A Framework to Determine Prominent Research Topics and Experts from Google Scholar

W. Sarasjati[1,2], Y. J. Kumar[2], O. S. Goh[2] and B. Raza[3]
[1]*Universitas Dian Nuswantoro, Jl. Nakula I No. 1-5, 50131, Semarang, Indonesia.*
.[2]*Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia.*
[3]*Department of Computer Science, COMSATS Institute of Information Technology, Islamabad, Pakistan*.
yogan@utem.edu.my

***Abstract*— In today's digital era, most scholarly publications are made available online. These include the data of a university's research publications which can be reached through Google Scholar. Determining the prominent research areas of a university and finding its experts is the motivation of this study. Although many people may be aware of the published articles of certain university researchers, however, there is little or no information on the main research areas of the university where the researchers belong to. Thus, this study will investigate how the prominent research areas can be determined by implementing Refined Text Clustering (RTC) technique for clustering scholarly data based on the titles of publications. Then, an expert search approach can be used to determine the key players who are the experts in each research cluster. The Expert Finding System (EFS) is proposed by applying statistical analysis based on the total of number researcher's publications and their number of citations.**

***Index Terms*—Clustering; EFS; K-means Algorithm; RCT; Text Mining.**

I. INTRODUCTION

Publication in academic research is growing rapidly and it has various research areas or domains. These areas can be obtained by research knowledge enhancement that mostly comes up from university. This situation brings different domain 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 through observing their research publications through online scholarly literature e.g. Google Scholar. Although people may be aware of certain published research, it does not specifically focus on research area of university. In order to discover prominent research area in a university, this study proposes a framework that applies clustering technique.

Universiti Teknikal Malaysia Melaka (UTeM) will be used as a case study to find the most popular research topics that have been published in Google Scholar. UTeM has several focuses on technology research known as Center of Excellences (CoEs), which includes 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 has been defined their critical focused area which is Advanced Manufacturing Technology (AMT). AMT has some thrust areas which are Green Technology, Systems Engineering, Human - Technology Interaction, Emerging Technology [1]. Therefore, this study can be used to ensure the validity of prominent research area using the proposed method.

This study proposes K-Means algorithm that use an iterative computation to discover cluster toward dataset. According to [2], K-Means is a suitable algorithm to determine topic detection for a large scale data and the research proved that performance increased 38.378% for large scale corpus when compared to a small corpus. For this study, we will implement Refined Text Clustering (RTC) algorithm based on Spherical K-Means in order to improve the performance of the algorithm [3].

Once the prominent areas have been discovered then we can determine the key players that are involved in those prominent research areas. There are some methods to discover the key players and one of them is expert search. It can be used for the identification of topic experts from the researchers who has a relevant expertise [4]. Expert search system shows the users about people's expertise: first is to denote the topic experts, 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. The system uses a profile of evidence for each candidate which point out their expertise.

In the following years, some studies were proposed for expert search and based on those research, ranking of candidates become the most effective approach used [5]. Thus, the proposed framework to identify the candidates who are the experts in a research topic is determined based on Expert Finding System (EFS) analysis.

II. RELATED LITERATURE

Clustering technique has been applied in some cases especially for topic mining. Referring to the work in [2], the authors proposed K-Means algorithm in order to detect some topics in corpus which consist of news using Chinese-language in CAS Institute of Computing. In [6], the author creates a cluster networks for related papers and categorize papers to groups on the same topic using Latent Dirichlet Allocation (LDA). In [7], the authors apply Vector-Space model to represent the documents and Newman fast clustering algorithm has been used to discover the cluster for each year that is indicated by yearly research fields. The results of cluster documents shows the entire status of research field per year and the analysis of connection of similarity among clusters give a stable research fields in clear representation.

An expert search which is also called expert finding is used in designing the relation among candidates and their expertise [8]. There are some other works such as [5], where an expert search identifies topic experts from the documents that are associated by implementing Voting Model for ranking those documents. According to [9], their research apply Citation Author Topic (CAT) in order to extract topic in academic research and determine various research expert by clustering experts in similar expertise and interests.

In this study, we propose RTC algorithm that is an enhancement of Spherical K-Means algorithm to be implemented in clustering process and then use EFS technique that will be applied in expert search.

III. Operational Framework

Several stages that should be done are mentioned in Figure 1 which depicts our research operational framework.

