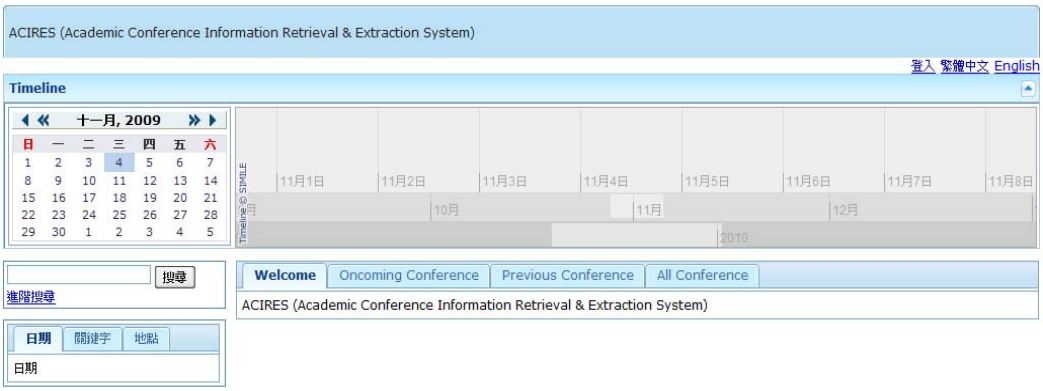

圖 7. ACIRES 首頁

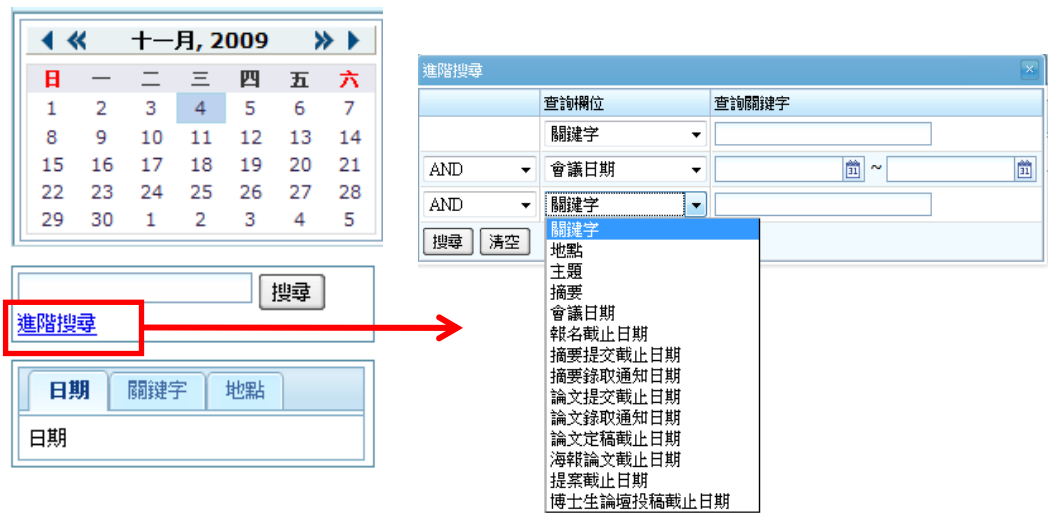

圖 8. 進階檢索

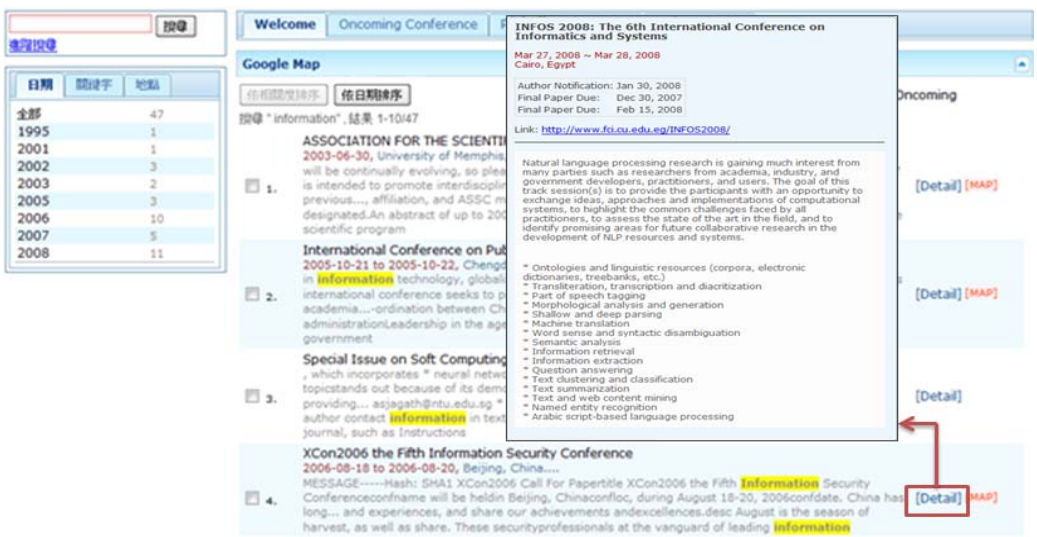

圖 9. 查詢結果清單及會議詳細資料

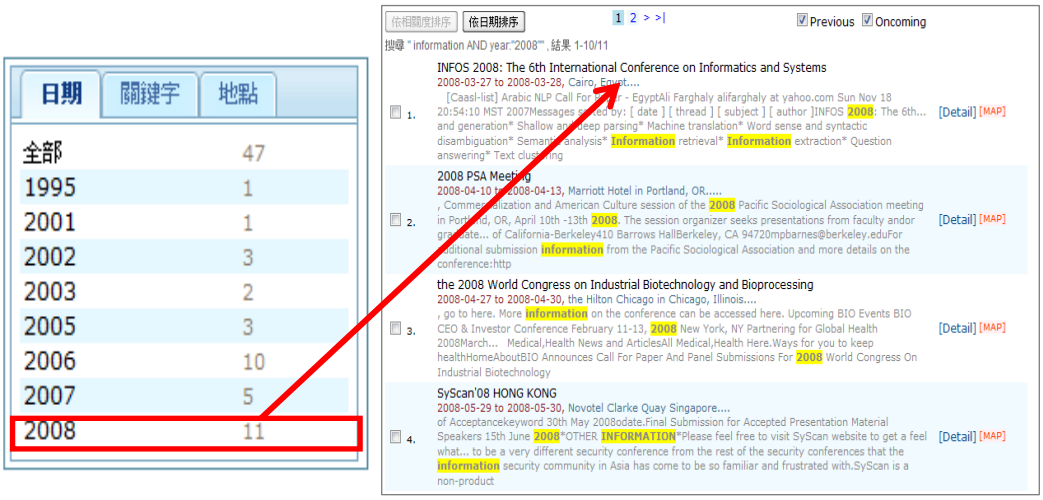

圖 10. 查詢結果分類瀏覽

5.2.3 分類瀏覽查詢結果

過去的檢索系統通常僅僅顯示檢索的結果，本系統則是進一步允許使用者依據會議舉行年份、會議舉行地點、及會議相關議題等不同觀點，更有意義地瀏覽結果。因此，本系統會爲每一次的查詢結果，進行自動分群的工作，讓使用者可進一步縮小檢索範圍，分類瀏覽查詢結果，請參見圖 10。

5.2.4 檢視會議時間資訊

以時序方式展示研討會事件，可以讓使用者更容易安排學術活動，本系統用時間軸移動的概念，表示每個研討會事件的先後順序，讓使用者可以清楚地了解不同時間中的會議舉行狀態。使用者可直接捲動時間軸改變呈現的時間點，或是利用左上角的日曆設定日期，時間軸會即時連動至對應的時間點。當使用者勾選查詢結果清單上任一筆會議資訊，時間軸也會自動捲至該會議舉行的時間點，讓使用者可於時間軸上檢視在同一時間舉行的其他會議。點選時間軸上的事件節點，則可檢視該會議詳細資訊，請參見圖 11、12、13。

