環境教育研究，第十七卷，第一期，2021 年，85-128 頁
10.6555/JEER.17.1.085

探討環境教育論文的文件自動分類技術—以 2013-2018 年環境教育研討會摘要為例

張益誠[1]、張育傑[2]、余泰毅[3]*

[1] 國立宜蘭大學環境工程學系副教授

[2] 臺北市立大學地球環境暨生物資源學系教授

[3] 銘傳大學風險管理與保險系教授

摘要

本研究收集中華民國環境教育學會歷年舉辦的環境教育研討會論文摘要，透過文件自動分類技術，探討環境教育領域文章的詞彙特色與分類的一致性，運用的技術涵括自然語言處理、二階段集群分析、文字雲、共詞分析與關聯規則分析。本文將研討會論文摘要導入中研院中文詞知識庫之自然語言處理演算法，進行斷詞處理，期間採用環境教育專家意見進行輔助斷詞文字處理，將語料庫製成量化的 TF-IDF (Term Frequency- Inverse Document Frequency，詞頻-反向文件頻率)結構化樣式。應用二階段集群分析技術對於 TF-IDF 權重矩陣進行文章自動分類，同時運用文字雲、共詞分析與關聯規則分析，展現類別文章的詞彙特色以及勾稽分類文章的一致性。透過 2013-2018 年的 561 篇研討會論文摘要結果發現，斷詞後的原始關鍵詞彙共計 4980 個，前 500 大(10%)詞彙可以解釋 74.1%的累積詞頻，TF-IDF 權重對於篩選環境教育專業詞彙的篩選，可以符合關鍵少數原則。分析階層式集群分析法的總殘差下降幅度，決定 K-means 集群數目為六類，與歷史文獻比對環境教育的主題，各集群文件的主題分類為：(1)環境政策法規；(2)永續發展；(3)環境倫理、能源資源永續利用；(4)災害防救、能源資源永續利用；(5)氣候變遷；(6)環境倫理。本研究運用文字雲列出各類別高 TF-IDF 權重的詞彙、文章數量及其比例；採用勾稽方式評估環境教育主題分類的一致性，列出各類別最小距離的前三名文章題目、關鍵詞以及距離，發現各類別的文章主題的確符合一致性。此外，依據分類結果進行 Web 圖的繪

製，篩選重要關鍵詞彙以及其關聯規則，進而建議不同環境教育主題類別的重要關鍵詞彙。對於環境教育領域的自然語言斷詞處理程序以及自動文件分類勾稽而言，必須仰賴領域專家協助，方可提供正確與一致的斷詞與分類結果。

關鍵字： 二階段集群分析、文字探勘、文字雲、共詞分析、關聯規則分析

投稿日期：2021 年 04 月 17 日；接受日期：2021 年 05 月 30 日 *為通訊作者

柒、參考文獻

丁怡婷、劉志光(2010)。文字探勘技術應用於中醫診斷腦中風之研究。**數據分析，5**(4)，41-64。

【Ding, Y. T., & Liu, C. K. (2010). Applying text mining to diagnosis stroke on traditional Chinese medicine. *Journal of Data Analysis, 5*(4), 41-64. doi: 10.6338/JDA.201008_5(4).0003】

尹其言、楊建民(2010)。應用文件分群與文字探勘技術於機器學習領域趨勢分析 以 SSCI 資料庫為例。**長榮大學學報，14**(2)，1-16。

【Yin, C. Y., & Yang, J. M. (2010). Trend analysis in machine learning research from SSCI Database by document clustering manipulation and text mining methodology. *Journal of Chang Jung Christian University, 14*(2), 1-16. doi: 10.30115/JCJCU.201012.0001】

方瑀紳、李隆盛(2014)。1994-2013 年國內外科技教育學位論文研究取向之比較。**科技管理學刊，19**(3)，33-61。

【Fang, Y. S., & Lee, L. S. (2014). A comparison of research orientations between domestic and international technology education theses and dissertations in 1994-2013. *Journal of Technology Management, 19*(3), 33-61.】

