氏　　名	INZALI NAING
授与した学位	博　士
専攻分野の名称	工　学
学位授与番号	博甲第　　７３８０　　号
学位授与の日付	２０２５年　９月　２５日
学位授与の要件	自然科学研究科　　産業創成工学専攻 （学位規則第４条第１項該当）
学位論文の題目	A Study of Reference Paper Collection System Using Web Scraping and BERT Model （Web スクレイピングと BERT モデルを用いた参考文献収集システムの研究）
論文審査委員	教授　舩曳　信生　　教授　田野　哲　　教授　野上　保之

学位論文内容の要旨

Research Background and Motivation

The manual collection of academic research papers presents significant challenges for researchers across all disciplines. Traditional methods of finding relevant literature through platforms like Google Scholar are time-consuming and frequently result in inaccessible papers due to paywalls, access restrictions, or broken links. Furthermore, researchers must manually verify the credibility and relevance of collected papers, creating troublesome in the research process. This thesis addresses these challenges by developing an automated reference paper collection system that leverages web scraping technology and natural language processing to streamline academic literature discovery. The system aims to reduce collection time greatly while improving paper discovery capabilities and relevance accuracy.

Primary Objective

The primary objective is to design and implement an automated reference paper collection system that significantly improves the efficiency and accuracy of academic literature discovery for researchers.

Key Contributions

System Design and Architecture

Development of a comprehensive web-based reference paper collection system that accepts thesis titles and keywords as input and automatically generates lists of relevant, downloadable research papers. The system architecture incorporates modern web technologies including Angular frontend, Python Flask backend while cooperation with Selenium and BERT model.

System Implementation

Implementation of web scraping functionality using Selenium automation framework, featuring reliable website navigation strategies, connecting PDF download mechanisms, and efficient content extraction algorithms. The system incorporates multithreading capabilities to achieve 2.5x performance improvements over sequential processing approaches. The system is carefully structured using Docker for easy installation.

Paper Filtering and Relevance Assessment

Development of advanced paper filtering algorithms includes continuously handling edge cases of pdf format and error cases. The assessment of paper is utilizing BERT (Bidirectional Encoder Representations from Transformers) models for similarity. The system achieves an average of 90% accuracy in identifying relevant papers across diverse academic disciplines, significantly outperforming traditional keyword-based approaches.

Improved Automation with Anti-Detection Capabilities

Implementation of an improved system version incorporating Selenium Stealth technology to overcome anti-scraping mechanisms deployed by academic databases. This enhancement ensures consistent system accessibility and prevents IP blocking issues that commonly affect traditional web scraping approaches.

Thesis Structure and Content Overview

Chapter 1: Introduction

This chapter introduces the evolution of the reference paper collection system and outlines the content of the thesis.

Chapter 2: Problem Statement

This chapter describes existing paper collection methodologies, web scraping technologies including Selenium framework, and related automated academic information retrieval systems.

Chapter 3: System Design and Architecture

This chapter presents the overall system architecture and design of the reference paper collection system, including core components, data flow mechanisms, and technology stack selection.

Chapter 4: System Implementation

This chapter presents the implementation of web scraping functionality using Selenium automation, covering PDF download mechanisms and content extraction algorithms.

Chapter 5: System Enhancements

This chapter proposes the system enhancements including multithreading implementation, Docker containerization and covering PDF file handling mechanisms for the reference paper collection system, focusing on performance optimization and deployment flexibility.

Chapter 6: Performance Evaluation

This chapter presents the experimental evaluation of system performance, including efficiency metrics, accuracy assessment and system usability scoring with manual collection methods.

Chapter 7: Improved Version with Selenium Stealth

This chapter reviews previous works related to the automated reference paper collection system and presents improvements using Selenium Stealth about anti-scraping mechanisms.

Chapter 8: Additional Contribution about Selenium as a Unit Testing Tool

This chapter presents that Selenium can be used as the automated web testing tool and reviews how it can cooperate with Allure and Behave frameworks to create an automatic testing in the web-based programming learning assistant system.

Chapter 9: Related Work

This chapter presents paper works related to this thesis.

Chapter 10: Conclusion

This chapter concludes this thesis with some future works.

論文審査結果の要旨

In this thesis, the applicant presented the studies on developments of a web-based reference paper collection system. The system accepts paper titles and keywords as input and automatically generates lists of relevant, downloadable research papers. The system architecture incorporates modern web technologies including Angular frontend, Python Flask backend while cooperation with Selenium and BERT model.

Firstly, she implemented web scraping functionality using Selenium automation framework, featuring reliable website navigation strategies, connecting PDF download mechanisms, and efficient content extraction algorithms. The system incorporates multithreading capabilities to achieve 2.5x performance improvements over sequential processing approaches. The system is carefully structured using Docker for easy installation.

Secondly, she developed paper filtering algorithms continuously handling edge cases of pdf format and error cases. The assessment of a reference paper is utilizing BERT (Bidirectional Encoder Representations from Transformers) models for similarity checking. The system achieves an average of 90% accuracy in identifying relevant papers across diverse academic disciplines, significantly outperforming traditional keyword-based approaches.

Thirdly, she implemented an improved system version incorporating Selenium Stealth technology to overcome anti-scraping mechanisms deployed at academic databases. This enhancement ensures consistent system accessibility and prevents IP blocking issues that commonly affect traditional web scraping approaches.

The applicant has published one journal paper, one international conference paper, and two domestic conference papers to present the contributions.

From the overall evaluation of this thesis, the applicant has satisfied the qualification condition for the doctor degree in Engineering from the Graduate School of Natural Science and Technology at Okayama University.