Future Generation Computer Systems 89 (2018) 584–605

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Analyzing scientific context of researchers and communities by using complex network and semantic technologies

Vitor Horta *, Victor Ströele, Regina Braga, José Maria N. David, Fernanda Campos

Federal University of Juiz de Fora (UFJF), Computer Science Postgraduate Program, Knowledge Engineering Research Group, Juiz de Fora – MG, Brazil

HIGHLIGHTS

- An overlapping community detection algorithm for bidirectional graphs is proposed.
- Scientific social network analyses over time are performed.
- Scientific context discovery using ontological terms and rules.
- Influential researchers, central connectors and information brokers are found.
- Semantic analysis shows that central connectors define their communities's context.

ARTICLE INFO

Article history:
Received 15 December 2017
Received in revised form 4 May 2018
Accepted 10 July 2018
Available online 25 July 2018

Keywords:
Scientific social network analysis
Semantic analysis
Overlapping community detection
Clustering algorithm
Ontology
Scientific context

ABSTRACT

Social network communities are composed of people with common interests who influence or are influenced by themselves. In the scientific context, Scientific Social Networks are characterized as social networks that represent the social relations established by researchers. Identifying and exploring these relationships are fundamental activities to support scientific experiments. In this study, we aim to discuss the use of complex networks combined with semantic analysis in a network of scientific publications called DBLP. DBLP can be classified as big data, and its use for the analysis of connections and influences among researchers can be considered a context-aware approach. Therefore, in the present study, concepts of complex network analysis are applied to verify the level of influence among researchers, by analyzing the structure of the scientific social network under study and its communities. A bidirectional graph-based model was proposed in order to evaluate the influence of researchers, in addition to algorithms to analyze the network structure and identify scientific communities, using ontological terms and rules, considering the scientific context, and identifying new connections to promote scientific collaboration. For the identification of scientific communities, we proposed an overlapping community detection algorithm, named NetSCAN. A large scientific database (DBLP) together with digital libraries were used to evaluate the model and the algorithms in a historical research experiment. The results point to the viability and effectiveness of the proposed solution.

1. Introduction

Complex networks represent a set of objects that are connected in a non-trivial way. In order to understand how objects are related and grouped, a careful analysis on the structure and characteristics of these networks is needed [1]. Objects, vertices or nodes can be used to represent various real-life scenarios, such as people through a social relationship, food chains, molecular interactions, among others. In this way, knowing the network structure in which its elements are included, it is possible to identify the main characteristics that influence their interactions.

Considering that complex networks of distinct domains often share similar behaviors [2], classifying networks based only on their topology can anticipate future links between objects, or can influence and motivate the creation of new interactions. These issues and the interdisciplinary nature of the subject have increasingly stimulated the study and development of algorithms and techniques to analyze network topology, define clusters for identification of communities, and locate influencing elements, connectors and information diffusers.

Semantic analyses can improve this topological analysis by using semantic characteristics to identify new connections that are not explicitly defined, in addition to help define new semantic contexts where these connections can be placed.

* Corresponding author.
E-mail address: vitor.horta@ice.ufjf.br (V. Horta).

https://doi.org/10.1016/j.future.2018.07.012

According to Parker and Cross [3], there are four kinds of people in a social network, as follows: (i) *central connectors*, who have large amount of relationships; (ii) *boundary spanners*, who connect different groups of people; (iii) *information brokers*, influential people who communicate across subgroups maintaining a large connected group, or connecting two groups; and (iv) *peripheral people*, who are in the border of the social network needing help to improve their connections. Analyzing all these kinds of people is important so as to characterize the social network in some way. In this vein, semantic meaning can also improve this characterization. Specific analyses can be used to identify semantic connections among people who share similar interests that are not explicitly stated. In addition, semantic meaning can be used to better characterize contexts and improve the connections among people related to these contexts.

Considering people in social networks, some studies focus on their direct connections [3,4], aiming to identify *central connectors*. However, indirect and implicit relationships are also important and should be considered in social network analysis [5,6]. In this sense, *information brokers* can influence people even if they are not directly connected. They help in the dissemination of certain kinds of information and in the connectivity throughout a network [3]. Semantic characterization can also improve this process, identifying semantic interests and helping communities to maintain collaboration. In this regard, some studies have proposed methods to extract semantic knowledge in social networks [7] and methods for detecting communities in networks based on relationships enriched with semantic context [8]. This semantic knowledge can be extracted by using keywords such as tags in forum posts [9]. However, none of these studies have used implicit knowledge to discover new connections and propose new semantic contexts.