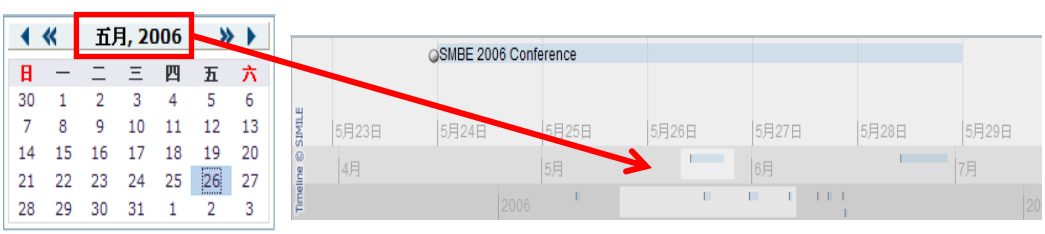

圖 11. 點選日曆捲動時間軸

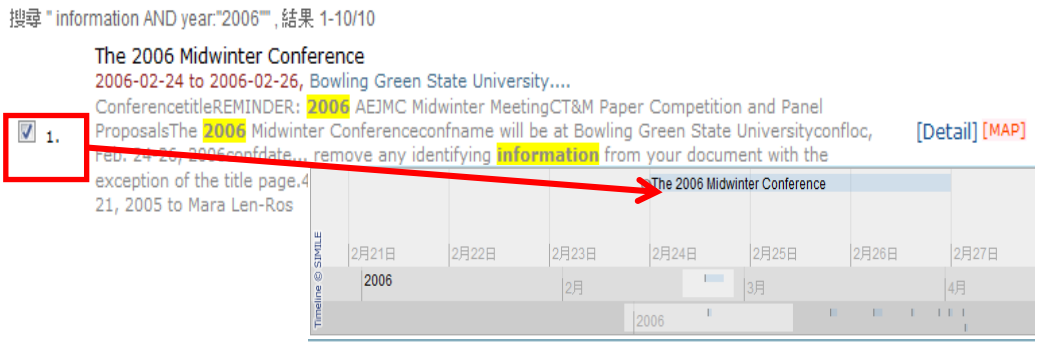

圖 12. 勾選會議項目自動捲動至對應時間

圖 13. 於時間軸上檢視會議資訊

5.2.5 瀏覽會議地點資訊

本系統整合 Google Map 服務，將查詢結果所得之會議舉行地點一一標示在地圖上，請參見圖 14。直接點選任一地點標示，即可檢視對應會議的相關資訊。當使用者勾選任一項會議資料，或點選檢索結果清單的[MAP]按鈕，地圖即自動將該地點放大特寫，請參見圖 15。

