國立中興大學圖書資訊學研究所

碩士學位論文

基於語義之多層式圖書自動分類實證研究

Empirical Study of Semantic-based Multi-layered Automatic Book Classification

指導教授：郭俊桔 June-Jei Kuo

研　究　生：陳雨沛 Yu-Pei Chen

中華民國一〇九年七月

摘要

本研究旨在探討多層式圖書自動分類系統應用於圖書館分類編目的工作流程，並嘗試導入語義概念作為改進分類成效之策略，目的是為解決人工分類的一致性、分類效率等問題，改善分類編目的品質，並針對資料量或文件特徵量可能不足之課題，利用 Word2Vec 能夠保留目標詞與其上下文之間的語義關係之特性，將帶有語義的詞彙擴展特徵詞彙，藉此改善分類成效。本研究蒐集的資料取自大學圖書館 HyRead ebook 中文電子書，共 17,650 冊，擷取圖書題名、目次、摘要等三種片段內容，使用 TF×IDF 加權統計挑選對於類別具代表性的特徵詞彙，分類類別為圖書分類號，依《中文圖書分類法》從中隨機挑選的圖書分類號涵蓋十大類，共 581 個圖書分類號，並使用四種分類器應用於多層式圖書自動分類系統；於語義方面，實驗使用 Word2Vec 詞向量工具訓練語料，並建構類索引典之同義詞詞典，再以擴充的方式擴展特徵詞彙，最後，以正確率評估分類效能。實驗結果顯示將多層式圖書自動分類系統應用於圖書館分類編目具有更好的分類成效，並且所提出的策略確實能夠提升圖書分類的準確度，經詞彙擴充之後，正確率為 62.50%，有效減少多層式分類最終仍無法判斷的數量，降低因無法分類而增加的錯誤率，由此實證 Word2Vec 適合作為圖書自動分類的輔助工具。然而，對圖書自動分類而言，存在多個主題的圖書內容可能是影響分類成效的因素之一，此現象亦是影響圖書分類號一致性的可能原因之一，若將其他圖書館系統所給予的圖書分類號視為正確類別，使用基於語義之多層式圖書自動分類系統得到的正確率接近 70%，正確率共提升了 7.14%。因此，在未來圖書館實務上，尚須考慮到存在多個主題的圖書內容對圖書自動分類的影響。

關鍵字：分類號；階層式；圖書自動分類；Word2Vec

Abstract

This study aims to explore the application of the multi-layered automatic book classification using the majority voting strategy in a library environment, and proposes a semantic-based approach as a strategy to improve the classification performance. The proposed strategy uses Word2Vec, which can extract the deep semantic relationships between words, to expand words features for improving the classification performance. The collection of data from HyRead ebook, which was ordered by the university library, was utilized as the training and testing corpus, including book titles, table of contents, abstracts. In total, 17,650 books were collected. And the classification classes were the book classification numbers. A total of 581 classification numbers were selected randomly from New Classification Scheme for Chinese Libraries. Naïve Bayes, SVM, Decision Tree C4.5, and kNN were applied to the multi-layered automatic book classification. Regarding the proposed semantic-based approach, this study used Word2Vec as a tool for training word embedding. First, a thesaurus was built by the training results, and next the word features of the data set for classification were expanded. Under the principle of accuracy, experimental results showed that the performance of the multi-layered automatic book classification outperformed the traditional automatic book classification in a library environment. And the strategy proposed in this study could indeed improve the accuracy of book classification, achieving 60%. After expanding the word features of the data set, the accuracy rate promoted 1.79%. The above results proved that Word2Vec tool was helpful for automatic book classification. It could effectively reduce the number of unable classified books, which were classified through the multi-layered automatic book classification. However, book content with multiple topics may be one of the factors that affect classification effectiveness for automatic book classification. If the book classification number, given by other library systems, was regarded as the correct class, the accuracy rate could be approximately 70% by the semantic-based multi-layered automatic book classification using the majority voting strategy. The accuracy rate could be further increased by 7.14%. Therefore, it is necessary to consider the effect of book content with multiple topics for automatic book classification in the future library practice.

Keywords: classification number; layer style; automatic book classification; Word2Vec