跨語系之學術搜尋引擎的開發及挑戰

李竺芸
資訊工程學系
國立中央大學
桃園，台灣
totoro2345678@gmail.com

林冠佑
資訊工程學系
國立中央大學
桃園，台灣
angus850211@gmail.com

陳弘軒
資訊工程學系
國立中央大學
桃園，台灣
hhchen@g.ncu.edu.tw

張嘉惠
資訊工程學系
國立中央大學
桃園，台灣
chiahui@g.ncu.edu.tw

***Abstract*—我們以人工智慧研討會 (Conference on Technologies and Applications of Artificial Intelligence, 簡稱 TAAI) 2015 年至 2019 年間的國際論文及本地論文的元資料 (metadata) 及論文的 PDF 檔案建構了一個跨語系 (中文及英文) 的學術論文搜尋引擎。本文敘述此系統的建構過程、過去五年的論文統計資訊與主題方向、及未來展望。我們公開此系統的部份原始碼，其他研討會 (尤其是未被數位文獻資料庫收錄的區域性或非英文的研討會) 可藉由我們公開的原始碼快速建構研討會專屬的學術搜尋引擎。**

1. 導論

研究者們通常利用學術論文闡述研究成果。為了讓論文便於取得及搜尋，出版商及相關產業推出各種數位資訊平台，這些資訊平台有些提供論文的全文搜索或摘要搜索功能，有些提供完整的參考及引用 (reference and citation) 資訊。這些工具或平台除了讓資訊的取得更便捷，也讓學術發表的資訊得以量化。例如：我們可以量化一篇論文的「被引用次數」、或計算一個學者、一個機構、甚至一個國家的「論文數量」或「高引用論文數量」等。誠然，這些量化數字都過度簡化甚至膚淺化了學術研究的本質，但至少它們提供了某種角度的衡量方式 (即使是種狹隘的角度)。總而言之，數位化的論文及收錄方式大幅降低了論文整理、論文搜尋所需耗費的人力成本，並得以應用資料科學的技術對大量的科學文獻進行分析。

然而，區域型學術會議所發表的論文 (特別是其中以英文以外的語言所撰寫的學術論文) 通常不被正式收錄在專門的學術搜尋引擎平台中。即使是鼓勵上傳並開放預印本論文 (preprint) 且不需同儕審查 (peer review) 的 arXiv 平台，[1]也要求非英文的稿件必需要有英文的摘要。[2]在各種限制下，區域型學術會議上所發表的非英文論文常淪為學術上的孤兒：發表後即被遺忘，甚至消失在網路上。

台灣的人工智慧研討會 (Conference on Technologies and Applications of Artificial Intelligence, 簡稱 TAAI) 自 1995 年起除 1997 年外每年舉辦，至 2020 年為第 25 屆的年會。此研討會的主會議分為國際軌 (International Track) 及本地軌 (Domestic Track)，自 2010 年起 (除 2014 年外) 國際軌的論文均收錄至 IEEE Xplore 學術論文資料庫，但本地軌的論文則大多隨著會議的結束而佚失。為了更好地保留及整理歷年的研究成果，我們建立了一個學術搜尋引擎[3]收錄 2015 至 2019 年間被接受的國際軌及本地軌論文。另外，由於 2015 年及 2019 年的部份特別議程 (special session) 與主會議議程採用相同的投稿平台，故這兩年的部份特別議程也被納入此學術搜尋引擎中。本學術搜尋系統的資料匯入與索引流程接近全自動化，本年度 (2020) 及未來年度的被接受論文預計也將繼續收錄至此平台上。另外，我們建議未來的特別議程與主會議議程使用同樣的投稿平台讓特別議程的論文也能被索引。

我們將部份的核心元件開源至 GitHub[4] 並採用 MIT 授權，其他學會或區域性的研討會若有類似的需求將可快速複製我們的成功經驗。

本論文將敘述此學術搜尋引擎的建構過程 (Section 3)、設計上的考量 (Section 2、Section 3)、並回報 2015 年至 2019 年的論文的總體統計資訊及論文主題方向 (Section 4)；同時，我們也將討論跨語系的學術搜尋引擎可能面臨的獨特挑戰，以及我們計劃採用的策略 (Section 5)。