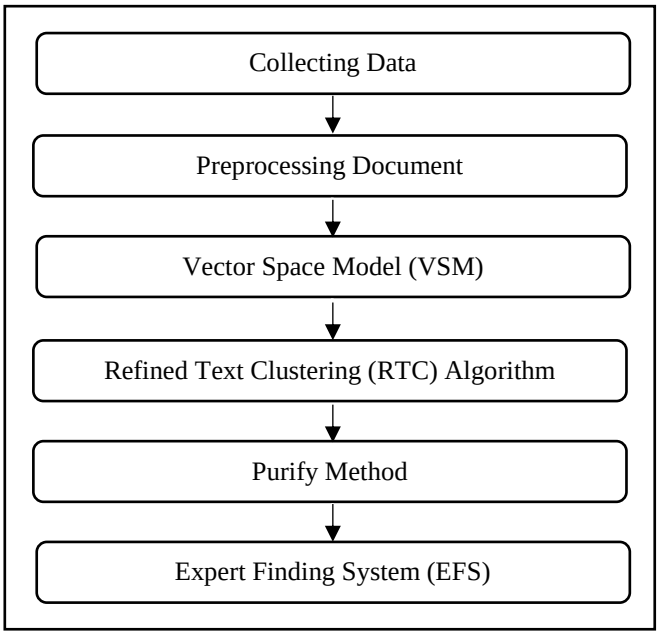

Figure 1: Research Operational Framework

A. Collecting Data

In this study, the data collection is obtained from Google Scholar by using crawling technique. The data content that were retrieved are the title of research, the researchers' names, number of citation, and their h-index.

B. Preprocessing Document

After the crawling process is done, the next phase is the preprocessing of document; that has several steps which is tokenization, stemming, and stopword removing.

1) Tokenization

In order to analyse the data, before entering the cluster process, data which is going to be used should have minimal noise. Tokenizing is the first step in preprocessing, which is a process to cut strings and convert into a basic unit word arrangements.

IV. Stemming

The second step in preprocessing is stemming, which is a method to reduce the amount of features called as spacing [10]. According to [11], the general idea of stemming is a process where users find information using one keyword as a query, for example the query word retrieval, but might contain some information which have the relation with other query words such as retrieving, retrieved, retriever, etc. Thus, in order to decrease the unique terms in words, this process mainly use the root of words.

V. Stopword Removing

The third step in preprocessing is stopword removing process. Based on the output of tokenizing, it determines some irrelevant words that will be removed. The features that have conjunctions and special characters will be also removed as it does not have an important meaning, for instance "a","in","the", "of", etc. [11]. From the list of stopword, each token (from tokenization) will be compared with the stopword list. If there is a token matched in stopword list then it should be removed until there is no more token available in the list.

A. Vector Space Model (VSM)

Documents are usually in textual form and should be represented as a mathematical form in order to be applied in cluster algorithm [12]. In this phase VSM is used to convert document from text into vector form. The technique works where the data have different weights for different words that are obtained for each document. In this model, the document will be represented as term document matrix. In order to determine the weights for each word in each document, Term Frequency (TF) technique will be implemented to find out the frequency of a term in the documents. These produce the calculation of the significance of each term in the document.

B. Refined Text Clustering (RTC) Algorithm

In this study RTC algorithm is proposed to cluster and determine the popular research topics in UTeM. RTC algorithm is an enhancement technique from spherical K-Means algorithm which is developed by [3]. Spherical K-Means algorithm is also an improvement of the traditional K-Means algorithm and implements cosine similarity [13]. RTC phases are explained next.

1) Refine the Initial Centres Choice

An initial centre choice is generally based on the centre election for each cluster randomly by using the furthest distance of each cluster [14]. Following are the principles of this algorithm [3]:

a. To improve the result in clustering process by producing un-similar initial centres, the value of cosine distance towards pair of initial centres should be in minor range.
b. To prevent selecting outliers which are obtained from an initial cluster, it should be dense.

There is no intersection between pair of initial clusters and eps nearest neighbors in certain condition; this situation will prevent different initial centres to belong in same cluster.

VI. Refine the Partition Adjustment

The best adjustment among border using {Pi'}k i=1 symbol is an adjustment which has the maximum function of quality value toward all of the adjustments. The following algorithm

in Figure 2 depicts the methodology of RTC algorithm in clustering process.

Algorithm : Refine of Text Clustering Algorithm
Input : Vector set *X* of text, number of cluster *k*, *MinPts* as a threshold of density, *iteration* as a halt of situation.
Output : Partition of cluster $\{P_i\}^k{}_{i=1}$
Steps :

1 Initialize $\{P_i^0\}^k{}_{i=1}$ as the partition, then use the algorithm of initial centre choice for determining *k* initial center c_i ($1 \le i \le k$) which has a relation with partitioning. Index of iteration should be define as j=0. Then calculate

$$Q(\{p_i^0\}_{i=1}^k) = \sum_{i=1}^{k} \sum_{x \in pi} \cos ine(x, c_i)$$

2 The process will begin when (j < *iteration*)

3 Determine the most similar centre for each text and arrange the *k* cluster p_i ($1 \le i \le k$) into $\{P_i^{j+1}\}^k{}_{i=1}$ as a partition

4 Using the formula to recondition the center of each cluster, which is

$$c_i = \frac{\sum_{xi \in pi} x_i}{|p_i|}$$

5 Measure the function of quality value by using this formula

$$Q(\{p_i^{j+1}\}_{i=1}^k) = \sum_{i=1}^{k} \sum_{x \in pi} \cos ine(x, c_i)$$

6 Go to the second step if the condition $Q(\{p_i^{j+1}\}_{i=1}^k) - Q(\{p_i^j\}_{i=1}^k) \ge tol$ then j=j+1; else if $Q(Adjustment(\{p_i^{j+1}\}_{i=1}^k), t\_fun, j) - Q(\{p_i^j\}_{i=1}^k) \ge to$ then go to the second step; else do halt algorithm;

