Online Social Networks and Media 7 (2018) 30–44

Contents lists available at ScienceDirect

Online Social Networks and Media

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

An evidential method for correcting noisy information in social network

Salma Ben Dhaou[a,*], Mouloud Kharoune[b], Arnaud Martin[b], Boutheina Ben Yaghlane[a]

[a] Laboratory of Operational Research of Decision and Process Control, ISG of Tunis, Tunisia
[b] Institute for Research in Computer Science and Random Systems, Rennes1, France

ARTICLE INFO

Article history:
Received 17 April 2018
Revised 18 July 2018
Accepted 22 July 2018

Keywords:
Theory of belief functions
Communities
Social networks
Attributed networks

ABSTRACT

Nowadays, social networks have become an important part of our daily lives. Hence, several researchers have been interested in the study and analysis of the interactions between the entities composing this type of networks. By modeling a social network, we can assign attributes to nodes and links based on network and community structure. These attributes which may be uncertain, imprecise or even noisy, involve obtaining a non-coherent network. In order to remedy this problem, we propose, in this paper, a method that corrects the noise in the network using the theory of belief functions.

1. Introduction

Nowadays, the use of computer technology and Internet has become essential. As a result, social networks became an important part of our daily lives. Therefore, it is interesting to study and analyze the types of relationships that exist in these networks. To do so, the study of the community structure as well as the nodes and links attributes represent main characteristics that must be taken into account to analyze these networks.

In social network analysis [1,2], the observed attributes of social actors are understood in terms of patterns or structures of ties among the units. These ties may be any existing relationship between units; for example friendship, material transactions, etc.

Currently, if we observe any social network, we will soon realize that the entities composing this network are grouped, for example, according to a center of interest, a category of age, a preference, etc.

In his work, Santo Fortunato [3] explained that communities, also called clusters or modules, represent groups of vertices which probably share common properties and/or play similar roles within the graph. He argues also that the word community itself refers to a social context. In fact, people naturally tend to form groups, within their work environment, family or friends.

In a social network, we can deal with missing or modified information. In addition, the information exchanged can be often imperfect, due to the heterogeneous nature of the sources. Therefore, it would be interesting to use a vector of values which represent the nodes and links attributes.

In the same context, many studies focus on modeling the uncertain social network. In fact, they represent an uncertain network by weighting the nodes or links with values in [0,1] to model uncertainties. Hence, it will be easier to monitor the behavior of the social network [4,5]. In addition, as shown in [6], the use of evidential attributes, from the theory of belief functions, gives better results compared to the probabilistic ones.

The theory of belief functions offers a mathematical framework for modeling uncertain and imprecise information [7]. It has been employed in different fields, such as data classification [8,9] and social network analysis [10].

Furthermore, the theory of belief functions provides a flexible way of combining information collected from different sources. In the majority of cases, this combination is followed by decision-making. It also allows conflict management.

The aim of this paper is to show that even with noise in the network, our algorithm is able to classify the nodes in their initial clusters. In the case of a large noise, the algorithm guarantees the coherence of the information of any network even when it is a network whose nodes and links attributes have been strongly modified.

In this paper, we focused on the use of a limited number of communities. In terms of scaling up, there are several strategies

* Corresponding author.
E-mail addresses: salma.bendhaou@hotmail.fr (S.B. Dhaou), mouloud.kharoune@univ-rennes1.fr (M. Kharoune), arnaud.martin@univ-rennes1.fr (A. Martin).

https://doi.org/10.1016/j.osnem.2018.07.002
2468-6964/

that can reduce complexity like the one presented in [11]. This will be the subject of future work.

This paper is structured as follows. In section 2, we remind some basic concepts of the theory of belief functions and review some community detection methods as well as some other related works. Section 3 is dedicated to our contribution. Section 4 is devoted to the experimentations and finally section 5 concludes the paper.

2. Background

In this section, we start by recalling some basis of the theory of belief functions, we use it in this paper in order to model uncertainties. Then we present some community detection methods that use both the structure and the attributes of the network.

2.1. Theory of belief functions

The theory of belief functions allows explicitly to consider the uncertainty of knowledge using mathematical tools [7,12]. It is a useful and effective way in many fields such as classification, decision making, representation of uncertain and inaccurate information, etc.

In fact, it is a suitable theory for the representation and management of imperfect knowledge. It allows to handle the uncertainty and imprecision of the data sets, to combine mass functions and make decisions.

The principle of the theory of belief functions consists on the manipulation of functions defined on subsets. However, it does not represent uncertainty using sets of probability measures. These functions are called mass functions and range from 0 to 1.