2. 相關研究

本節介紹幾個資訊領域較著名的學術論文搜尋引擎或資料庫。

資訊領域最重要的兩個學術組織可能是 ACM 與 IEEE，兩者也都有相應的論文搜尋系統：ACM Digital Library[5] 及 IEEE Xplore[6]。由於這兩個組織通常可向學術會議的主辦單位或期刊的編輯直接取得元資料，故其相應的論文資料庫的元資料相當精確。然而，這兩個系統只允許該組織的會員或團體會員方能閱讀全文。此外，雖然我們不確定 ACM 或 IEEE 是否明文規定只收錄英文論文，但我們的確從未看過 ACM 或 IEEE 的論文使用英文以外的語言撰寫。因此，要讓本地軌的論文集收錄在 ACM 或 IEEE 的論文系統中可能比較困難。

某些學術搜尋引擎採用網路爬蟲來蒐集科學文獻，並利用自然語言處理、資訊擷取、資料檢索、及其他人工智慧的相關技術來獲得論文的元資料, 代表的網站包括 CiteSeerX [1]、Google Scholar [2], Microsoft Academic [3] 等。這類型的平台海納百川，只要是網路

[1]https://arxiv.org/
[2]https://arxiv.org/help/faq/multilang
[3]http://search.taai.org.tw/
[4]https://github.com/taai-taiwan/academic-search
[5]https://dl.acm.org/
[6]https://ieeexplore.ieee.org/

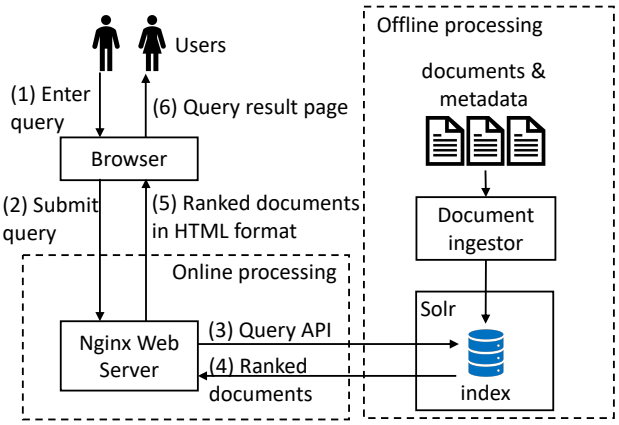

圖 1. TAAI 學術搜尋引擎系統架構圖

上的科學文獻都有可能被這類平台自動收錄並索引。然而，由於科學文獻的格式並不統一，這類型網站所得到的論文元資料 (如：作者、會議或期刊名等) 可能比較雜亂，同時，這類平台對於非英文的科學文獻也處理得比較差。其中，CiteSeerX 公開其原始碼，[7]其技術架構可做為我們的參考。

另一個值得參考的網站是 DBLP [4]。[8]。這個網站收錄了超過 500 萬篇電腦科學相關論文的標題、作者、會議或期刊等資訊。DBLP 從出版商取得元資料後，經過嚴格的人工編輯及清理，故 DBLP 的元資料比較可信。然而，DBLP 的做法需要大量的人力，且 DBLP 並不提供論文的全文搜尋或下載服務，故與我們的目標仍有出入。

對於我們的目標 — 建構 TAAI 會議的論文搜尋平台，這些著名的學術搜尋平台的做法各有其優缺點。我們最後採用綜合的作法：仿 DBLP 的型式直接從各屆的 TAAI 主辦方取得論文的元資料，但搜尋平台採用類似 CiteSeerX 的架構進行全文檢索，且允許全文下載本地軌論文或連至 IEEE Xplore 下載國際軌論文。另外，此平台上中文和英文論文各半，將帶來獨特的挑戰。