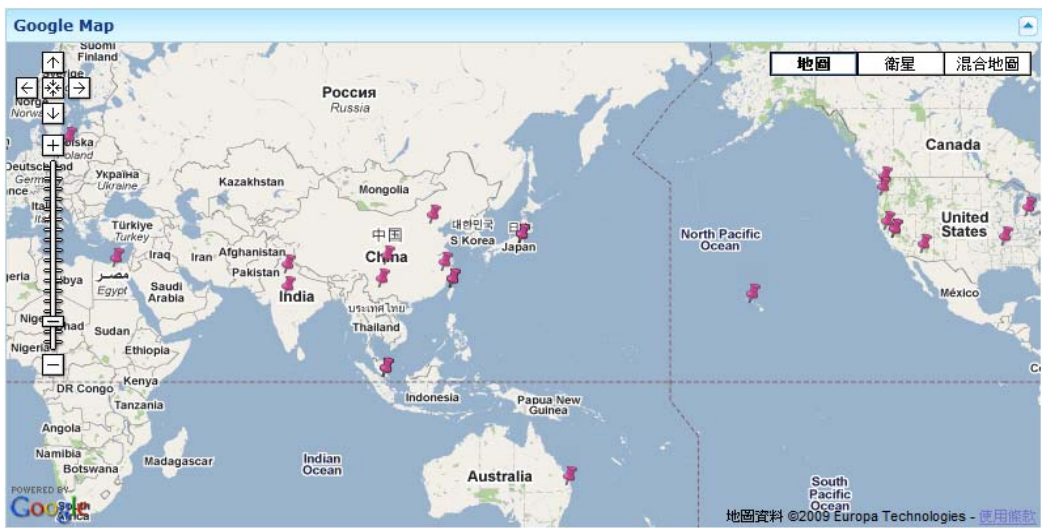

圖 14. Google Map -- Global View

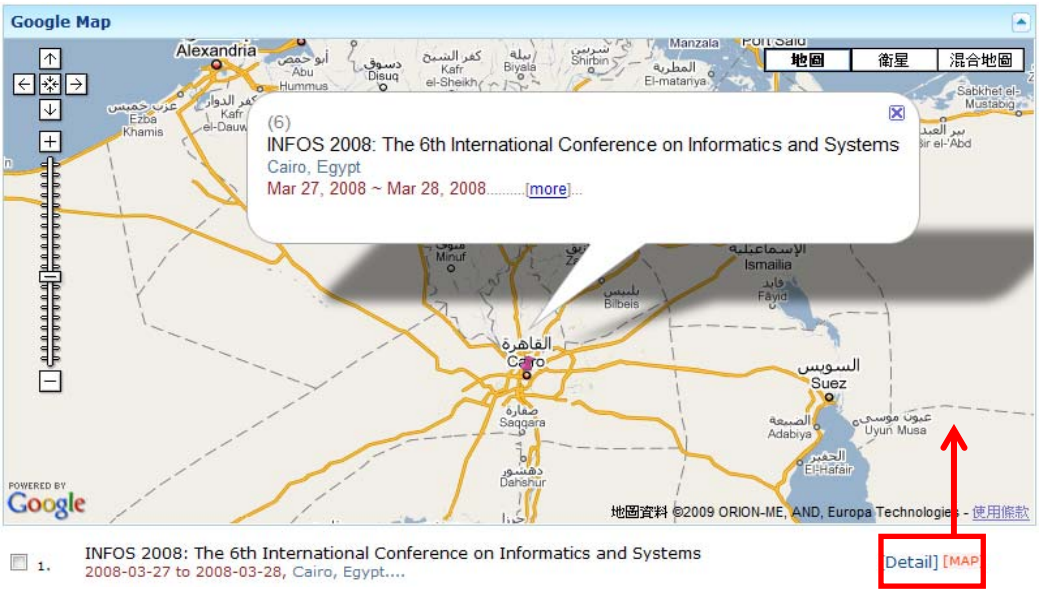

圖 15. Google Map -- Single Spot View

5.2.6 個人行事曆

本系統進一步讓使用者儲存並記錄想繼續追蹤的學術會議資訊。由於 Google Calendar 的使用者眾多，也已經有許多應用程式可執行於多種不同的智慧型資訊裝置，使用者可以很方便地使用各種裝置查詢行事曆。因此本研究整合 Google Calendar 個人行事曆服務，讓使用者將學術會議加入行事曆，並可直接在 ACIRES 檢視個人的行事曆內容。使用者於檢索結果清單中勾選感興趣的會議項目後，所勾選項目即加入左方的書籤清單，亦可點選[x]按鈕刪除對應的會議項目，書籤清單可記錄多次檢索結果所勾選的項目，使用者可隨時新增或刪除選取的項目，請參見圖 16。

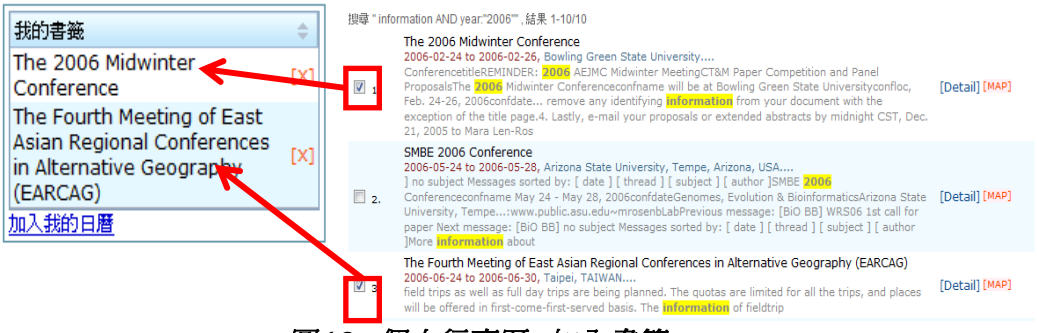

圖 16. 個人行事曆--加入書籤

會議資訊在放在「我的書籤」的清單後，尚未真正進入 Google Calendar，此時使用者可以預覽每筆會議項目資料，並可以編輯修改寫入行事曆的相關說明，請參見圖 17，當一切就緒後，可以點選「加入我的日曆」，相關會議資訊才是真正地寫入 Google Calendar。

