Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 124 (2017) 29–37

Procedia Computer Science

www.elsevier.com/locate/procedia

4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia

Community Detection On Citation Network Of DBLP Data Sample Set Using LinkRank Algorithm

Satrio Baskoro Yudhoatmojo*, Muhammad Arvin Samuar

Faculty of Computer Science, Universitas Indonesia, Kampus UI, Depok, 16411, Jawa Barat, Indonesia

Abstract

This paper describes the application of a community detection algorithm, namely LinkRank algorithm, on a citation network. Community detection is a task in network analysis which aims to find sets of tightly connected nodes that are loosely connected with other nodes outside of those sets. In our study, we focused on a citation network which depicts relationships between cited papers and the papers which cite those papers. The objectives of our study are to identify communities of papers based on the citation relationships and analyze the similarities of topics within each community. The approach of our study to reach the objectives is by applying LinkRank algorithm to a citation network. LinkRank algorithm is chosen because it can be applied to a directed network where other algorithms that we have surveyed can only be used on undirected network. The citation network that we used in our study is from Aminer website. In applying the algorithm, we had to port the original source code which is written in C programming language into Python programming language for our convenience in doing the experiment. The result shows that the algorithm able to detect 10,442 communities from 188,514 nodes. Once the communities have been detected, we sampled top three communities (the ones with the largest number of members) and took the top 10 nodes with the highest PageRank score in each of those communities. The samples show that most of the nodes have similar topic, but there are still some nodes with different topics mixed inside the same community. We found the ratio between nodes with similar and different topics to be 7 to 3, that is 70% of the nodes have similar topic while the other 30% have different topics. Thus, the homophily of each community does not reach 100%. Nevertheless, our study confirms that LinkRank algorithm can be used for community detection on directed network.

Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017.

Keywords: Citation Network; Community Detection; Complex Network; Directed Network; LinkRank Algorithm; Social Network Analysis

1. Introduction

A task of citing earlier research publications is a common task and it is a way of acknowledging previous achievement made by other researchers that may become foundations of new researches in the future for discovering new knowledges. Analysis on the citation relationships could reveal several interesting things, one of them is detecting

* Corresponding author. Tel.: +62-21-7863419 ; fax: +62-21-7863415.
E-mail address: satrio.baskoro@cs.ui.ac.id

1877-0509
Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017
10.1016/j.procs.2017.12.126

communities of citations. This detection can show how research publications are related to previous research publications which may due to similar research topic or known as topic discovery according to [1]. In Complex Network, researchers take the citations of research publications and model them into networks of citations. This is commonly known as *Citation Network*. Citation network is a graph where the nodes/vertices are composed of a set of documents and the links/edges are the citation relationships of documents to other documents [2]. In a citation network, the links between the cited publications and the publication that cites them are directed and not reciprocals. Directed means a publication cites older publications and an older publication cannot cite a newer publication. Networks with directed links, such as citation network, are called directed networks.

For computer science publications, there is a web application that serves as an open bibliographic information on major computer science journals and proceedings. This web application is called DBLP and the URL is: http://dblp.uni-trier.de. DBLP is a joint service provided by University of Trier and Schloss Dagstuhl [3]. A web site called *Aminer* has kindly pre-processed the DBLP data and format them into a citation network. The resulted data is available at this URL: https://aminer.org/citation [4]. In our work, we analyzed the community structure of the citation network from this data set using a community detection algorithm named LinkRank. We also observed the members of the identified communities to identify whether the publications within those communities are homogeneous or heterogeneous.

This paper is organized as follows. Section 2 describes literature reviews that we have conducted to several academic conference papers, journal papers and textbooks. Section 3 provides the research methodology that we used to work on this research. Section 4 shows the result of our analysis. Section 5 concludes our work.

2. Literature review

This section describes the literature reviews of foundation theories and related topics of our research. There are seven topics that we reviewed. These seven topics are: (a) Social Network Analysis, (b) Network-based Unsupervised Learning, (c) Community Detection, (d) Modularity, (e) PageRank Algorithm, (f) LinkRank Algorithm, and (g) Pajek Data Format. Each of the seven topics is described in the following paragraphs.

Social Network Analysis is a field that takes social relationships in the form of network of nodes and edges. The nodes represent actors and the edges represent relationships between those actors [5]. According to [6], Social Network Analysis is part of a model development, specification and testing process for conveying relationships theoretically by purveying formal definition and measurement for expressing structured output. The goal of Social Network Analysis is to find ways of analyzing network data to reveal the structure of the underlying communities that they represent [7]. According to [8], networks in general have four common properties: (a) Small-World Property, (b) Power-Law Distribution Property, (c) Network Transitivity Property, and (d) Community Structure Property. This paper is about the community structure property. The community structure property implies there are subsets of vertices in which within the subsets, the vertex-to-vertex connections are *dense* but between the subsets are less dense [8].

Unsupervised learning is process of learning patterns from a given data which has no prior knowledge about its structure and information [9]. According to [9], two classical example of unsupervised learning are clustering and dimensionality reduction. Clustering is one of the main topics in unsupervised learning and it can be called Community Detection for clustering data in which the data is in the form of network [10].

Communities are sets of nodes in a network which share common characteristics or similar properties [11]. *Community detection* is a common problem in graph data analysis in which the goal is to find sets of tightly connected nodes that are loosely connected with other nodes outside of those sets [12]. Santo Fortunato in [13] also stated that community detection's goal in graph is to identify modules and hierarchical organizations based only on the topological information of the network. Community structure is an important characteristic of networks because it is the key to learn about the complex network topology, to understand the network functions, to find hidden pattern, to do link prediction and to expand the detection [1][14]. The problem of detecting communities can be formalized as an optimization problem by defining and optimizing an appropriate criterion function that catches the intuitive concept of community [14]. According to [15], there are four categories in finding communities of a directed network. Those four categories are: (a) naive graph transformation, (b) transformations maintaining directionality, (c) extending clustering objective functions and methodologies to directed network, and (d) alternative approaches.