3. 系統總覽

本節介紹此學術搜尋引擎的架構，包括線下處理 (offline processing) 的資料預處理與資料索引，及線上處理 (online processing) 則將使用者的查詢送至索引並回傳結果，流程如圖 1 所示。以下詳述線下處理及線上處理的過程。

A. Offline Processing

我們聯繫了過去五年 (2015 - 2019) 的會議主辦人，從投稿平台 (EasyChair[9] 或 Microsoft CMT[10]) 下載各年度被接受論文的稿件 PDF 檔及元資料，由於每年用的投稿平台不同，元資料的格式也不同，因此剛開始得到的元資料比較凌亂且沒有統一的格式，必須要人工整理元資料，才能將資料匯入。然而，為了讓平台的自動化程度更高，我們撰寫了程式將 Microsoft CMT 投稿平台上提供的元資料從 xlsx (Microsoft Excel) 格式轉成本學術搜尋引擎後台所需要的格式。今年 (2020) 的 TAAI 會議也因此使用 Microsoft CMT 做為投稿平台，以利加入新的論文。

在論文全文檢索的部份，我們進行以下的線下處理。首先，學術論文的投稿格式為 PDF 檔，但 PDF 格式對於程式來說較不好處理，因此我們將 PDF 檔一律先轉為純文字檔，過程中也針對論文內文過濾了停用字及標點符號，以便於後續的處理。最後，我們將論文全文連同此論文的重要資料欄位 (包括：Paper Title、Abstract、Author Names、Track Name) 一併匯入 Apache Solr 建立索引。

學術搜尋引擎另外有一個一般搜尋引擎較少研究的主題：作者消歧義 (author disambiguation) [5], [6]。作者可能發生歧義的狀況有兩者：類型一，不同的人可能具備同樣的姓名 (e.g., 於撰寫本文時，DBLP 認為署名為 “Wei Zhang” 的論文來自於 149 個英文姓名相同的人)，因此，系統可能將多個同名的學者誤認為是同一位學者；類型二，同一個人有時候會有不同的姓名呈現方式 (e.g., 根據 DBLP，Clyde Lee Giles 與 C. Lee Giles 是同一個人)，單純採用字串比對的方式可能將同一作者的多種姓名表達錯認為是多位學者。目前本學術搜尋引擎僅以字串匹配的方式分辨不同作者，因此可能發生上述的作者歧義問題，最明顯的例子是：同一個作者發表在本地軌及國際軌的論文可能因為「中文」姓名及「英文」姓名的差異而被誤認為兩個不同的作者。然而，我們在系統設計上已經為這個問題預留了彈性：我們為每一篇論文的每一個作者建立鍵 (key) 和值 (value)，針對上述類型一的歧義，相同姓名的兩個人只要給予不同的鍵編號，則會被系統認為是不同的人；對於類型二的歧義，單一作者的不同姓名寫法給予相同的鍵編號，則系統會認為他們是同一個人。一旦具備好的消歧義方式，系統可直接支援消歧義後的結果。

B. Online Processing

我們向 DigitalOcean[11]租用虛擬實體雲端主機 (VPS)，採用 Nginx 非同步框架的網頁伺服器，後台搜尋引擎採用 Apache Solr，並對 Solr 後台加裝 Basic Authentication Plugin，使得訪問後台、更新資料都需要帳號密碼驗證。此外我們開發 API 連接網頁前後端，用 Asynchronous JavaScript and XML (AJAX) 的方式發出 HTTP POST 請求，得到使用者搜尋的資料，並利用跨來源資源共享 (Cross-Origin Resource Sharing)，限定只有此學術搜尋引擎的域名，才能存取資料。

首頁的部分顯示前十篇瀏覽次數最高的論文，每篇的論文資訊列表有論文名稱、年份、會議名稱、作者、論文瀏覽次數、論文摘要片段。在統計論文的瀏覽量用了兩種方式統計：Solr 後台統計數據及 Google Analytics 的統計資訊，此學術搜尋引擎顯示的瀏覽次數是 Solr

