Neural Networks 58 (2014) 14–28

Contents lists available at ScienceDirect

Neural Networks

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

2014 Special Issue

An incremental community detection method for social tagging systems using locality-sensitive hashing☆

Zhenyu Wu*, Ming Zou

School of Computer Science and Engineering, Beihang University, Beijing, 100191, China

ARTICLE INFO

Article history:
Available online 2 June 2014

Keywords:
Community detection
Big social data
Locality-sensitive hashing

ABSTRACT

An increasing number of users interact, collaborate, and share information through social networks. Unprecedented growth in social networks is generating a significant amount of unstructured social data. From such data, distilling communities where users have common interests and tracking variations of users' interests over time are important research tracks in fields such as opinion mining, trend prediction, and personalized services. However, these tasks are extremely difficult considering the highly dynamic characteristics of the data. Existing community detection methods are time consuming, making it difficult to process data in real time. In this paper, dynamic unstructured data is modeled as a stream. Tag assignments stream clustering (TASC), an incremental scalable community detection method, is proposed based on locality-sensitive hashing. Both tags and latent interactions among users are incorporated in the method. In our experiments, the social dynamic behaviors of users are first analyzed. The proposed TASC method is then compared with state-of-the-art clustering methods such as StreamKmeans and incremental *k*-clique; results indicate that TASC can detect communities more efficiently and effectively.

1. Introduction

The rise of Web 2.0 has dramatically changed the way information is stored and accessed. Social network is changing traditional communication methods by providing users with a platform for various online social activities, such as forming connections, status updating, and annotating social media at any time and place. Users often self-organize into different communities to share similar interests and contents, such as bookmarks, web blogs, photographs, music, and videos. Social tagging has become a popular means of social network. For example, Del.icio.us supports bookmark annotating, Flickr supports picture annotating, and MovieLens supports movie annotating. During the annotation process, users (so-called taggers) can freely choose tags (words or phrases) to annotate their resources of interest. Annotation provides an efficient way for users to describe, categorize, search, discover, and navigate content. These annotations generate a large amount of social data with highly dynamic characteristics.

Unstructured "big social data" is suitable for human consumption; however, it remains almost inaccessible to machines. Therefore, big social data analysis focuses on meaningful disciplines, such as social network analysis, multimedia management, social media analytics, trend discovery, and opinion mining (Cambria, Rajagopal, Olsher, & Das, 2013). Discovering common interests is a fundamental objective in big social data analysis because it can help build user communities of shared interests, identify domain experts in different subjects, determine popular social topics, and recommend relevant personalized contents (Li, Guo, & Zhao, 2008). Community analysis plays an important role in elucidating the creation, representation, and evolution of knowledge among individuals, while helping foster an understanding of their opinions. In addition, it can be used to organize and track content in online social media (Sundaram, Lin, Choudhury, & Kelliher, 2012). In social tagging systems, because many users independently assign shared tags to the same resource, a significant opportunity exists for investigating hidden community structures and the corresponding impact of these communities on information sharing.

Distilling communities from unstructured data is an extremely difficult task. Most studies on community detection focus on analysis of the graph structure, which captures the perception of a community as a set of nodes with better internal connectivity than external connectivity. Content information is typically ignored or not well integrated into graph-based methods. Particularly in

☆ This work is supported by National Natural Science Foundation of China (No. 61035004, No. 61273213).

* Corresponding author. Tel.: +86 18901386032.
E-mail address: zhenyu.wu@cse.buaa.edu.cn (Z. Wu).

http://dx.doi.org/10.1016/j.neunet.2014.05.019
0893-6080/

social tagging systems, graph-based methods become problematic in the following scenarios (Sun & Lin, 2013). For one, if a user has broad interests, he or she may annotate numerous diverse resources. This will generate many edges between the given user and others, making it difficult to determine the community to which the user should belong. In addition, if two users share only a few common resources, they may be divided into different communities because the edge weights are small. However, these two users may have similar interests and use many similar tags, indicating that they should be in the same community. Moreover, even if two users are interested in the same resource, they are likely to be interested in different aspects of that resource.

Furthermore, community detection is closely related to opinion mining. The models and techniques inspired by natural mechanisms, such as those studied in biology, can be used to discover user opinions. For example, Cambria et al. propose a novel cognitive model based on the combined use of multi-dimensional scaling and artificial neural networks to model the way multi-word expressions are organized in a brain-like universe of natural language concepts (Cambria, Mazzocco, & Hussain, 2013). Although the model is effective for opinion mining, capturing the opinions of users in real time is likewise important. However, traditional machine learning algorithms are not suitable for real-time processing. For example, many dynamic community detection methods divide data into successive time slices; communities are then detected in each time slice. In this way, the community detection method is repeatedly applied in every time slice, which is time consuming. Therefore, algorithms should be considered in a scalable way.