王惠嘉、黃天祥、劉姿蘭(2013)，探討文字探勘方法對電子病歷進行 ICD-9-CM 自動編碼之研究。**醫療資訊雜誌，22**(1)，35-50。

【Wang, H. C., Huang, T. H., & Liu, T. L. (2013). Exploring text-mining methods to predict ICD-9-CM codes using electronic patient records. *The Journal of Taiwan Association for Medical Informatics, 22*(1), 35-50.】

吳家豪、馬麗菁(2017)。線上健康類新聞之分析與預測—巨量資料架構。**企業管理學報，113**，1-29。

【Wu, J. H., & Ma, L. C. (2017). On-line health news analysis and prediction: a framework of big data. *Journal of Management and Business Research, 113*, 1-29. doi: 10.3966/102596272017060113001】

吳慧珉、楊小億、施淑娟、許天維(2019)。一對一畢氏定理對話式智慧家教系統之建置與成效評估。**數位學習科技期刊，11**(3)，1-28。

【Wu, H. M., Yang, H. Y., Shih, S. C., & Sheu, T. W. (2019). Developing one-to-one dialogue based intelligent tutoring system for Pythagoras theorem and its effectiveness. *International Journal on Digital Learning Technology, 11*(3), 1-28. doi: 10.3966/2071260X2019071103001】

李宜玫(2012)。數學低成就學習動機之類型與區別分析：中小學弱勢學生與一般學生之比較。**教育科學研究期刊，57**(4)，39-71。

【Lee, Y. M. (2012). Discriminating math low-achievement motivation patterns: comparing disadvantaged and other students in elementary and junior high school. *Journal of Research in Education Science*s, *57*(4), 39-71.】

汪憶湘(2019)。應用文字探勘技術探討學生之戒菸經驗。**健康生活與成功老化學刊，11**(1)，1-13。

【Wang, Y. H. (2019). Applying text mining technology to explore student's experiences of quitting smoking. *Journal of Healthy Life and Successful Aging, 11*(1), 1-13.】

辛懷梓、張自立、王國華(2011)。內容分析 10 年間環境教育的研究方法與趨勢。**東海大學教育評論，6**，24-46。

【Hsin, H. T., Chang, T. L., & Wang, K. H. (2011). Content analysis of a 10-year retrospective and prospective view of research and tendency of environmental education. *Tunghai Educational Review, 6*, 24-46.】

周智勳、丁泓丞(2013)。基於關聯度指標之網路文件語意分析與文句摘要。**資訊科技與應用期刊，7**(3)，89-94。

【Chou, C. H., & Ding, H. C. (2013). Semantic analysis and text mining of web documents basing on the relation indices. *Journal of Information Technology and Applications, 7*(3), 89-94. doi: 10.6302/JITA.201309_7(3).0004】

林佳慶、謝雨蓁(2019)。以集群分析方法探討臺灣大學生進行開放式網路資源探究之個人化數位內容策展模式。**數位學習科技期刊，11**(2)，37-55。

【Lin, C. C., & Sie, Y. J. (2019). Employing cluster analysis to explore behavioral patterns of Taiwanese college students' digital content curation in the open-ended online resource inquiry. *International Journal on Digital Learning Technology, 11*(2), 37-55. doi: 10.3966/2071260X2019041102002】

林宜歆、林嶔、葉釋仁、蘇遂龍(2018)。利用文字探勘建立醫學主題詞與基因名稱之關聯性。**台灣公共衛生雜誌，37**(1)，12-23。

【Lin, Y. H., Lin, C., Yeh, S. J., & Su, S. L. (2018). Associations between medical subject headings and gene names based on text-mining in PubMed. *Taiwan Journal of Public Health, 37*(1), 12-23. doi: 10.6288/TJPH.201802_37(1).106086】

林俊成、王培蓉、詹為巽(2018)。運用共詞與社會網絡分析探討 2008-2017 年臺灣林學研究期刊重點主題與結構。**中華林學季刊，51**(3)，217-229。

【Lin, J. C., Wang, P. J., & Chan, W. H. (2018). Exploring main themes and structures in Taiwan forestry research journal from 2008 to 2017 using co-word and social network analysis. *Quarterly Journal of Chinese Forestry, 51*(3), 217-229.】

