DETERMINING PROMINENT RESEARCH AREA AND EXPERTISE FROM SCHOLARLY DATA USING SPHERICAL K-MEANS ALGORITHM AND SCHOLAR RANKING MODEL

WENDY SARASJATI

MASTER OF COMPUTER SCIENCE
(SOFTWARE ENGINEERING AND INTELLIGENCE)

2017

Faculty of Information and Communication Technology

DETERMINING PROMINENT RESEARCH AREA AND EXPERTISE FROM SCHOLARLY DATA USING SPHERICAL K-MEANS ALGORITHM AND SCHOLAR RANKING MODEL

Wendy Sarasjati

Master of Computer Science in Software Engineering and Intelligence

2017

DETERMINING PROMINENT RESEARCH AREA AND EXPERTISE FROM SCHOLARLY DATA USING SPHERICAL K-MEANS ALGORITHM AND SCHOLAR RANKING MODEL

WENDY SARASJATI

A thesis submitted
in fulfilment of the requirements for the degree Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I declare that this thesis entitled “Determining Prominent Research Area and Expertise from Scholarly Data Using Spherical K-Means Algorithm and Scholar Ranking Model” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name : .....

Date : .....

APPROVAL

I hereby declare that I have read this dissertation / report and in my opinion this dissertation / report is sufficient in terms of scope and quality as a partial fulfilment of Master of Computer Science (Software Engineering and Intelligence).

Signature : .....

Supervisor Name : .....

Date : .....

DEDICATION

This study is dedicated to my beloved parents, my little brother, and also my friends who are always supporting me to accomplish this study.

To my supervisor, Dr. Yogan Jaya Kumar who is always gives the support, suggestion, and motivation.

ABSTRACT

There are lots of information given through the website or online media nowadays. These include data of research publication such as the data available on scholarly data e.g. Google Scholar. Determine the prominent research area and finding the key players is the motivation of this study. Despite many people may know about the published articles of certain researchers, however there are no information on the research areas of an institute or university where the researchers belong to. Thus, this study will investigate how the prominent research area can be determined by using Spherical K-Means algorithm to cluster the topics. Likewise the proposed expert search approach could determine the key players who are the experts in certain research area. In order to identify the experts, this study identifies the prominent of research area using novel ranking measure i.e. scholar ranking model. This study applies top-down approach in order to solve the problem. This top-down approach initially represents UTeM research areas then followed by the prominent research study. Thus, based on this prominent research study then it comes up with the experts which have related to each field. This study achieves the first objective using spherical K-Means that is determine the prominent topics in UTeM such as Advanced Computing Technology, Telecommunication Research and Communication, Advanced Manufacturing Technology, and Robotic Industrial Automation. Besides, this study also completes the second objective which is identifying the experts based on the prominent research study. The outcome to achieve the second objective is ranking each candidate expert based on the citation by using scholar ranking model.

ABSTRAK

Terdapat banyak maklumat yang terdapat di laman web atau media dalam talian pada zaman kini. Ini adalah termasuk data penerbitan penyelidikan seperti data yang boleh didapati di data penerbitan seperti Google Scholar. Menentukan bidang penyelidikan terkemuka dan mencari pemain utama ialah motivasi kajian ini. Walaupun ramai orang mungkin mengetahui tentang cara menerbitkan artikel penyelidikan yang tertentu, bagaimanapun tiada maklumat tentang bidang penyelidikan sesebuah institut atau universiti di mana ia adalah kepunyaan penyelidik. Maka, kajian ini akan menyelidik bagaimana bidang penyelidikan terkemuka boleh ditentukan dengan menggunakan algoritma Spherical K-Means dalam mengumpulkan sesebuah topik. Begitu juga cadangan pendekatan pencarian pakar boleh digunakan bagi menentukan penyelidik utama yang merupakan pakar dalam bidang penyelidikan tertentu. Untuk mengenal pasti kepakaran seorang penyelidik, kajian ini mengenal pasti bidang penyelidikan menggunakan ukuran kedudukan yang terbaru iaitu meletakkan model kesarjanaan. Kajian ini berkaitan pendekatan dari top-down supaya dapat menyelesaikan masalah ini. Pendekatan dari top-down ini pada mulanya mewakili bidang penyelidikan UTeM kemudian diikuti oleh kajian penyelidikan terkemuka. Maka, berdasarkan penyelidikan kajian terkemuka ini kemudian ia dinaik taraf dengan pakar yang mana berkaitan dengan setiap bidang. Kajian ini mencapai objektif pertama menggunakan K-Means bulat yang menentukan topik-topik terkemuka dalam UTeM seperti Advanced Computing Technology, Telecommunication Research and Communication, Advanced Manufacturing Technology, dan Robotic Industrial Automation. Selain itu, kajian ini juga mencapai matlamat kedua yang mana sedang mengenal pasti pakar berdasarkan penyelidikan kajian terkemuka. Keputusan untuk mencapai objektif yang kedua itu meletakkan kedudukan setiap pakar calon berdasarkan petikan dengan menggunakan ranking model sarjana.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. Yogan Jaya Kumar of the Faculty of Information and Communication Technology and at Universiti Teknikal Malaysia Melaka (UTeM) who gives the support, suggestion, and motivation about my research or writing.