The streaming-based method is one of the most efficient ways to analyze big social data. In social tagging systems, a sequence of annotations can be regarded as a stream. Therefore, a real-time community detection method can be designed based on the stream. Meanwhile, two challenges are considered in the algorithm: efficiency and effectiveness. That is, the algorithm should process data in a single pass, use a limited amount of memory, and work in a limited amount of time. Furthermore, detected communities should accurately reflect the temporal interests of users.

Natural language tags can be incorporated in the community detection method to improve its accuracy because content information can help strengthen the community signal (Ruan, Fuhry, & Parthasarathy, 2013). In fact, tags have temporal properties; a more recently used tag can more accurately reflect the latest interest of a user. In addition, common interests, or latent interactions, exist between two users if they annotate the same resource. Similarly, latent interactions that most recently occur can more accurately reflect the latest user interests. Based on tags, latent interactions, and their temporal properties, user interests can be modeled by the vector space model and are referred to as the user profile vector. Dimensions in the user profile vector include tag frequency and latent interaction frequency. Values of all dimensions fade over time in correlation with the fading of interests. To efficiently cluster high dimensional user profile vectors, locality-sensitive hashing (LSH), proposed by Indyk and Motwani (1998), is applied to locate the nearest neighbor. Furthermore, an incremental method based on LSH is proposed to detect communities by processing the social tagging stream.

The main contributions of this paper are as follows. (1) A user profile vector based on the vector space model is proposed to represent the interests of a user. (2) An incremental community detection method based on LSH is proposed. This method can efficiently and effectively detect communities in social tagging systems. (3) Two measures are proposed to evaluate the community detection method in dynamic environments where traditional static measures will fail. (4) Social dynamics are analyzed in social tagging systems to elucidate the characteristics of big social data.

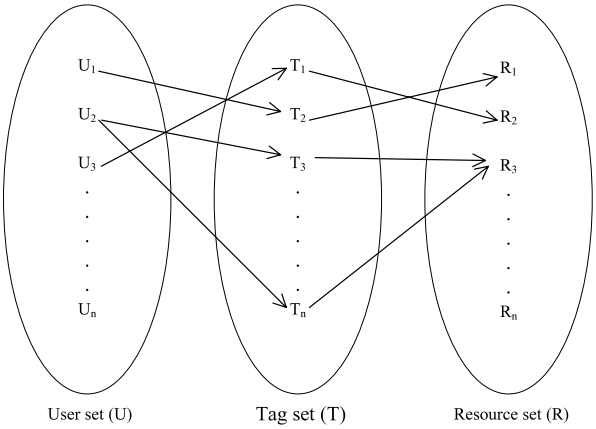

Fig. 1. Entities in social tagging systems.

The remainder of this paper is organized as follows. In Section 2, the background and motivation of this paper are discussed. In Section 3, related work is reviewed. In Section 4, the LSH-based community detection method is presented. In Section 5, experimental results are presented from the social dynamics analysis and comparison of the proposed method with the two state-of-the-art methods. This paper is concluded in Section 6.

2. Background

There are three types of entities in social tagging systems: user set (U), resource set (R), and tag set (T). U includes all users in the given tagging system. T includes the tags (words or phrases) employed by users. R includes resources of the tagging system. For example, in the present case, R refers to movies in MovieLens and artists in LastFM.

Definition 1. Tag assignment Y is a ternary relation between $U, R, T, Y \subseteq U \times R \times T$.

For example, in Fig. 1, $\langle U_1, T_2, R_1\rangle$, $\langle U_2, T_3, R_3\rangle$, and $\langle U_3, T_1, R_2\rangle$ are three tag assignments.

Definition 2. Tag assignment stream $AssignStream = (\langle Y_1, S_1\rangle, \langle Y_2, S_2\rangle \ldots)$ is a sequence of tag assignments in a time series, where Y_i is a tag assignment and $S_i (S_i \leq S_{i+1}, i = 1, 2, \ldots)$ is the timestamp of the corresponding tag assignment.

In one of the earliest studies of social tagging systems, Golder and Huberman determined that there are structural patterns, including the stabilization of tags over time, in Del.icio.us, even in the presence of large and heterogeneous communities (Golder & Huberman, 2006). This stabilization suggests a shared knowledge among users. Halpin et al. determined that power law distribution can be used to describe the stabilization (Halpin, Robu, & Shepherd, 2007). The same results were found by Cattuto, Loreto, and Pietronero (2007) and other researchers (Li et al., 2008). That is, an underlying social collective intelligence is embedded in the uncoordinated annotations of users, and the social collective intelligence can be used to explore the interests of users in social tagging systems.

Most previous studies on community detection are based on social link graphs. However, interactions can express the interests or opinions of users better than social link patterns. Many activities can generate information about interactions, such as annotating bookmarks, which implies user interests, and frequently commenting on the homepages of friends, which suggests the strength of connections. In other words, interactions can be explicit (such as a direct email exchange between two