林柏宇、謝邦昌、廖佩珊(2016)。基於 Python 之文字探勘平臺。**數據分析，11**(6)，35-61。

【Lin, P. Y., Shia, B. C., & Liao, P. S. (2016). Text mining platform with Python. *Journal of Data Analysis, 11*(6), 35-61. doi: 10.6338/JDA.201612_11(6).0003】

林效荷、江志民、夏學理(2009)。複合式休閒運動市場區隔之研究。**數據分析，4**(5)，165-195。

【Lin, H. H., Chiang, C. M., & Hsia, H. L. (2009). A study on the segmentation of complex leisure athletics market. *Journal of Data Analysis, 4*(5), 165-195. 10.6338/JDA.200910_4(5).0007】

林頌堅(2017)。以開放資料的教師學術專長彙整表為基礎之學科標準分類分析。**教育資料與圖書館學，54**(1)，69-95。

【Lin, S. C. (2017). Analyses of the standard classification of fields based on the directory of faculty expertise open data. *Journal of Educational Media & Library Sciences, 54*(1), 69-95. doi: 10.6120/JoEMLS.2017.541/0046.RS.AM】

邱登裕、潘雅真(2006)。結合資訊檢索與分群演算法建構知識地圖。**資訊管理學報，13**(S)，137-160。

【Chiu, D. Y., & Pan, Y. C. (2006). Combining information retrieval and clustering algorithm to construct a knowledge map. *Journal of Information Management,13*(S), 137-160.】

邵軒磊(2019)。當代西方民主研究論述分析：知識系譜與文字探勘。**哲學與文化，46**(2)，33-56。

【Shao, H. L. (2019). Discourse analysis of contemporary western research of democracy: knowledge genealogy and text mining. *Universitas-Monthly Review of Philosophy and Culture, 46*(2), 33-56.】

翁政雄(2011)。從購買意願資料中挖掘高度相關性的關聯規則。**資訊管理學報，18**(4)，119-138。

【Weng, C. H. (2011). Mining association rules with high correlation from the purchasing intension data. *Journal of Information Management, 18*(4), 119-138.】

郝沛毅、歐仁彬、黃天受、林振穎、吳建生(2018)。透過新聞文章預測股價漲跌趨勢—結合情緒分析、主題模型與模糊支持向量機。**資訊管理學報，25**(4)，363-395。

【Hao, P. Y., Ou, J. B., Huang, T. S., Lin, Z. Y., & Wu, J. S. (2018). Sentiment and topic analysis on financial news for stock movement prediction by using fuzzy support vector machine. *Journal of Information Management, 25*(4), 363-395.】

馬桂新(2007)。**環境教育學(第二版)**。北京：科學出版社。

【Ma, G. X. (2007). *Environmental Education (2nd edition)*. Beijing: Science Press.】

高翠霞、張子超(2016)。環境教育的發展脈絡與融入十二年國教的方法。**課程與教學，19**(2), 27-51.

【Kao, T. S., & Chang T. C. (2016). Making sense of environmental education: Key themes for infusion into the curricula in new education reform. *Curriculum & Instruction Quarterly, 19*(2), 27-51. doi: 10.6384/CIQ.201604_19(2).0002】

張心馨(2006)。消費者對 Internet 智慧代理人的科技特性、任務特性及任務—科技配適度之實質接受度。**資訊管理學報，13**(1)，271-308。

【Chang, H. H. (2006). The application of the task-technology fit model to consumer acceptance of internet intelligence agent. *Journal of Information Management, 13*(1), 271-308.】

曹修源、方鄒昭聰、林慶昌、吳采軒(2019)。創新的社群文字探勘方法分析 2018 台北市市長候選人形象定位。**電子商務研究，17**(4)，277-293。

【Tsao, H. Y., Fangtsou, C. T., Lin, C. C., & Wu, T. H. (2019). The positioning analysis for the candidates of Taipei city mayor election in 2018 via text mining online. *Electronic Commerce Studies, 17*(4), 277-293.】