I express my very profound gratitude to my parents and my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.0 Introduction

Publication in academic research is growing rapidly and it has various research areas. These areas can be obtained by research knowledge enhancement that mostly comes up from the university. This situation brings different major of research area in the university. Meanwhile, a university should have at least one prominent research area. The way to find out the prominent area in the university is observing the research publication through online scholarly literature. Despite people may know about the certain published research, it does not specifically focus on research area of university. In order to analyse this, there is a way for grouping the prominent in publication research by using clustering which is applied in research topic area.

A prominent in publication research which is done by university can be determined from the most popular topics that published in online scholarly literature e.g. Google Scholar. Thus, in order to discover the most frequently performed in university which have some various major of fields this study uses clustering process. It can identify the features which is prominent by processing the collection of object that has similarities with each other in several groups (Ramkumar, 2016). The process does not need information about the collection of the object to be categorized, hence, clustering process can be defined as unsupervised. There are some studies which concern toward clustering process that produce different kind of techniques for grouping objects in the same categorize.

Analysis of clustering is one of the data mining techniques which are being used in some areas such as knowledge discovery, pattern recognition, and so on (Jindal & Kharb, 2013). The aim of clustering is to find out the similarity of huge dataset which is transformed into different groups in a particular way (Zhenpeng et al., 2014). There are two kinds of data which used; structured data and unstructured data. Since finding a prominent research area is categorized as unstructured data because it has no well-defined arrangement of text document. In this case, a simple data mining cannot performed in unstructured data, then there are some key tasks (Musembi Kwale, 2013); document representation, definition of similarity measure, and clustering logic.

In summary, document representation is a process to convert unstructured data into structured data, so that simple or traditional data mining clustering can be applied. The use of document representation in the simplest way is Boolean model which is in every document will be simplified as a bag of words. The implementation of this model is Vector Space Model (VSM) which belong to distance based algorithm (Rehurek, 2011). Definition of similarity measure is discovering the structured data are related to another, thus there are some data which go to the identical cluster. Cosine measure can be applied to find the relations between data are implied to cosine of the angle between a pair of vectors data. The last is clustering logic or algorithm, clustering process using algorithm i.e. K-Means which is a popular algorithm that use iterative computation to discover a cluster toward dataset. According to (Zhang & Li, 2011), K-Means is a suitable algorithm to determine topic detection with a large scale data, the research proved that performance increased 38.378% from large scale corpus than a small one. In addition, K-Means algorithm has the advantages there are theory reliability, suppleness, fast convergence speed, small time complexity, and so on (Zhenpeng et al., 2014).

Based on the explanation of clustering process by using some key tasks, we can do the implementation toward finding the prominent publication research by grouping the related data in similar topic and select the famous category between those groups. The aim of text data mining is finding unknown information which is no one knows and written yet (Gupta et al., 2009). Thus, by processing the data, we can get the information which people did not know yet about the prominent research area in a university.

K-Means algorithm typically is a popular clustering method toward a set of data vectors i.e. Spherical K-Means algorithm, the form of this algorithm uses a Euclidean distance. This algorithm is simple to apply and understand, by measuring the Euclidean and primary data using matrices. Moreover, this algorithm has the effectiveness on memory requirements usage, it only requires storing data points and the distances between data points and centroids (Musembi Kwale, 2013). In addition, there is another proposed algorithm to be used in this study which is Spherical K-Means that is an enhancement from the traditional K-Means algorithm. Meanwhile Spherical K-Means implements cosine similarity rather than using Euclidean distance to determine the similarity distance.

