CHAPTER 1: INTRODUCTION

This chapter presents an overview of the thesis. It discusses background, problem statement and research question, objectives and hypothesis, research methodology, contribution, and thesis overview.

1.1 Motivation

In recent years, there has been a significant surge in the number of academic publications, with many research papers being published and found online [1]. In 2018, the number published per year is estimated to be more than 3 million [2], [3]. As the number of published papers continues to grow exponentially, researchers face challenges in effectively discovering and accessing papers that are directly relevant to their research interests. This issue becomes even more complex when researchers seek papers that bridge multiple disciplines, as traditional keyword-based search approaches may need help to capture the multidimensional nature of such cross-domain research topics.

Despite the availability of the CSO (Computer Science Ontology), there remains a pressing need for a more advanced and robust approach to ontology-based keyword detection. Existing methods, such as the CSO Classifier [4], may not fully capture the nuanced relationships between concepts in cross-domain research. This limitation becomes particularly evident when considering the semantic variations and complexities that often characterize interdisciplinary research. A more sophisticated approach is required to identify research topics that transcend disciplinary boundaries effectively.

Many studies on paper recommendation systems have been carried out with methods, approaches [5], in the ontology approach to recommendation systems, the results showed how using ontology hierarchies in the profiling process resulted in superior performance compared to regular topic lists, including getting more general topics not suggested directly [6]. The challenges in research on recommender systems for ontology-based papers include the selecting representative keywords [7] and the availability of ontology that supports exploring the linkages between papers in the same discipline and between disciplines [8].

1.2 Problem Statement and Research Question

In the paper on CSO Classifier, the semantic topic search process classifier was carried out using the word2vec word embedding method using Computer Science ontology. The performance of using CSOC in the Computer Science domain to expand the search for specific keywords in this field has an f-measure value of 74.1 [4]. Organizing scholarly papers by their relevant research topics is an essential task that facilitates various functions, including the generation of recommendations. The results of keywords and terms generated from CSOC can be good enough to produce paper recommendations in the Computer Science domain, but what if there are other terms outside that domain which may be related to other papers, for example papers at Bioinformatics conferences, Computational Biology, and Biomedicine which includes several computer science and biology terms. One potential improvement that could be made is to use a multidisciplinary ontology to enrich the reference. Ontologies offer an extensive vocabulary that includes related terms and hierarchical relationships. This expansion allows for more thorough searches, capturing relevant papers that might use varied terminology.

Currently, the CSO Classifier is limited to detecting or generating keywords from research papers within the Computer Science domain, both syntactically and semantically. This limitation hinders its ability to extend keyword detection to other cross-domains, such as Biology, which is necessary for recommending relevant multidisciplinary research papers. The current method lacks the capability to identify and recommend papers that span multiple domains, thereby limiting the discovery of interdisciplinary research connections. This limitation underscores the need for improvement, the method should be able to extend more keyword from biology term to provide references regarding the paper's relationship with other domains, enabling researchers to find recommendations for papers that are found because they have similarities in discussing the domain of Computer Science and Biology. This limitation underscores the need for improvement, particularly in the context of the recommendation system. The question then arises: what approach can be used to extend keyword detection from the Computer Science domain to other research cross-domains, such as Biology, to enhance the identification and recommendation of multidisciplinary research papers?

1.3 Objective

The purpose of this research is to Generate extended keywords for multidisciplinary research papers, particularly those spanning Computer Science and Biology, and to leverage these keywords to enhance the identification and recommendation of related papers across these domains. Several reasons for taking the biology domain as an example of a cross-domain case are as follows.

a. Relevance to Practical Applications. Medical and biological terms matter in real-world applications like healthcare and biomedical research. Studies often combine biology and computer science to introduce new algorithms or methods, especially in Computational Biology and Bioinformatics [9].

b. Complexity and Diversity of Terms. These domains offer a rich testing ground for computational techniques due to their intricate relationships and hierarchical structures [10].

