Expert Systems with Applications 42 (2015) 366–375

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A semantic overlapping community detection algorithm based on field sampling

Yu Xin [a], Jing Yang [a,*], Zhi-Qiang Xie [b]

[a] College of Computer Science and Technology, Harbin Engineering University, Heilongjiang 150001, China
[b] College of Computer Science and Technology, Harbin University of Science and Technology, Heilongjiang 150001, China

ARTICLE INFO

Article history:
Available online 12 August 2014

Keywords:
Semantic social network
Community detection
Overlapping communities
Semantic modularity

ABSTRACT

The traditional semantic social network (SSN) community detection algorithms need to preset the number of the communities and could not detect the overlapping communities. To solve the issue of presetting the number of communities, we present a clustering algorithm for community detection based on the link-field-topic (LFT) model suggested. For the process of clustering is independent of context sampling, the number of communities is not necessary to be preset. To solve the issue of overlapping community detection, we establish the semantic link weight (SLW) depending on the analysis of LFT, to evaluate the semantic weight of links for each sampling field. The proposed clustering algorithm is based on the SLW which could separate the SSN into clustering units. As a result, the intersection on several units is the overlapping part. Finally, we establish semantic modularity (*SQ*) involving *SQ*1 and *SQ*2 for the evaluation of the detected semantic communities. The efficiency and feasibility of the LFT model and the semantic modularity is verified by experimental analysis.

1. Introduction

In accordance with the development of network communication, the electronic social network, such as Facebook and Twitter, has played an important part in people's daily social communication. Many social networking sites have launched the Community Recommended and Friend Circle Service to enrich people's web life. Thus, the community detection and recommendation algorithms have become the focus on social networks data mining. To date, community detection researching includes the following three aspects: hard community detection, overlapping community detection and semantic community detection.

The hard and overlapping community detection belongs to the topological community detection. The objective of these algorithms is to detect the communities with close internal relationships utilizing the properties of the relationships. The hard community detection is the pioneer work, and the ultimate goal of which is to divide the social networks into several separate networks (Newman, 2006; Newman & Girvan, 2004). The representative algorithms include GN (Girvan & Newman, 2002) and FN (Newman, 2004). In accordance with the development of hard community detection, researchers gradually focus on the case that a node belongs to several communities. Therefore, Palla, Derenyi, Farkas, and Vicsde (2005) suggested the CPM algorithm to detect the overlapping structures. After that, overlapping community detection research became the major concern in social networks and many representative algorithms were proposed, such as EAGLE (Shen, Cheng, & Cai, 2009), LFM (Lancichinetti, Fortunato, & Kertesz, 2009), COPRA (Gregory, 2010), UEOC (Jin, Yang, & Baquero, 2011), et al. The objective of semantic community detection is to cluster the nodes with similar semantic context (microblogging and social labels) into the same community. Since the semantic communities are detected by both context and relationship of the nodes, the result could represent the cohesion of communities more efficiently. For the semantic data mining must be based on the text analysis, many semantic community detection algorithms exploited the latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) model as the core model. According to the applied manner of LDA model, semantic community detection algorithms can be summarized as the following three categories:

(1) The LDA semantic analysis in terms of relationship. Such algorithms treated the topology of the social networks as semantic context, utilizing an improved LDA model to analyze the semantic similarity of nodes. Zhang, Qiu, and Giles (2007) proposed the SSN-LDA algorithm, regarding the ID and relationship as semantic context, the similarity of nodes as the training result. Henderson and Eliassi (2009) proposed the LDA-G algorithm to extend the

[*] Corresponding author.
E-mail addresses: xinyu@hrbeu.edu.cn (Xin Yu), yangjing@hrbeu.edu.cn (J. Yang), xiezhiqiang@hrbust.edu.cn (Z.-Q. Xie).

http://dx.doi.org/10.1016/j.eswa.2014.07.009
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

SSN-LDA model with infinite relational models (IRM) (Kemp, Tenenbaum, & Griffiths, 2006). The LDA-G combined the LDA model with graph model, allowing it to predict the potential links among the detected communities. Then Henderson et al. (2010) proposed the HCDF algorithm, extending the LDA-G with multiple attribute analysis and increasing its stability. The GWN-LDA (Zhang, Giles, & Foley, 2007) devoting to the directed networks and the HSN-PAM (Zhang, Li, & Wang, 2007) to the hierarchical networks were proposed based on the SSN-LDA. The advantage of such algorithms is the simply structure and the less requirement for input parameters, suitable for handling large-scale data. The disadvantages are that the semantic of such algorithms is not context and the detected community lack of the real semantic relevance.

(2) The LDA semantic analysis in terms of relationship-topic. Such algorithms treat the context of nodes as semantic context, analyzing the similarity of the nodes with semantic context. Most of such algorithms utilize the AT (Steyvers, Smyth, & Rosen, 2004) model as the basic model. The ART (McCallum, Corrada, & Wang, 2005) proposed by McCallum is the representative model, which added the recipient sampling into the AT model. The ART promoted the AT research into the field of SSN. After that, McCallum, Wang, and Corrada-Emmanuel (2007) designed the role analysis model (RART) based on the ART, extending the application fields of ART into the Social Computing. Zhou, Manavoglu, and Li (2006) applied the user distribution sampling to the AT model, suggesting the CUT model. Cha and Cho (2012) proposed the HLDA model which extract the relational tree model from online social networks on the basis of the relationship of reply context and design a hierarchical LDA to simulate the context relation tree. The advantages of such models are the extension of the context analysis into topological analysis for each node, and the detected community having a higher internal similarity. The disadvantages are that such models merely consider the relationship properties of the social networks, lacking of the consideration on the feature of local field. That would result in the disconnected community.