[7]https://github.com/SeerLabs/CiteSeerX
[8]https://dblp.uni-trier.de/
[9]https://easychair.org/
[10]https://cmt3.research.microsoft.com/
[11]https://www.digitalocean.com/

後台統計儲存的數據，在 Solr 後台更新瀏覽量的方式是利用 Atomic update[12]更新點擊量，此方式更新數據時，不必重新索引整個文檔，因此大幅減少處理時間。Google Analytics 的統計數據主要僅用來驗證 Solr 的統計值，Google Analytics 是利用點擊的方式觸發事件統計瀏覽量，Google Analytics 的好處是高度可視覺化的統計報表，後台並有即時數據顯示 30 分鐘內被點擊的論文，然而總數據的更新時間是 24-48 小時才能查看。

此學術搜尋引擎雖支援全文檢索，但礙於版權問題，我們只開放國內軌的論文全文下載，國際軌的論文則直接指向 IEEE Xplore，故 IEEE 的會員或有訂閱 IEEE Xplore 的機關團體依然可下載全文。

4. TAAI 歷年發表趨勢 (2015 - 2019)

表 I
2015 - 2019 每年主會議議程的國際軌及本地軌論文數及特別議程論文數統計 (特別議程僅考慮收錄至本系統的論文)

年度	2015	2016	2017	2018	2019
國際軌	40	40	36	38	54
本地軌	39	31	33	42	32
特別議程	42	-	-	-	87

表 II
2015 - 2019 出現次數最高的 20 個關鍵字詞

排名	Keyphrase	出現次數
1	neural network	916
2	machine learning	576
3	time series	348
4	deep learning	306
5	data mining	265
6	artificial intelligence	232
7	social network	225
8	natural language	212
9	computer vision	175
10	sentiment analysis	170
11	domain adaptation	168
12	convolutional neural network	155
13	reinforcement learning	152
14	big data	139
15	evaluation function	138
16	pattern recognition	137
17	decision tree	132
18	natural language processing	122
19	random forest	118
20	loss function	114

本節展示針對 2015 年至 2019 年 TAAI 會議上的論文初步的分析成果。

表 I 展現過去五年國際軌與本地軌接受的論文數量。整體而言，在過去五年主議程的論文變化量不大，但 2019 年接受了較大量的特別議程論文。

我們另外針對這五年的論文主題進行簡單的分析。由於中文涉及斷字等處理，我們目前只對論文中的英文用詞進行分析。過去的研究曾顯示文章中的關鍵字通常是名詞 [7]，故我們首先嘗試利用英文的詞性解析器找出

[12]https://lucene.apache.org/solr/guide/8_6/updating-parts-of-documents.html

表 III
2015 - 2019 出現在最多檔案的 20 個關鍵字詞

排名	Keyphrase	出現的檔案數
1	neural network	83
2	machine learning	51
3	deep learning	49
4	convolutional neural network	39
5	computer vision	38
6	pattern recognition	35
7	artificial intelligence	33
8	loss function	21
9	learning rate	20
10	big data	19
11	data mining	18
12	natural language	18
13	object detection	17
14	recurrent neural network	17
15	natural language processing	16
16	computational intelligence	15
17	transfer learning	15
18	network model	14
19	mean square	14
20	signal processing	14