曹開明、黃鈴媚、劉大華(2017)。數位語藝批評與文本探勘工具—以反核臉書粉絲團形塑幻想主題為例。**資訊社會研究，32**，9-49。

【Tsao, K. M., Huang, L. M., & Liu, T. H. (2017). An exploration of text mining tools on digital rhetoric criticism: how anti-nuclear Facebook fanpages shaped fantasy themes. *The Journal of Information Society, 32*, 9-49. doi: 10.29843/JCCIS.201701_(32).0002】

許曉霈、李子奇、張瑞瑤、於淑娟、黃久美(2018)。應用文字探勘探索痛經婦女之疾病經驗。**健康生活與成功老化學刊，10**(1)，39-53。

【Hsu, H. P., Lee, T. C., Chang, J. Y., Yu, S. C., & Huang, C. M. (2018). Using text mining techniques to explore the illness experience among women with dysmenorrhea. *Journal of Healthy Life and Successful Aging, 10*(1), 39-53.】

陳譽晏(2015)。運用 R Shiny 建立文字探勘平台之語意分析及輿情分析。**數據分析，10**(6)，51-78。

【Chen, Y. Y. (2015). Semantic analysis and public opinion analysis under the R Shiny text mining platform. *Journal of Data Analysis, 10*(6), 51-78. doi: 10.6338/JDA.201512_10(6).0003】

曾元顯、林瑜一(2011)。內容探勘技術在教育評鑑研究發展趨勢分析之應用。**教育科學研究期刊，56**(1)，129-166。

【Tseng, Y. H., & Lin, Y. I. (2011). The application of content mining techniques to the analysis of educational evaluation research trends. *Journal of Research in Education Sciences, 56*(1), 129-166.】

舒玉、陳鈺潔、黃天麒(2019)。護理教育未來式—以虛擬實境誘發動機之整合學習模式。**護理雜誌，66**(2)，22-28。

【Shu, Y., Chen, Y. J., & Huang, T. C. (2019). Exploring the future of nursing education: an integrated motivation learning model based on virtual reality. *The Journal of Nursing*, *66*(2), 22-28. doi: 10.6224/JN.201904_66(2).04】

黃俊英(2000)。**多變量分析**。臺北：華泰書局。

【Huang J. Y. (2000). *Multivariate Analysis*. Taipei: Hwa Tai Publishing.】

黃嘉郁(1999)。台灣地區環境教育學位論文研究主題之分析。**中師數理學報，2**(2)，69-92。

【Huang, C. Y. (1999). The analysis of graduation thesis subject on environmental education in Taiwan. *Chung-Shi Bulletin of Mathematics and Science*, *2*(2), 69-92.】

楊冠政(1997)。**環境教育**。臺北市，明文書局。

【Yang, G. Z. (1997). *Environmental Education.* Taipei: Ming Wen.】

楊錦生、謝佩芸、施曉萍(2017)。社群媒體中顧客知識之挖掘：意見探勘技術開發。**臺大管理論叢，27**(2S)，1-28。

【Yang, C. S., Xie, P. Y., & Shih, H. P. (2017). Mining consumer knowledge from social media: development of an opinion mining technique. *NTU Management Review*, *17*(2S), 1-28. doi: 10.6226/NTUMR.2017.JUN.F104-008】

趙妤瑄、王豐緒(2017)。情緒詞權重計算與分類演算法對於情緒分析結果之影響—以臉書粉絲團議題分析為例。**電子商務研究，15**(2)，147-166。

【Zhao, Y. H., & Wang, F. H. (2017). On the effects of emotional term-weight calculation and document classifiers on sentiment analysis: taking topic analysis of Facebook fan groups as an example.*Electronic Commerce Studies*, *15*(2), 147-166.】

蔡介元、張百棧、王錫中(2003)。運用關聯法則技術與類神經網路於產品開發設計之研究。**工業工程學刊，20**(2)，101-112。

【Tsai, C. Y., Chang, P. C., & Wang, S. J. (2015). Applying association-rule techniques and artificial neural networks to product development. *Journal of the Chinese Institute of Industrial Engineers*, *20*(2), 101-112.】