1.4 Hypothesis

The hypothesis for this study is by using the CSO Classifier with the Cross-Domain Ontology relation approach, leveraging resources such as WordNet, it is possible to extend keyword detection to other domains like Biology. This method will enhance the ability to recommend related papers across multidisciplinary fields.

Premise 1: Ontology approach to the recommendation system, the results showed how the use of ontology hierarchies. This includes being able to find more general topics that are not suggested directly [6], [7]

Premise 2: The use of CSOs in the Computer Science Ontology Classifier has high accuracy compared to others, so a list of terms or related topics can be used [4]

Premise 3: Semantic similarity, using ontology for term interpretation, has been implemented for computing the conceptual similarity between natural language terms using WordNet by examining their relationships in that ontology [11], [12].

1.5 Research Methodology

This chapter describes the comprehensive research methodology employed in this study, including the research design, data collection techniques, implementation, data analysis procedures, and experiment outcomes used in this study. The steps in the methodology are as follows:

a. Problem Identification

   This step aims to identify issues in extending keywords in multidisciplinary domains for paper recommendation, define the problem addressed in this thesis, and suggest potential improvements for the chosen problem.

b. Model Design

   Model design defines functional requirements and design implementation, describing how a system is formed. It can be defined as a depiction and arrangement of several separate elements, including Cross-Domain Ontology, a new method to expand the keyword's result. The model design proposes problem solving logic.

c. Data Collection and Processing

   The data collection process for this study involved sourcing academic papers from the ACM Digital Library, specifically focusing on conferences within the realms of Bioinformatics and Computer Science. Subsequently, the compiled data underwent meticulous processing to ensure uniformity and pertinence. Papers were selectively filtered based on predefined inclusion criteria aligned with our research objectives. The ontological data from CSO and Wordnet was instrumental in structuring and categorizing the information, thus providing a robust framework for analysis.

d. Implementation

In this phase, the model designed in the previous phase is translated into a program.

e. Experiment

This stage proves the hypothesis about using the CSO Classifier and the ontology relation approach in other multidisciplinary domains found on WordNet, that can expand keywords in others domain for recommendation papers.

f. Analysis of experiment results

This section analyzes the Cosine Similarity scores obtained by the proposed method and compares them with previous research. It also analyzes the results of the survey conducted with expert respondents to measure the relevance of those papers using Precision at K evaluation.

1.6 Contribution

The main contribution of this study is to form a cross-ontology from WordNet, in this case from the biology domain and combined with CSO, which uses the CSO Classifier to produce undiscovered and new terms related to biology domain where the results of generating extended keywords can be used to recommend papers. While this study focuses on the biology domain as a case study, the proposed method is not limited to this domain alone. The adaptable approach can be applied to other domains by leveraging the availability of domains on WordNet. This evaluation could be valuable for bringing researchers together to collaborate on interests in the computer science and biology domains and to compare the differences between methods in the future.

1.7 Thesis Overview

This thesis comprises five chapters: introduction, literature review, algorithm design and implementation, experiment and data analysis and conclusions. The explanation of each chapter is as follows:

a. Introduction

This chapter discusses the background of the problem addressed in this thesis, the problem definition and research questions, the research question, objectives that will be reached, the research methodology briefly and the thesis overview.

b. Literature Review

This chapter meticulously delves into the theoretical basis of the concepts and theories that underpin and are utilized in this final project. The literature review serves as a comprehensive reference guide, ensuring the robustness of the final project's theoretical foundation.

c. Research Methodology

This chapter presents a meticulous description of the system and the proposed model, designed to address the existing problems in this thesis. The design provides a detailed overview of the system's structure, the proposed model, data flow, and system usage scheme, ensuring the precision and validity of the research methodology.

d. Experiment and Analysis

This chapter describes the testing purpose and the scenario for completing existing problems in this thesis. At the stage of the testing results analysis, the data from the testing results, the graph of the test results, and its analysis were explained.

e. Conclusions

This chapter explains the conclusion of this study. Conclusions include the results of the final analysis and explain the answer to the problem statement formulated in this thesis. The recommendation point describes suggestions regarding the possibilities that can be developed from the results of this research.