Both of the algorithms are including as unsupervised learning which is need to be validated the quality after clustering process. The other way, there will face any difficulties to use different clustering results. Hence, this study prefers to use an internal clustering validation. The main reason of using this technique because the internal validation is no need to calculate the entropy which is evaluating based on the given class label and must be understand the right cluster in advance, it sometimes used to choose the optimal cluster on a particular dataset. Regarding to (Zhao, 2012) toward this study, the information of data set sometimes is not available at in beginning, and the internal validation is more profitable to be used. Meanwhile, internal validation only calculate toward information of

dataset without any additional information (Liu et al., 2010). Besides, the external validation needs prior information toward data set i.e. the optimal number of cluster, this information used for selecting the best clustering method in particular data set. In conclusion, the internal validation has several methods, this study intents to apply Silhouette method to validate the K-means and Spherical K-Means clustering algorithm.

Once we discover about the research area that has been held in the university, then we can determine who are the key players that involve in the prominent publication research. There are some methods to discover the key players one of them is expert search, it can be used for identification topic interest from the research who has a relevant expertise (Macdonald & Ounis, 2006). Expert search system shows to the users about people's expertise: first is to denote the topic interest toward system, users do a formulation for the query; then, available documentary evidence is used for ranking a candidate person by referring to their expertise toward the query that has been predicted. System especially use a profile of evidence for each candidate which point out their expertise.

In the following years, some studies are proposed to do an expert search, based on the research, ranking of candidates is become the most effective approaches used (Macdonald & Ounis, 2009). Thus, the method used in this study is scholar ranking model to determine the experts who are doing research with the related expertise in research topic. Refer to (Wu et al., 2011), this method produces: (1) A recommendation of well-known and important scholar searching in web mining approach; (2) Flexible ranking function development in research field toward scholar ranking; (3) The researcher was complete the construction and demonstration of scholar ranking model toward scholar searching mechanism. This method establish a ranking of the expert system by doing a

computation of query in a scholarly literature (i.e. Google Scholar) based on the title of publications, researcher's name, and citations.

There are some results which followed by this chapter, such as: section 1.2 explains about the problem background of this research; section 1.3 declares about problem statement; section 1.4 maintains the objectives of research; section 1.5 describes the scope of the research; section 1.6 explains the significance of research; section 1.7 contains the expected contribution of research.

1.1 Problem Background

There are lots of information given through website or online media nowadays. Including the data of publication research that have been done in UTeM which become our concern study. Most of the lecturers are publishing their research there because Google Scholar is a web search engine which have full of article index from scholarly literature such as online academic journals and books, conference papers, technical reports, and it find more cited references than another bibliographic database which means that Google Scholar has more indexing journal that exceeded from another database.

On the other hand, to define the prominent research area which conducts in the university by doing observation through the topic one by one at Google Scholar is very difficult. It provides some data from the university which has some names of researcher with the title of publication research, citation per year, and year of publication. However, reader may know about the publication research of certain researchers in their university but not the focus research area toward the university. Pointed to the statement in section 1.0, there is a need to determine a prominent publication research area which has been held in UTeM.

Since UTeM has several focuses on technology research named Center of Excellences (CoEs) shown in Figure 1.1, there are Advanced Manufacturing Center (AMC), Center for Telecommunication Research and Innovation (CeTRI), Center for Advanced Computing Technology (C-ACT), Center for Robotic Industrial Automation (CeRIA) and Center of Advanced Research on Energy (CARE) (UTeM, 2014). UTeM has been defined their critical focused area which is Advanced Manufacturing Technology (AMT). AMT has some thrust area which are Green Technology, Systems Engineering, Human - Technology Interaction, Emerging Technology (UTeM, 2015). Therefore, this research attempts to investigate whether UTeM's research publications are within these prominent research areas or not.