Let Ω be a finite and exhaustive set whose elements are mutually exclusive, Ω is called a frame of discernment. A mass function is a mapping

$$m : 2^{\Omega} \rightarrow [0, 1]$$

such that

$$\sum_{X \in 2^{\Omega}} m^{\Omega}(X) = 1 \text{ and } m^{\Omega}(\emptyset) = 0 \tag{1}$$

The mass $m^{\Omega}(X)$ expresses the amount of belief that is allocated to the subset X. We call X a focal element if $m^{\Omega}(X) > 0$.

A categorical mass function is a mass function with an unique focal element such that $m^{\Omega}(A) = 1$.

In this work, we used also another interesting concept which is the distance of Jousselme [13]. This distance represents the degree of similarity between bodies of evidence. It is defined by:

$$d_j(m_1^{\Omega}, m_2^{\Omega}) = \sqrt{\frac{1}{2}(m_1^{\Omega} - m_2^{\Omega})^T \mathbf{Jac}(m_1^{\Omega} - m_2^{\Omega})} \tag{2}$$

where the elements ***Jac*(*A*, *B*)** of Jaccards weighting matrix **Jac** are defined as

$$Jac(A, B) = \begin{cases} 1 \text{ if } A = B = \emptyset \\ \dfrac{|A \cap B|}{|A \cup B|}, A, B \in 2^{\Omega} \setminus \emptyset \end{cases} \tag{3}$$

We also consider the normalized conjunctive rule called the Dempster rule [14], given for two mass functions m_1^{Ω} and m_2^{Ω} for all $X \in 2^{\Omega}$, $X \neq \emptyset$ by

$$m_{\oplus}(X) = \frac{1}{1-k} \sum_{A \cap B = X} m_1^{\Omega}(A) m_2^{\Omega}(B) \tag{4}$$

where $k = \sum_{A \cap B = \emptyset} m_1^{\Omega}(A) m_2^{\Omega}(B)$ is the global conflict of the combination. The Dempster combination rule reinforces the mass values of the elements on which the sources are agree. This rule is adapted when the combined mass functions are cognitively independent. In the case of dependent mass functions, one can use the mean rule given for two mass functions m_1^{Ω} and m_2^{Ω} for all $X \in 2^{\Omega}$, $X \neq \emptyset$ by

$$m^{\Omega}(X) = \frac{1}{2}(m_1^{\Omega}(X) + m_2^{\Omega}(X)) \tag{5}$$

In order to make decision, we use the pignistic probability introduced by Smets in [15] for normal mass functions by

$$BetP(X) = \sum_{Y \in 2^{\Omega}, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} m^{\Omega}(Y) \tag{6}$$

2.2. Some community detection methods with graphs structure and attributes

In this section, we introduce some community detection methods based on graph structure and attributes.

According to [16], an attributed graph $G_a = (V_a, E_a)$ can be defined as a set of attributed vertices $V_a = \{v_1, \ldots, v_p, \ldots, v_q, \ldots, v_n\}$ and a set of attributed edges $E_a = \{\ldots, e_{pq}, \ldots\}$. The edge e_{pq} connects vertices v_p and v_q with an attributed relation.

The presented model in [17] uses both information. In fact, an unified neighborhood random walk distance measure allows to measure the closeness of vertex on an attributed augmented graph. Then, the authors use a k-Medoids clustering method to partition the network into k clusters.

A second method presented in [18] consists on a model dedicated to detect circles that combine network structure and user profile. The authors learn for each circle, its members and the circle-specific user profile similarity metric. They model the membership of a node to multiple circles in order to detect overlapping and hierarchically nested circles.

A third method presented in [19] consists on dealing with the uncertainty that occurs in the attribute values within the belief function framework in the case of clustering. In this work, the authors present a new version of decision trees with the theory of belief functions to handle the case of uncertainty present only in attribute values for both construction and classification phases.

Thus, it is important to consider both information structure and attributes in order to detect the network communities. In fact, if one source of information is missing or noisy, the other can solve the problem.

The works cited above [17,18] use only a probabilistic attributes as well as the structure of the graph to do the clustering. In our previous work [6], we show that the use of evidential attributes gives better results than the probabilistic ones in the clustering.

The works cited [17–19] are interesting, but they do not assume that network information can be noisy or perturbed. In addition, they do not consider the use of node and link attributes simultaneously to do clustering.

2.3. Other related works: homophilic behaviors in social networks

In addition of the presented community detection methods above, there are works that are related to our research such as the reconstruction of an initial network and the propagation of labels.

In [20], the authors present a new method using the theory of belief functions that aims to detect communities on graphs after the stabilization of the label propagation process. In fact, SELP permits to propagate the labels from the labeled nodes to the unlabeled ones based on a propagation rule. The proposed algorithm computes the dissimilarities between nodes based on the graph structure. The main advantage of the proposed algorithm is that it can effectively use limited supervised information to guide the process of the detection.