CHAPTER 1 : THE PROBLEM

1.1 Motivation

Commonly, one way to find out the interests of researchers is cited paper. It is an effective practice, but it limits the researchers to a particular community, and that is biased because of the widely cited papers. A statistician may miss the relevant papers in economics or biology because the two pieces of literature rarely quote each other and he may miss relevant papers in statistics. [1]

The other way to find out the interests of researchers is the relationships between researchers. Such author relationships could be useful for recommending articles that satisfy the information needed. It is conjectured that an article may be of interests to the user if the authors of the article under discussion are professionally close to those of the set of articles.

With data that have been archived online, it is easier to know the interests of researchers search from conferences and not only based on certain communities. Using research content the only way to understand the meaning of paper using text mining. A disadvantage of using text mining is that it could not get the context of what researcher's want to look for, but only get similar words. For example, the phrase "Semantic Similarity" and "Semantic Relatedness" have the same meaning context but if using text mining will not appear when looking for "Semantic Similarity" then the paper that says "Semantic Relatedness" will not appear.

Improvements in recommending relevant papers have been proposed in some previous studies. SHUCHEN, et al. [1] uses content and co-authorship information to identify relationships between researchers. This method extracts relationship (related)??? papers and researcher-citation, network co-author, and content. It also uses bookmarks which indicates co-authorship preferences by the way the author marks users in each conference. By using these parameters, increasing the performance of recommenders is up to 26% at some conferences. Lack of this method is not handling the context of paper.

Furthermore, Anastasios, et al. [4] uses content-based text mining and collaborative filtering, then classifies it according to the subject, which later will be used as a graph and then see the closeness between the graphs. Lack of this method is to consider rating as the highest

parameter of recommendations. Furthermore, Chandra, et al. [5] The recommendation process uses Approximate nearest neighbor search algorithm and binary-tree up to 100 trees to get the best results. The recommendation process will be more maximal when in the same domain. On the other hand, Xuan-You Liu, et al. [6] uses a cross-crawling strategy and a paper citation network analyzer. Cross crawler to search for and collect relevant search based on searches desired in paper indexing sites. Search seen similarities between terms and words based on available textual information. Lack of this method is only considering textual of a paper not instead of context.

Parul, et al. [9] proposes an indexing structure in which index is built on the basis of context of the document rather than on the terms basis using ontology. The ontology-based collection uses context to describe collections and search engines. The context of the documents being collected by the crawler in the repository is being extracted by the indexer using the context repository, thesaurus and ontology repository and then documents are indexed according to their respective context. This aids in improving the quality of the retrieved results 80% when the keywords of the index using context.

One potential improvement for paper recommender systems is by using ontology to find relevant paper based-on related meaning of keywords. Xuebo Song, [11] proposes techniques to measure the semantic similarity of objects in multiple domains. By utilizing the structured knowledge such as ontologies could greatly capture the semantics of text objects that has already been established, it explores the domain knowledge from the existing lexical resources and incorporate it into specific applications within different domains. Based on the evaluation, the system has an accuracy of 0.725, Recall 0.788, Precision 0.867, And F1-Score 0.825.

One study that applied ontology in paper recommender systems was proposed by Modhi, et al [3]. It uses the DNTC or Dynamic Normalized Tree of Concepts method and the 2012 version of the ACM Computing Classification System (CCS) ontology to recommend papers. This research provides an average accuracy of up to 87%. Lack of this method is not sharing the complex ontologies for mapping paper. Furthermore, Anna Rozeva. [8] approach implemented in the current work uses the rules model obtained by processing the text with a machine learning algorithm for instantiation of domain ontology. This provides for enhancing the efficiency of the representation of the analysed domain by turning the mined model into a context model. The context model being ontologically based ensures logically validated classification and logical? reasoning.

1.2 Problem Statement

SHUCHEN, et al. [1] extracted the abstract, citation paper, and co-authorship relationship using marked other authors in each conference to getting the list of terms. The terms only discusses based on textual information and not to understand the context of the paper. The output recommendations that will appear when based on textual information only words that are similar to the other words. For example: “Semantic Similarity” and “Semantic Relatedness” will not appear in the recommendation. It is a problem considering that some researchers have an interest that turns out to be in a similar context to other researchers.