7 End process.

Figure 2: Refine Text Clustering Algorithm

A. *Purify Method*

This phase performs an evaluation of the result by testing the performance of RTC clustering algorithm in percentage of accuracy. Thus, we present Purify as the method to evaluate the clustering quality. If clustering performance is bad then the Purify value is close to 0 (zero) and vice versa. Following is the formula of Purify method that will be implemented in this study:

$$Purify = \frac{1}{N} \sum MAX\ (€ \cap £) \qquad (1)$$

Where: N = The total of member clusters (1,2, ... , k)
€ = The total cluster
£ = The total class

B. *Expert Finding System (EFS)*

This technique applies statistical analysis based on the total of number researcher's publications and their number of citations. The ranking system is used in this stage to represent the top 10 experts based on the popular research domain in UTeM.

C. *Estimate the Candidate Ranking*

To estimate the rank of candidates (in this case, the researchers), this study will measure the ranking to estimate the impact of individual candidate, called p-index, that contains the total citations of papers which are published by its candidate and Θ as the total citation of UTeM's publications in certain research clusters or domain. The formula to find the value of p-index as follows [15]:

$$C_i^{p-index} = \frac{\sum citations \in C_i}{\sum citations \in \theta} * 100 \qquad (2)$$

where: Ci = candidate i
Θ = set of published literatures.

VII. Conclusion And Future Work

Determining the prominent research areas of a university and finding its experts is the motivation of this study. Although many people may be aware of the published articles of certain university researchers, however there are little or no information on the main research areas of the university where the researchers belong to. This paper provides a framework on how the prominent research areas can be determined by implementing Refined Text Clustering (RTC) technique for clustering scholarly data based on the titles of publications. Based on the research clusters, expert search approach can be used to determine the key players who are the experts in each research cluster or domain. In our ongoing work, we aim to test our proposed framework on UTeM publication data. In addition, we plan to investigate the findings based of different time frame to analyse the research direction in UTeM. This will make our findings more conclusive as it can give some insights whether UTeM's research direction is in line with its thrust areas.

Acknowledgment

The authors gratefully acknowledge the support of Universiti Teknikal Malaysia Melaka (UTeM) under the University Research Grant Scheme (PJP) No. PJP/2015/FTMK(5B)/S01438 and Universitas Dian Nuswantoro and Ministry of Education, Indonesia in sponsoring research authors.

References

[1] UTeM, "Excellence in focused research and innovation," 2015. [Online]. Available: http://www.utem.edu.my/portal/fast-facts.html.
[2] D. Zhang and S. Li, "Topic detection based on K-means," 2011 Int. Conf. Electron. Commun. Control, pp. 2983–2985, Sep. 2011.
[3] Y. Liu, J. Cai, J. Yin, and Z. Huang, "An Efficient Clustering Algorithm for Small Text Documents," 2006.
[4] C. Macdonald and I. Ounis, "Voting for Candidates : Adapting Data Fusion Techniques for an Expert Search Task," no. November, pp. 387–396, 2006.
[5] C. Macdonald and I. Ounis, "On perfect document rankings for expert search," Proc. 32nd Int. ACM …, pp. 3–4, 2009.
[6] R. Nakazawa, T. Itoh, and T. Saito, "A Visualization of Research Papers Based on the Topics and Citation Network," 2015 19th Int. Conf. Inf. Vis., pp. 283–289, Jul. 2015.
[7] Z. Xuan, L. Chen, Y. Dang, and J. Yu, "Research Field Discovery Based on Text Clustering," 2008 4th Int. Conf. Wirel. Commun. Netw. Mob. Comput., pp. 1–4, Oct. 2008.
[8] Y. Fang, L. Si, and A. P. Mathur, "Discriminative models of integrating document evidence and document-candidate associations for expert search," in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10, 2010, p. 683.
[9] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier, "Citation Author Topic Model in Expert Search," no. August, pp. 1265–1273, 2010.
[10] A. I. Kadhim, Y.-N. Cheah, and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," 2014 4th Int. Conf. Artif. Intell. with Appl. Eng. Technol., pp. 69–73, Dec. 2014.
[11] A. Singhal, "Modern Information Retrieval : A Brief Overview," IEEE Data Eng. Bull., vol. 24, pp. 35–43, 2001.
[12] Latika, "International Journal of Advanced Research in Computer

Science and Software Engineering An Effective and Efficient Algorithm for Document Clustering," vol. 5, no. 5, pp. 449–455, 2015.
[13] R. Duwairi and M. Abu-Rahmeh, "A novel approach for initializing the spherical K-means clustering algorithm," Simul. Model. Pract. Theory, vol. 54, pp. 49–63, May 2015.
[14] S. Sujatha and A. S. Sona, "New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method," vol. 2, no. 2, pp. 1–9, 2013.
[15] C. Wu, J. Chung, C. Lu, H. Lee, and J. Ho, "Using Web-Mining for Academic Measurement and Scholar Recommendation in Expert Finding System," 2011.