蔡逸芬、陳品華(2015)。國小高年級學童課外閱讀自我決定動機之研究。**教育心理學報，46**(3)，425-448。

【Tsai, Y. F., & Chen, P. W. (2015). Self-determined motivation for extracurricular reading among fifth and sixth graders. *Bulletin of Educational Psychology, 46*(3), 425-448. doi: 10.6251/BEP.20140627】

盧姝如、朱慶雄、盧昉暄(2013)。數位化桌上遊戲創新學習模式之開發設計—以國小中年級生海洋教育為例。**國民教育，53**(4)，45-55。

【Lu, S. J., Chu, C. H., & Lu, F. H. (2013). Designing an innovative learning digital board game: a case study of marine education for middle-grade elementary. *Elementary Education, 53*(4), 45-55.】

謝元晟、程美華、張光昭(2016)。運用 R 建立文字探勘平台應用於電視收視率預測。**數據分析，11**(3)，109-134。

【Hsieh, Y. C., Cheng, M. H., & Chan, K. C. (2016). Predicting TV ratings under the R to build the text mining platform. *Journal of Data Analysis, 11*(3), 109-134. doi: 10.6338/JDA.201606_11(3).0007】

謝吉隆、楊芯淳(2018)。從「應變自然」到「社會應變」：以文字探勘方法檢視國內風災新聞的報導。**教育資料與圖書館學，55**(3)，285-318。

【Hsieh, J. L., & Yang, B. C. (2018). A content analysis and comparison of typhoon news in early and recent periods based on the text-mining approach. *Journal of Educational Media & Library Sciences, 55*(3), 285-318. doi: 10.6120/JoEMLS.201811_55(3).0022.RS.BM】

羅鳳珠(2011)。以語言知識庫為基礎的智慧型作詩填詞輔助系統。**教學科技與媒體，95**，36-52。

【Lo, F. J. (2011). Empowering classical poetry and lyrics composition: an intelligence model based on language knowledge database. *Instructional Technology & Media, 95*, 36-52.】

Worawut, D., & Wirot, Y. (2015)。以兩階段集群分析方法之比較：以泰國普吉島遊客資訊管理為例。**島嶼觀光研究，8**(4)，32-48。

【Worawut, D., & Wirot, Y. (2015). Comparison of two-stage clustering methods: SOM and K-means algorithm and hierarchical clustering and K-means algorithm in tourist information management in Phuket. *Journal of Island Tourism Research*, *8*(4), 32-48.】

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarter*ly, *36*(4). 1165-1188. doi: 10.2307/41703503

Chen, K. J., & Liu, S. H. (1992, August). *Word identification for Mandarin Chinese sentences.* Proceedings of the 14th conference on Computational linguistics, Volume 1, pp. 101-117. doi: 10.3115/992066.992085

Chen, S. Y., & Liu, S. Y. (2020). Developing students' action competence for a sustainable future: a review of educational research. *Sustainability*, *12*(4), 1374. doi: 10.3390/su12041374

Chen, Y. L., & Weng, C. H. (2008). Mining association rules from imprecise ordinal data. *Fuzzy Sets and Systems*, *159*(4), 460-474. doi: 10.1016/j.fss.2007.10.005

Chen, Y. L., Liu, Y. H., & Ho, W. L. (2013). A text mining approach to assist the general public in the retrieval of legal documents. *Journal of the Association for Information Science and Technology*, *64*(2), 280-290. doi: 10.1002/asi.22767

Corrales-Garay, D., Ortiz-de-Urbina-Criado, M., & Mora-Valentín, E. M. (2019). Knowledge areas, themes and future research on open data: A co-word analysis. *Government Information Quarterl*y, 36(1), 77-87. doi: 10.1016/j.giq.2018.10.008

Dijcks, J. (2013). *Oracle: Big data for the enterprise.* Redwood Shores, CA: Oracle Corporation.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*(6), 817-842. doi: 10.1016/S0306-4573(00)00051-0

Entwistle, N., Tait, H., & McCune, V. (2000). Patterns of response to an approaches to studying inventory across contrasting groups and contexts. *European Journal of Psychology of Education*, 15(1), 33-48. doi: 10.1007/BF03173165