This thesis, co-authorship recommendation in research publication, proposes a method to provide relevance for each interest of the author. The raw data is in the paper perspective, which will be extracted into the author perspective. The data will go through several pre-processing, such as paper created by the author, paper co-author, and citation list on the paper created or co-author paper. The data such as paper, co-author paper, and citation list on the paper already created or co-author paper would be extracted by CSO to get a list of interests of the author based on the data. CSO will expand more data because it will check the concept in ontology. For example: The concept of the relation between “Semantic Similarity” and “Semantic Relatedness” is “Semantic Similarity” is equivalent with “Semantic Relatedness. The output of pre-processing will be trained in unsupervised learning. The data could not work in supervised learning cause there is no baseline for justify label or class of the data. The method chosen for unsupervised learning uses K-means. The previous research is tested using this unsupervised learning and compare it with our approach.

This thesis proposes to consider context for co-authorship recommendations using an ontology approach based on user preferences which the list of terms is getting from the abstract, the citation of abstract, and the title of papers. The terms would be expanded and make weighted using ontology, when the terms child or the terms equivalent or parents one level of the terms, it indicates have same context then the output recommendations will appear.

Could an ontology will improve the silhouette score and expand the recommendation?

1.3 Objectives

The objective of this thesis is to:

1. Generate co-authorship recommendation using ontology.
2. Integrate the methods with Ontology CSO that would expand more authors interest.

3. Measure the silhouette score of the proposed method and comparing with the previous method.

1.4 Hypothesis

Previous research has attempted to use all of the paper attribute combined with bookmarking conference attendees as paper recommendation [1]. The experimental using classifier result shows the precision 35%, Recall 25% and F1 Score 29%. One of the previous research method applies enhance Content-based algorithm. The algorithm extracts abstract, list of citation and co-authorship relationship to obtain the terms. The paper terms would be compared to another papers terms. When the paper terms are similar with other paper terms then the paper would be the paper recommendation and set the weighted for the paper recommendation. The previous research has not been handled the context of the text and the previous research is paper recommendation and this study focus is co-authorship recommendation. To handle the context of the text using ontology. The ontology used is CSO. Authors interest would be expanded using ontology then it would checked by the concept of ontology. The output would be different because clustering approach is used and there is no baseline to justify such as true or false. The evaluation measurement used is Silhouette Score. It is hypothesized that considering co-authorship recommendation improves the expand of the recommendation and improves Silhouette Score.

1.5 Research Methodology

The research methodology applied in this thesis consists of the following steps.

1. Problem Identification

   This step aims to identify problems in coauthorship recommendation, define a problem addressed in this thesis, and make potential improvements for the chosen problem.

2. Model Design

   Model design defines the functional requirement and design implementation that describes how a system is formed. It could be defined as a depiction and making arrangement of several separate elements including ontology as new method to expand the recommendation. Model design proposes problem solving logic.

3. Data Collection and Processing

   Data are collected from aminer website. The collected data then run in to data conversion to expected format dataset.

4. Implementation

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

5. Experiment

   It aims to prove the hypothesis in section 1.4 that considering whether the ontology improves and expands the recommendation.

6. Analysis of experiment results

   This section analyzes the SIL score gained by the proposed method and compares it with the previous research.

1.6 Scope and Delimitation

The problem scope of this research are:

1. This research focuses on co-authorship recommendation based on research and publication.
2. The dataset does not have available keywords or list of terms.
3. This research using third party tool named CSO for getting terms list from the papers content.
4. This research focuses on computer science domain.

1.7 Contribution

The main contribution of this study is an improved method for the recommender system. Recent studies have focused on their methods by combining abstract, list of citation and coauthorship relationship. This study proposes a combination of the same property as recent studies and ontology. This thesis applies SIL score because there is no baseline to set the classifier. Using SIL score makes people bring together researchers to collaborate on computer science domain interests from these researchers and easily compare the difference between methods in the future.