國立臺灣師範大學教育學院圖書資訊學研究所

碩士論文

Graduate Institute of Library and Information Studies

College of Education

National Taiwan Normal University

Master’s Thesis

中文期刊論文資訊擷取之研究

—以圖書資訊學領域為例

Information Extraction From Chinese Scientific Article

— A Case Study of Library and Information Science

黃冠綸

Huang, Guan-Lun

指導教授：曾元顯 博士

Advisor：Tseng, Yuen-Hsien, Ph.D.

中華民國一一二年六月

June 2023

摘要

目前的科學文獻數量以相當驚人的速度在成長當中，如何將這些巨量、富含知識的科學文獻內容從 PDF 中剖析出來，是當前相當重要的課題。然而在臺灣鮮少看到有相關的研究，本研究的目的在於提出臺灣中文學術期刊資訊擷取的解決方案，並以圖書資訊學領域期刊論文為例。

本研究透過重新訓練開放原始碼科學文獻剖析工具 GROBID，達成擷取中文學術期刊資訊（篇名、作者、摘要、關鍵字、具章節邏輯的內文等）的目的，並透過十倍交叉驗證法(10 Fold Cross-Validation)來評估訓練成效。本研究透過重新訓練後的模型剖析 725 篇台灣圖書資訊領域期刊論文，觀察與分析可能影響剖析成功率的原因。

本研究發現，三個模型（Segmentation、Header、Fulltext）在訓練資料 n = 100 與 n = 250 時， F1 score 沒有特別明顯的成長。相同期刊的論文會因為不同年代出版而有不同的版型，這個現象對於剖析成功率有影響。

本研究透過將剖析後的科學文獻內文匯入ＱＡ系統中，使得ＱＡ系統可以回答更專業的問題，作為對剖析科學文獻後的加值利用範例。

關鍵字：資訊擷取、開放原始碼、GROBID、全文資料集

Abstract

The current volume of scientific literature is growing astonishingly, and the extraction of the vast amount of knowledge-rich content from scientific article PDFs has become a critical issue. However, there is a scarcity of research focusing on this area in Taiwan. This study aims to propose a solution for extracting information from Chinese academic journals in Taiwan, using the field of library and information science as an example.

This study successfully extracts Chinese academic journal information by retraining the open-source scientific literature parsing tool GROBID, including article titles, authors, abstracts, keywords, and structured full text with logical sections. The effectiveness of the training is evaluated using a ten-fold cross-validation method. The retrained model is applied to analyze 725 journal articles in the library and information science field in Taiwan, observing and analyzing factors that may affect the success rate of parsing.

The study found that the three models (Segmentation, Header, Fulltext) did not significantly improve the F1 score when trained on n = 100 and n = 250 data samples. The variation in document layouts due to different publication years of articles within the same journal impacts the parsing success rate.

Finally, we incorporate the parsed scientific literature into a Question-Answering (QA) system, making an example of the added value of parsed scientific literature.

Keywords: Information Extraction, open source, GROBID, full text dataset