Gao, Y., Xu, Y., & Li, Y. (2015). Pattern-based topics for document modelling in information filtering. *IEEE Transactions on Knowledge and Data Engineering, 27*(6), 1629-1642. doi: 10.1109/TKDE.2014.2384497

Garcia, S., Derrac, J., Cano, J., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(3), 417-435. doi: 10.1109/TPAMI.2011.142

Girmay, G., & Bhaskari, D. L. (2018). Big data analytics and security: a big choice and challenge for the generation. In S. Satapathy, V. Bhateja, & S. Das (Eds.), *Smart computing and informatics* (pp. 209-217). Springer: Singapore. doi: 10.1007/978-981-10-5547-8_22

Gunter, B., Koteyko, N., & Atanasova, D. (2014). Sentiment analysis: A market-relevant and reliable measure of public feeling? *International Journal of Market Research, 56*(2), 231-247. doi: 10.2501/IJMR-2014-014

Guo, D., Chen, H., Long, R., Lu, H., & Long, Q. (2017). A co-word analysis of organizational constraints for maintaining sustainability. *Sustainability*, *9*(10), 1928. doi: 10.3390/su9101928

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33*(3), 464-472. doi: 10.1016/j.ijinfomgt.2013.01.001

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML, 98*, 137-142. doi: 10.1007/BFb0026683

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). *Big data: Issues and challenges moving forward*. Proceedings of the 2013 46th Hawaii

International Conference on System Sciences, pp. 995-1004. IEEE. doi: 10.1109/HICSS.2013.645

Keim, D., Qu, H., & Ma, K. L. (2013). Big-data visualization. *IEEE Computer Graphics and Applications, 33*(4), 20-21. doi: 10.1109/MCG.2020.3043984

Khasseh, A. A., Soheili, F., Moghaddam, H. S., & Chelak, A. M. (2017). Intellectual structure of knowledge in iMetrics: A co-word analysis. *Information Processing & Management, 53*(3), 705-720. doi: 10.1016/j.ipm.2017.02.001

Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications, 40*(5), 1847-1857. doi: 10.1016/j.eswa.2012.09.017

Lai, C. H., & Liu, D. R. (2009). Integrating knowledge flow mining and collaborative filtering to support document recommendation. *Journal of Systems and Software, 82*(12), 2023-2037. doi: 10.1016/j.jss.2009.06.044

Lavie, T., Sela, M., Oppenheim, I., Inbar, O., & Meyer, J. (2010). User attitudes towards news content personalization. *International Journal of Human-Computer Studies, 68*(8), 483-495. doi: 10.1016/j.ijhcs.2009.09.011

Liao, S. H., & Wen, C. H. (2007). Artificial neural networks classification and clustering of methodologies and applications: literature analysis from 1995 to 2005. *Expert Systems with Applications, 32*(1), 1-11. doi: 10.1016/j.eswa.2005.11.014

Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham, MA: Academic Press. doi: 10.1016/B978-0-12-386979-1.09001-0

Shen, L., Xiong, B., & Hu, J. (2017). Research status, hotspots and trends for information behavior in China using bibliometric and co-word analysis. *Journal of Documentation, 73*(4), 618-633. doi: 10.1108/JD-10-2016-0125

Wang, S. I., & Manning, C. D. (2012, July). *Baselines and bigrams: Simple, good sentiment and topic classification.* Proceedings of the 50th Annual Meeting of

the Association for Computational Linguistics, Volume 2, pp. 90-94. Association for Computational Linguistics.