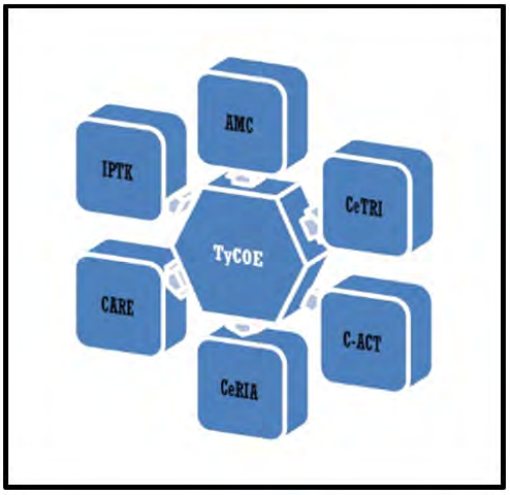

Figure 1.1 : UTeM's Center of Excellences (UTeM, 2014)

A purpose of this research is to determine a domain or prominent of publication research area which is being conducted in UTeM, this research will implementing an information extraction technique in order to cluster the research based on the topic. Clustering technique that purposed in this study is K-Means and Spherical K-Means algorithm. These algorithms include as an Information Extraction which is the task to

discover structured information from unstructured or semi-structured form (Fader et al., 2011).

In this study, we use Google scholar as a data resource because it is a bibliographic database which is open source that presents data where everyone can access the information about the academic research or journal, besides that, people can also get information about the name of researcher, institution, citation index, and etc. That is the reason why this research uses Google Scholar as a resource of data.

An expert search to be used for identification topic interest from the research who has a relevant expertise (Macdonald & Ounis, 2009). This research will use a novel topic model which is scholar ranking to determine key player of experts based on the relevant expertise research topic. Data to be used in this process is making up from the result of clustering which has been done in a previous section. This technique is suitable to be used in this research because the components to find the expert search are query topics that obtained from clustering process, the author name or the experts which has been extracted in crawling process, and citations for each publication research were analysed in crawling process (Wu et al., 2011).

The implementation of expert search toward to this research can show the key players who are the experts of certain topic in research area (for example machine learning) in Universiti Teknikal Malaysia Melaka (UTeM).

1.2 Problem Statement

The prominent research areas in some of universities are different with other, each of them is doing in different major and field, and it also will be different prominent area in

the range of time. Data of publication research which is done in Universiti Teknikal Malaysia Melaka (UTeM) that available on Google Scholar. Most of people are using Google scholar for both of the researcher and the reader because it is a web search engine which have full of article index from scholarly literature such as online academic journals and books, conference papers, technical reports, and it find more cited references than another bibliographic database which means that Google Scholar has more indexing journal that exceeded from another database. However, people may know about the published research of certain researchers in UTeM but not specifically focus on research area of university.

Difficulty to discover the research area requires an inventive solution. Hence, this research will give the information about prominent area of research and also the experts that belongs to. This study implements K-Means and Spherical K-Means algorithm for clustering toward corpus from Google Scholar in order to define the most prominent research area that is being conducted by UTeM and visualize a time frame to point out a research area. Implementation of expert search toward to this research is showing the key players who are the experts of certain topic in research area in UTeM. By applying a novel ranking measure which is scholar ranking for expert ranking, finding the research topic of academic and finding different type of groups in research that will be done using cluster which implemented in experts through the same of expertise and interest. The implementation of expert search toward to this research can show the key players who the experts of certain topic are in research area (for example data mining) in UteM.

There are some research questions that will be examined in this research:

1. How to determine the most prominent research area in Universiti Teknikal Malaysia Melaka (UTeM) using Spherical K-Means algorithm?
2. How to find the expert in related expertise of research topic based on the prominent of research area in UTeM?
3. How to determine ranking of the expert toward prominent research topic using scolar ranking model?

1.3 Objectives of Research

The aim of this research is to determine a domain or prominent of publication research area which is being conducted in UTeM, this research will implementing an K-Means algorithm in order to cluster the research based on topic and find the expert of in related expertise of research topic based on the prominent of research area using scholar ranking based for expert ranking. In order to achieve the aim that has been mentioned directly above, the following objectives are listed below:

1. To determine the most prominent research area in Universiti Teknikal Malaysia Melaka (UTeM) by clustering research topic using Spherical K-Means algorithm.
2. To identify the expert in related expertise of research topic based on the prominent of research area using novel ranking measure i.e. scholar ranking model.

1.4 Scope of Research

This research has some limitations in order to perform clustering the prominent research area in Universiti Teknikal Malaysia Melaka (UTeM) using K-Means algorithm