當使用者使用 ACIRES 系統時，若已經使用 Google 帳號登入，即可在首頁直接檢視個人的行事曆內容，請參見圖 18 與圖 19。

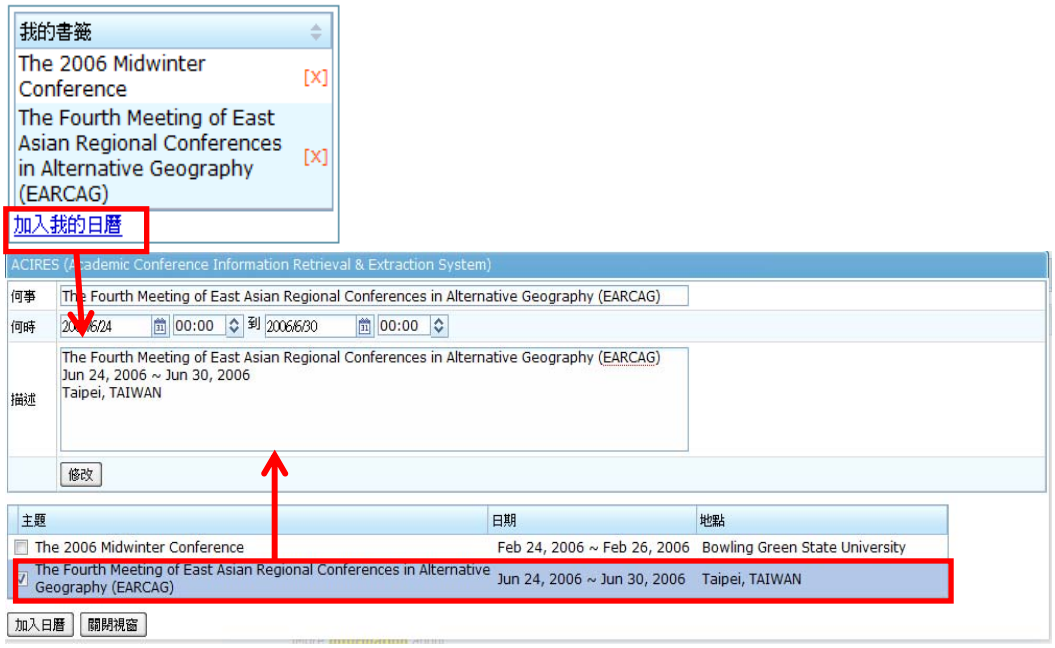

圖 17. 個人行事曆--加入行事曆前預覽

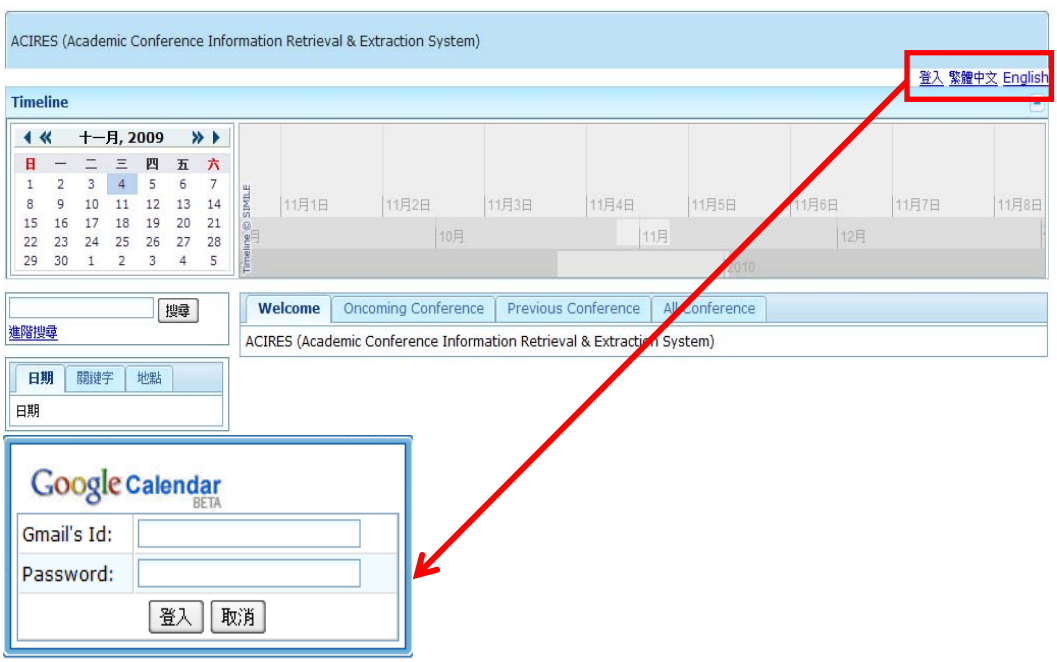

圖 18. 以 Google 帳號登入

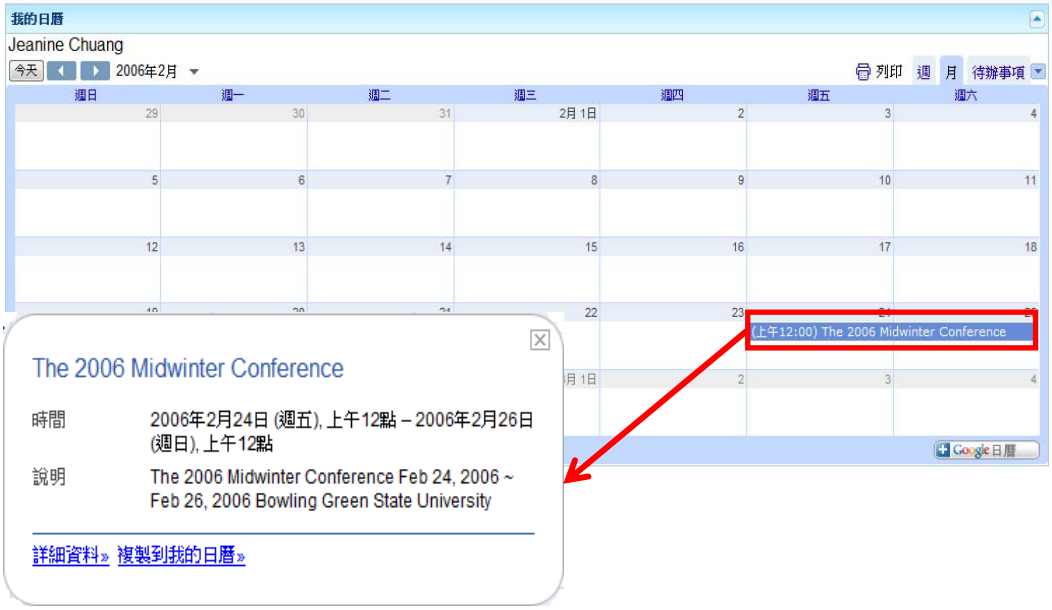

圖 19. 個人行事曆-總覽

6. 結論

本文提出學術會議資訊檢索的自動處理程序，以因應廣泛的學術會議資訊檢索需求。爲了確認處理程序的可行性，本研究首先進行了一系列分類績效與擷取績效的實驗，實驗顯示分類績效 F1 measure 超過 80%；擷取績效 Recall 超過 86%，F1 measure 超過 70%。第二步則是實作學術會議資訊檢索與擷取系統平台。本研究不僅提供傳統的主題檢索功能，考慮使用者的時間考量與空間考量，允許使用者由時間軸線與空間地圖瀏覽學術會議資訊，並應用檢索後分類的策略，讓使用者可以分類瀏覽檢索結果，更有意義地看待學術會議資訊。本研究同時整合了個人行事曆功能，讓學術會議資訊檢索融入研究人員的活動行程，使得前述的系統平台更具實用性。