Scientific Social Networks are characterized as specific types of complex networks that represent the social relations established between researchers. In this context, the characterization of researchers can help in understanding how they collaborate with each other, and how research communities are related. The problem is how to identify researchers who help maintain the network connectivity, disseminating information and linking research groups/subgroups, and how to discover semantic connections (research interests) between them, also including the connection of researchers based on their scientific context, even though such a connection is not explicit.

Thus, this study aims to find central connectors and research-information brokers in an attempt to identify influential researchers in scientific social network, in order to connect people that have similar research interests, even though they are not explicit. For this, complex network concepts and techniques will be used to analyze the interaction between researchers, identifying (i) influential researchers who work in two or more communities simultaneously, i.e., research-information brokers connecting two or more groups; and (ii) potential influencers in specific scientific communities, i.e., central connectors from a group who connect subgroups. Ontological terms and rules will be used to discover semantic meaning and, based on that, propose new connections between researchers, also including their research contexts. Therefore, with this proposal, it is possible to extract semantic meaning from heterogeneous distributed information sources. In this study, we are using scientific repositories, specifically DBLP, digital libraries and CSS (ACM domain taxonomy). These scientific repositories are accessed through their APIs and processed using context data, by our proposal, to deliver information for adequate communities and researchers decision-making. The proposed ontology reconciles the data from the different scientific repositories by mapping their outputs to defined terms in the domain taxonomy, described in Section 6. For this, the present study advances the research work of the NEnC (Knowledge Engineering Research Group) in scientific social network [10,11] and in the scientific ecosystem approach [12–15].

In this vein, we defined two main research questions: (i) *How to identify researchers who help maintain the network connectivity, disseminating information and linking research groups/subgroups?* (ii) *How to discover semantic connections (research interests) between researchers, also including their connections based on their scientific context, even though such connections are not explicit?*

For this purpose, the DBLP database [16] was used, considering that it is one of the databases most used in the study of large scientific social networks and because it contains scientific contributions that exceed one million researchers and more than nine million interactions between them [16], thus characterizing it as big data. Scientific social network is represented by a bidirectional graph whose vertices represent the researchers and the edges correspond to the scientific relations between them. A directed graph is used so that it is possible to define the degree of influence among researchers, allowing the identification of influential and non-influential researchers in their research areas. The ontology is used to identify research interests and, by using ontological rules, identify new interest connections and new research contexts. Therefore, we are using complex networks to analyze researchers' connections and ontologies mostly focused on the researchers' semantic context.

In addition to the modeling of a bidirectional scientific social network, which allows analyzing levels of influence among researchers, this study proposes a clustering algorithm that considers the characteristics of multidisciplinary researchers, allowing them to take part in multiple groups. Through this approach, it is possible to identify people who have more than one activity area and who participate in two or more scientific communities. Another important feature of the proposed algorithm is the subgrouping function, whose objective is to identify subgroups of research communities.

As regards the contributions of this study, we can highlight: (i) the application of clustering techniques in a large-volume database; (ii) the definition of a model for the analysis of research-information brokers aiming to identify influence among researchers based on a bidirectional graph; (iii) the detection of scientific communities and their research subgroups; (iv) the identification of multidisciplinary researchers and their different levels of influence, (v) the extraction of semantic information from the network and the use of this information to identify new connections and contexts, and (vi) the use of context-aware data, extracted from scientific heterogeneous repositories, to deliver strategic information to scientific communities and researchers.

This study is organized as follows: Section 2 presents some related work; Section 3 describes the proposed model that supports the development of this study; Section 4 analyzes the topology of the scientific social network; Section 5 describes the NetSCAN clustering algorithm developed; Section 6 presents the semantic analysis; the proposal evaluation is carried out in Section 7; and finally, Section 8 makes the final considerations.

2. Related work

Several algorithms for detecting overlapping communities in social networks have already been developed with the aim of identifying groups whose members have greater similarity among themselves and greater dissimilarity from the members of other groups [17].

A tertiary study was conducted to find the methods most commonly used to detect overlapping communities in social networks. The aim of that study was to find out secondary studies that in turn could help reveal the state of the art of a research area. Accordingly, five results were obtained [18–22]. These articles presented methods with different approaches. Some examples are