Wang, W., Chen, X., Zou, Y., Wang, H., & Dai, Z. (2010, April). *A focused crawler based on Naive Bayes classifier*. 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 517-521. IEEE. doi: 10.1109/IITSI.2010.30

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems, 26*(3), 1-37. doi: 10.1145/1361684.1361686

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758-2765. doi: 10.1016/j.eswa.2010.08.066

作者簡介：

張益誠　國立宜蘭大學環境工程學系副教授
電話：03-9357400 ext. 7586
電子郵件：icchang@niu.edu.tw
通訊處：260 宜蘭縣宜蘭市神農路一段一號

張育傑　臺北市立大學地球環境暨生物資源學系教授
電話：02-2311-3040 ext. 3151
電子郵件：yjchang@utaipei.edu.tw
通訊處：100 臺北市中正區愛國西路一號

余泰毅　銘傳大學風險管理與保險系教授
電話：02-28824564 ext. 2619
電子郵件：yti@mail.mcu.edu.tw
通訊處：111 台北市士林區中山北路五段 250 號

Chang, I-Cheng
Associate Professor, Department of Environmental Engineering, National Ilan University
Tel: 03-9357400 ext. 7586
Email: icchang@niu.edu.tw
Address: No.1, Sec. 1, Shennong Rd., Yilan City, Yilan County 260, Taiwan (R.O.C.)

Chang, Yu-Jie
Professor, Department of Earth and Life Science, University of Taipei
Tel: 02-2311-3040 ext. 3151
Email: yjchang@utaipei.edu.tw
Address: No.1, Ai-Guo West Road, Taipei 100, Taiwan (R.O.C.)

Yu, Tai-Yi
Professor, Department of Risk Management and Insurance, Ming Chuan University
Tel: 02-28824564 ext. 2619
Email: yti@mail.mcu.edu.tw
Address: No. 250, Zhong Shan N. Rd., Sec. 5, Taipei 111, Taiwan (R.O.C.)

Exploring Automatic Document Classification of Environmental Education Research Papers Using Text Mining Manners: An Analysis on Abstracts from the International Conference on Environmental Education between 2013-2018

I-Cheng Chang[1], Yu-Jie Chang[2], Tai-Yi Yu[3]*

[1.] Associate Professor, Department of Environmental Engineering, National Ilan University

[2.] Professor, Department of Earth and Life Science, University of Taipei

[3.] Professor, Department of Risk Management and Insurance, Ming Chuan University

Abstract

This research collects abstracts from the International Conference on Environmental Education Academia and Practices held by the Chinese Society for Environmental Education (CSEE) between 2013-2018. Through the automatic topic classification techniques, it explores the vocabulary characteristics of classified articles in the field of environmental education and the consistency of classification. Techniques applied include natural language processing, two-step cluster analysis, word cloud, co-word analysis and association rules analysis. In this study, the research abstracts from the conference papers have been imported into the natural language processing algorithm of the CKIP Chinese Lexical Knowledge Base of Academia Sinica for word segmentation. The opinions of environmental education experts have been applied for auxiliary word segmentation, and corpora of abstracts from conference papers have been made into quantitative Term Frequency-Inverse Document Frequency (TF-IDF) weights. Afterwards, two-step cluster analysis technology has been performed to automatically classify articles clusters; the techniques of word cloud, co-word analysis and association rule analysis have been used to show the vocabulary characteristics of distinct clustered articles and the consistency of the classified articles. Based on the results of 561 abstracts of conference papers from 2013 to 2018, the number of original keywords after word

segmentation is 4,980. The top 500 (10%) words account for the 74.1% of the cumulative word frequency. The selection of professional vocabularies can match the Pareto principle. The two-step cluster analysis classifies the number of K-means clusters into six categories, namely (1) environmental policy and regulation; (2) sustainable development; (3) environmental ethics and sustainable use of energy and resources; (4) disaster prevention and response, sustainable use of energy and resources; (5) climate change; (6) environmental ethics. This study applies the word cloud to enlist the dominant words with high TF-IDF weights, word frequency and proportions for distinct clusters; utilizes the cross-check method to assess the consistency of topic classification and enlists the top three article titles and keywords with the smallest distance in each category. In addition, the web map is drawn in accordance with classification results, and dominant keywords and their association rules are screened, and then dominant keywords of different themes have been suggested. For natural language word segmentation process and automatic document classification in topic modeling, the assistance of domain experts for environmental education plays a crucial role in providing correctness and consistence in aforementioned academic tasks.

Keywords: two-step cluster analysis, text mining, word cloud, co-word analysis, association rules