一個相對完整的資訊系統，除了提供檢索功能外，還應該提供系統內的知識框架，允許使用者應用瀏覽的方式，檢視系統提供的各項資訊，未來本研究提出的學術會議資訊檢索與擷取系統平台將加入這樣的知識框架，以及使用者個人化的功能，例如個人專題資訊選萃（資訊過濾）。

誌謝

本研究部分研究成果獲得國科會專題研究計畫「個人化資訊服務：學術資訊之擷取」的補助，計畫編號爲 NSC 96-2413-H-002-018。感謝研究助理莊雅蓁小姐、陳新瑋先生、與陳瑞呈先生的協助。

參考文獻

Apache Software Foundation. (2010). *Apache Lucene - Overview*. Retrieved Oct. 1, 2010, from http://lucene.apache.org/java/docs/index.pdf

ARWU (2010). *Academic Ranking of World Universities - 2010*. Retrieved Oct. 1, 2010, from http://www.arwu.org/

Brennhaug, K. E. (2005). *EventSeer: Testing Different Approaches to Topical Crawling for Call for Paper Announcements*. Unpublished Thesis. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway. Retrieved Oct. 1, 2010, from http://ntnu.diva-portal.org/smash/get/diva2:348108/FULLTEXT01

Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In *Proceedings of the 8th International World Wide Web Conference* (Vol. 31, pp. 1623-1640). Retrieved Oct. 1, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.1111&rep=rep1&type=pdf

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A brief History. In *Proceedings of the 16th International Conference on Computational Linguistics* (pp. 466-471). Retrieved Oct. 1, 2010 from http://www.aclweb.org/anthology/C/C96/C96-1079.pdf

Kudo, T. (2010). *CRF++: Yet Another CRF Toolkit Version 0.54*. Retrieved Jun. 2, 2010 from http://crfpp.sourceforge.net/

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 282-289). Retrieved Oct. 1, 2010, from http://www.cis.upenn.edu/~pereira/papers/crf.pdf

Lazarinis, F. (1998). Combining Information Retrieval with Information Extraction for Efficient Retrieval of Calls for Papers. In *Proceedings of IRSG98*. Retrieved Oct. 1, 2010, from http://www.cs.strath.ac.uk/~mdd/research/publications/98lazarinis.pdf

McCallum, A. (1996). *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering*. Retrieved Aug. 4, 2009, from http://www.cs.cmu.edu/~mccallum/bow.

MUC (2001). *Message Understanding Conference Evaluation*. Retrieved Oct. 1, 2010 from http://www-nlpir.nist.gov/related_projects/muc/

Pink, B., & Bascand, G. (2008). *Australian and New Zealand Standard Research Classification (ANZSRC)*. Retrieved Mar. 2, 2010, from http://www.arc.gov.au/pdf/ANZSRC_FOR_codes.pdf

QS (2010). *World University Rankings*. Retrieved Oct. 1, 2010, from http://www.thes.co.uk/worldrankings/

Schneider, K.-M. (2005). An Evaluation of Layout Features for Information Extraction from Calls for Papers. In *Proceedings of Lernen, Wissensentdeckung und Adaptivitat* (pp.

111-116). Retrieved Oct. 1, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.118&rep=rep1&type=pdf

Sutton, C., Rohanimanesh, K., & McCallum, A. (2004). Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. In *Proceedings of the 21st International Conference on Machine Learning*. Retrieved Oct. 1, 2010, from http://www.cs.umass.edu/~mccallum/papers/dcrf-icml04.pdf

Takada, T (2008). ConfShare: A Unified Conference Calendar that Assists Researchers in the Tasks for Attending an Academic Conference. *Journal of Information Processing Society of Japan*, 49(12), 4093-4104.

Wallach, H. M. (2004). *Conditional Random Fields: An Introduction*. Technical Report MS-CIS-04-21, University of Pennsylvania. Retrieved Oct. 1, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.436&rep=rep1&type=pdf

Xin, X., Li, J., Tang, J., & Kuo, Q. (2008). Academic Conference Homepage Understanding using Constrained Hierarchical Conditional Random Fields. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1301-1310). Retrieved Oct. 1, 2010, from http://doi.acm.org/10.1145/1458082.1458254