(3) The LDA semantic analysis in terms of community-topic. Such algorithms add the local field sampling into the relationship-topic model, developing the adjacency sampling to local area sampling. These algorithms avoid the case of disconnection in local field. The GT model (Wang, Mohanty, & McCallum, 2005) suggested by Wang, extending the ART model by replacing the recipient sampling with group recipient, is the representative model. Then, Pathak, DeLong, and Banerjee (2008) discussed the necessity of recipient sampling and proposed the CART model, adding the community sampling into the ART model. Recently, community-topic model has become the focus on SSN research. Mei, Cai, and Zhang (2008) combining the topic distribution in local field with the modularity, proposed the TMN model and established the topic-community correlation function to optimize the process of community detection. Sachan, Contractor, and Faruquie (2011, 2012) and Yin, Cao, and Gu (2012) proposed the TURCM and LCTA model, in terms of topic-community and community-topic distribution respectively. The both models above not only increased the difference of the topic distributions in different communities, but also made the result more reasonable. The advantage of such models is the high accuracy of the result. The disadvantages are not only the complex structure and the easy of getting over-fitting result, but the number of communities needs to be preset as the basic LDA model requires the prior parameters. The result tends to be different as the difference of presetting parameter.

Allow for the advantage of LDA analysis of community-topic on semantic community detection, we utilized the sampling manner of community-topic as the basis sampling manner. To avoid the problem of presetting the number of communities, we separated the community-topic detection into LDA sampling and semantic community detection. In the process of LDA sampling, we designed the sampling field which has a great weight in the central area and a low weight in marginal area, according to the semantic attenuation in the topic propagation. For the sampling manner replace the community sampling with the field sampling, it has not to preset the number of communities. In the process of semantic community detection, we designed the community clustering algorithm. The clustering element is link_block which is the smallest community. There would be an intersection part among different link_blocks, thus, overlapping nodes belonging to different clusters may exist. For that, the overlapping communities could be found. For the clustering process have no requirement for the number of clusters, the semantic community detection could be achieved without presetting the number of communities. For the measurement, we designed the semantic link weight (SLW) to evaluate the semantic weight of links, and the SQ ($SQ1$, $SQ2$) model to evaluate the detected semantic communities.

2. Link-field-topic (LFT) model

The representative semantic community detection algorithms such as AT, ART and HLDA sample the context of nodes in the form of point, surface and radiation, respectively. The sampling process of the three models is illustrated in Fig. 1. Fig. 1(a) shows the sampling process of AT model. In the AT model the node G_i and G_j are sampled separately without the consideration on relationships. Therefore, the sampling process of AT model is specific to node. Fig. 1(b) shows the sampling process of ART model. In the ART model the nodes directly adjacent to the sampling node are treated as the recipients. One of the recipients (G_1, G_2 and G_j) of G_i is sampled at random in sampling the node G_i. Separately, one of the recipients (G_3, G_4, G_5 and G_i) is sampled at random in sampling the node G_j. Essentially, the sampling process of ART model is in the form of the field around the node sampled. Fig. 1(c) shows the sampling process of HLDA model. In the HLDA model each node is sampled in a hierarchical manner. After sampling the node G_i, the 1-step distance nodes G_1, G_2 and G_j are sampled, then the 2-step nodes G_3, G_4, G_5 and G_7, and so on. Obviously, the sampling process of HLDA model is in the form of the radiate field around the node sampled.

The ART and HLDA model are the application of AT to the non-semantic network. As the sampling radius of ART is 1, the sampling field is relatively small. The sampling result merely representing the direct relationship, could not reflect community's block characteristic. The sampling process of HLDA is in the form of radiate field without weight, ignoring the impact of distance on sampling. For that, we suggest the LFT sampling model exploring the radiate sampling manner of HLDA, however, the sampling weight is given according to the distance step between nodes. The distance will decrease as the distance gets further. Therefore, the sampling field appears an internal compact block similar to the ART model. The sampling process of LFT model is specific to link, which is adjacent to more neighbors than node, and so, could sample the context more sufficiently. Unlike ART and HLDA, the sampling process of LFT is in the form of the field around the link sampled. Fig. 1(d) shows the sampling process of LFT model. In the sampling process of $link_{i,j}$, two points G_i, G_j of $link_{i,j}$ is sampled with weight for distance = 0, while the $G_1 \sim G_5$ for distance = 1, then the G_6, G_7 for distance = 2 and so on. The relevant mathematical symbols for illustrating the LFT model are given in Table 1.

For the semantic context is spread by message in SSN practically, the semantic context of message will get weak with the distance increase. Therefore, the weighting coefficient in LFT could be modeled by the Gaussian field (Zhu, Lafferty, & Ghahramani, 2003). The sampling weight of G_r in $link_{i,j}$'s sampling field can be obtained in Eq. (1).