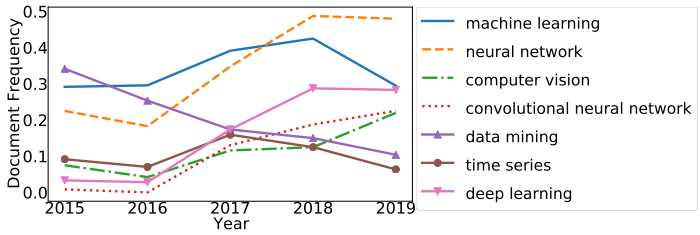

圖 2. 七個熱門字詞歷年的檔案頻率變化趨勢

名詞，並在刪除停用字 (stop words) 後，將剩餘的名詞及名詞片語當作文章可能的關鍵字詞。然而，這個方法所產生的字詞仍有相當的部份屬於意義比較模糊的常見字詞 (例如：“experimental result”, “previous study”)。由於 TAAI 論文主題聚焦於電腦科學當中的人工智慧技術及應用，我們最後決定仿 [8] 的做法：以維基百科做為關鍵字的參照字典。具體而言，我們爬取維基百科上電腦科學領域類別下所有文章的標題，建立了專屬於電腦科學的字典，再用公共最長子字串 (longest common substring) 的匹配的方式匹配出 TAAI 的論文裡曾出現在字典中的字詞。我們人工檢視了部份的論文後，發現這個方法所找到的關鍵字詞能相當程度地表示論文的關鍵技術，以下的論文主題分析均以這個方法所產生的關鍵字詞為基礎所展開。

表 II 及表 III 展示這五年間出現過最多次的關鍵字詞及出現在最多論文中的關鍵字詞，從這兩張表可以看出這五年中 TAAI 論文有相當大的比例集中在機器學習 (尤其是深度學習) 相關的技術。就應用領域而言，則大多集中在社群網路分析、自然語言處理、及電腦視覺。

我們選出表 II 前五名及表 III 前五名的聯集，得到以下七個字詞：neural network, machine learning, time series, deep learning, data mining, convolutional neural network, computer vision。為了比較這七個字詞在過去

五年的消長，我們計算了每個詞 w 在每一年度 t 的檔案頻率 (document frequency) $\mathrm{DF}_t(w)$，如式 1 所示。

$$\mathrm{DF}_t(w) = \frac{n_t(w)}{N_t}, \tag{1}$$

其中 $n_t(w)$ 是 w 在年度 t 出現在多少篇論文中，N_t 是年度 t 總共接受的論文數。

圖 2 展示這七個詞彙歷年的檔案頻率變化趨勢圖。其中，有較明顯增長的詞彙包括："neural network", "deep learning", "convolutional neural network", 和 "computer vision"；有較明顯的下降趨勢的是 "data mining"；而 "machine learning" 和 "time series" 則大致持平。

5. 結論與未來展望

本節討論此學術搜尋引擎的已知缺陷、可能的研究方向，以及我們計劃加強的部份或增加的功能。

首先，目前的設計相當依賴元資料的正確性，但元資料的訊息是由論文的上傳者提供的，雖然大部份的上傳者不會刻意登入錯誤的元資料，但人為疏失在所難免。我們目前發現一個比較常見的問題是上傳者在作者欄僅列出自己的姓名，其他作者的資訊雖然出現在 PDF 檔案中但未紀錄在元資料中，因此系統在搜尋結果的頁面只能顯示不完整的資訊；基於相同的原因，使用「作者搜尋」時系統可能會漏掉一部份的作者。

其次，我們希望能夠更多元地量化國內人工智慧相關研究的能量。Section 4 呈現了初步的結果，但仍有許多可繼續的課題，例如：使用其他主題模型 (如：Topic over time [9]) 來量化研究趨勢。另外，我們也計劃讓使用者能更方便地取得這些量化的結果：目前若要產生如圖 2 的字詞趨勢圖需要撰寫程式從後台撈資料，故僅限於有伺服器訪問權限的人才能繪製結果。我們考慮提供使用者界面讓用戶能直接搜索各關鍵字詞的歷年趨勢。

目前搜尋結果的排序只依賴字詞在各欄位 (如：標題、內文) 出現的頻率及檔案頻率做計算，並讓標題的權重略高於內文的權重。然而，目前的權重大小未經較嚴格的實驗，因此，我們計劃實驗不同的權重組合。另外，目前的預設排序也沒有考慮每篇論文的重要性或品質等指標 [1]，我們考慮未來以下面三種方式做為論文品質的參考指標：其一，論文被下載次數。其二，審稿者評分，TAAI 的每篇論文通常會被三位以上的審稿者評分，可參考評分來評價論文品質。其三，論文被引用數，但由於目前我們僅能取得 TAAI 的論文全文，故可能只能統計論文被其他 TAAI 論文所引用的次數。

作者消歧義方面，目前主要問題是同作者的中英文姓名對應。由於本會議的投稿者大部份是台灣人工智慧領域的研究者，故我們考慮人工建立此領域學者的中英文姓名對照表，一個包含數百人的中英姓名對照表應可解決大部份的作者歧義問題。另一種可能的方式是利用電子郵件來比對：若兩個不同的姓名字串對應到相同的電子郵件，則這兩個姓名應對應到同一個作者。

由於 TAAI 會議接受中英兩種語言的論文，同樣的技術或概念至少有中英文兩種不同的稱呼方式，倘若採用字串比對的搜尋方式，則只能得到其中一種語言的結果。例如：當使用「搜尋引擎」當作關鍵字，將無法找到國際軌中討論 "search engine" 的文章。這方面我們有兩項構想：其一，以外部資料 (如：維基百科的中英文條目對應、國家教育研究院的雙語詞彙對照等) 建立中英文對應字典，搜尋時也將另一種語言同義詞彙一併搜尋並用所有結果共同排序。其二，建立跨語言的字詞嵌入式表示法 (word embedding)，搜尋時，先將搜尋字詞轉為嵌入式表示法，再將此嵌入式表示法鄰近的其他字詞都當作搜尋的關鍵字一併搜尋並將結果共同排序。

最後，我們可藉由實際的平台進行各種研究實驗，有不少資訊檢索的相關任務只用線下的日誌 (log) 分析會產生系統性的偏誤 [10]，實際的平台將讓我們可以採用更公平的線上 A/B 分流測試。我們也歡迎研究者共同合作實驗各種設計在實際系統上的效果。

致謝

我們感謝歷屆 TAAI 主辦單位協助取得論文的元資料及 PDF 全文，特別是李健興教授 (TAAI 2015)、陳宜欣教授與沈之涯教授 (TAAI 2016)、林守德教授及楊得年教授 (TAAI 2017)、朱學亭教授 (TAAI 2018)、洪宗貝教授 (TAAI 2019)。我們感謝中央大學資訊系楊佳誠同學協助整理元資料。我們感謝中華民國人工智慧學會第十三屆理監事對此系統的原型給予的建議及支持。

References

[1] J. Wu, K. M. Williams, H.-H. Chen, M. Khabsa, C. Caragea, S. Tuarob, A. G. Ororbia, D. Jordan, P. Mitra, and C. L. Giles, "CiteseerX: AI in a digital library search engine," *AI Magazine*, vol. 36, no. 3, pp. 35–48, 2015.

[2] A.-W. K. Harzing and R. Van der Wal, "Google scholar as a new source for citation analysis," *Ethics in Science and Environmental Politics*, vol. 8, no. 1, pp. 61–73, 2008.

[3] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of Microsoft Academic Service (MAS) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 243–246.

[4] M. Ley, "DBLP: some lessons learned," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1493–1500, 2009.

[5] P. Treeratpituk and C. L. Giles, "Disambiguating authors in academic publications using random forests," in *Proceedings of the 9th ACM/IEEE Joint Conference on Digital Libraries*, 2009, pp. 39–48.

[6] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles, "CollabSeer: a search engine for collaboration discovery," in *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries*, 2011, pp. 231–240.

[7] S. N. Kim and M.-Y. Kan, "Re-examining automatic keyphrase extraction approaches in scientific articles," in *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, 2009, pp. 9–16.

[8] H.-H. Chen, J. Wu, and C. L. Giles, "Compiling keyphrase candidates for scientific literature based on wikipedia," in *TDDL/MDQual/Futurity@TPDL*, 2017.

[9] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 424–433.

[10] H.-H. Chen, C.-A. Chung, H.-C. Huang, and W. Tsui, "Common pitfalls in training and evaluating recommender systems," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 37–45, 2017.