Expert Systems with Applications 38 (2011) 3902–3908

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Role defining using behavior-based clustering in telecommunication network

Tian Zhu [a,*], Bai Wang [a], Bin Wu [a], Chuanxi Zhu [b]

[a] Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] School of Science, Nanchang University, Nanchang, Jiangxi 330031, PR China

ARTICLE INFO

Keywords:
Network
Node role
Customer behavior
k-means clustering

ABSTRACT

Understanding the individual behavior has shown to be of paramount importance to the triumph of the telecommunication operators to retain customers, enhance their purchasing capacity, and predict the churn rate. Different behavior patterns can be observed for different groups of users. Hence, there is an interesting problem posted in telecommunication network that how to define the users' role according to their behavior patterns. Traditionally, user behavior characterization methods generally based on their call detail record (CDR), which are user's individual features, are not appropriate to identify the role in network. In this paper, we develop a new methodology for identifying users' role based on their behaviors in telecommunication network using the social features instead of their individual features. Experiments have tested on synthetic data and large real datasets, and reveal good results on both of them. Finally, the methodology is not only limited to call graphs but also apply to other networks for role defining.

1. Introduction

Many datasets can be described in the form of graphs or networks where nodes in the graph represent entities and edges in the graph represent relationships between pairs of entities. A social network is a graph that the nodes generally represent individuals or organizations and the edges represent relationships among individuals (Wu, Wang, & Zhu, 2008). There is an increasing interest recently in developing algorithms and software tools for exploratory and interactive analysis of social networks, especially on online network and telecommunication network (Anagnostopoulos, Kumar, & Mahdian, 2008; Crandall, Cosley, & Huttenlocher, 2008).

Social network analysis methods try to achieve different goals. A well-known issue is the computation of the importance or centrality of vertices. For instance, Google's Pagerank defines the importance of Web-pages relative to an user's query and thereby facilitates seeing the most important results first. Other examples are popularity-rankings of users in social networking sites. A different goal in network analysis is the computation of densely connected groups of vertices, which is defined as community discovery. The hope is that these clusters correspond to natural divisions of the network into similar partitions. A third issue in social network analysis aims at computing groups of vertices that occupy the same structural position or play the same role in the network.

In the recent years, influence ranking and community discovery have received much interesting, while the notion of role assignment is not as well-known as them. Role assignment try to identify classed of vertices that occupy the same social position, play the same role, or have the same function in the network. It is different from the problem of community partition, for instance, densely connected employees may occupy different positions (like, e.g. manage or secretary).

There are not much research work focusing on role analysis in these years. Therefore, according to Scott (2000), social role is the core element of social network analysis. Thus, in this study, we focus on the problem that is to analyze the customer role in telecommunication network based on user behavior pattern. The comprehension of peculiarities of the user behavioral patterns in telecommunication systems are of significant importance in the sense they can give useful information for the operator to better understand their customers. Partition the different customers into various roles enable the operators to build strategies according to different behavior features. The present partition methods might only considering to the customers' CDR data (Lin, 2007; Yan, Fassino, & Baldasare, 2005) which are mainly standing for individual features. In contrast to these static records, the social features carry information not only about the individual users but also implicitly about their neighbors. These aggregated information are stronger and outweighs CDR records, thus could define better user behavioral pattern.

To achieve different goals that specialists in their different designing tasks, user behavior models have been extensively studied and are able to capture particularities from a group of users with a single behavior to multiple classes of users (Maia, Almeida, & Almeida, 2008), including Web, media, E-business, etc. There are also several research works concentrated on the telecommunication field (Dasgupta, Singh, & Viswanathan, 2008; Du, Faloutsos,

* Corresponding author. Tel.: +86 15810601863; fax: +86 791 8305885.
E-mail address: zhutian.bupt@gmail.com (T. Zhu).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.051

& Wang, 2009; González, Hidalgo, & Barabási, 2008), while few targeted on partition different kinds of customer roles.

The characteristic of our approach which different from the other behavior-based methods in that it combines both of the node identity and the relational data. By taking into account of the network structure, the role defining process can benefit from two main respects: firstly, it takes account of the interaction data of the nodes which is very useful for understanding the node's value; secondly, it can give a global view instead of a static view of the data.

The contributions of this paper include the following: providing a proper method, that is bringing in the structure information to the telecommunication network analysis; observed from the user behavior patterns form by the social features, we can partition the users into different roles according to which operators can make different strategies; we can also apply the methodology used in this paper for user role identification to social networks in general.

The remaining of this paper is organized as follows: Section 2 presents the related work. The data collection and feature sets are described in Section 3. Section 4 presents the role analysis process. Section 5 describes the experiment results and Section 6 concludes this work and expresses future directions.

2. Related work

2.1. Role analysis

A node role is a subjective characterization of the part it plays in a network structure, knowing the role of a node is important for making decision in many applications. To the best of our knowledge, the role defining problem based on relational behavior has been addressed only in a few works.

In their seminal paper, Lorrain and White (1971) proposed that vertices have the same role if they are structurally equivalent, i.e. have identical neighborhoods. This rather restrictive requirement has later been weakened in many ways (Everett & Borgatti, 1994). However, practical applicability of these kinds of vertex partitions to social networks or other irregular graphs is severely limited. Firstly, it is NP-complete to decide whether a graph has a regular, or equitable, partition with a given quotient graph. Secondly, most graphs analyzed in the social sciences have small or even trivial automorphism groups, irregularly distributed vertex degrees, etc., so that they hardly ever admit a non-trivial equitable partition anyway, while tractable regular partitions are often trivial as well. In this case, we should define another methodology of role analysis to apply to the practical large scale network.

2.2. Behavior pattern analysis

A number of previous research work focus on telecommunication network. Nanavati, Gurumurthy, and Das (2006) Nanavati, Singh, and Chakraborty (2008) have studied properties of the telecommunication network's structure and temporal properties. Structure and link properties in mobile communication networks have been discussed by Onnela, Saramäki, and Hyvönen (2007). Tapping into the user behavior, González et al. (2008) present the human behavior with whose position been tracked for a six-month period and give the conclusion that human trajectories show a high degree of temporal and spatial regularity.

Besides telecommunication network, specialists in other fields have done various research in user behavior. Agichtein et al. incorporate the user behavior to significantly improve the accuracy of a web search ranking algorithm (Agichtein, Brill, & Dumais, 2006). Maia characterizes different classes of user behavior in online social network (Maia et al., 2008). Backstrom introduces the three new concepts: thriving groups, k-core user and long-core user to define different kinds of user groups and users, and also investigate engagement and relationship between users and groups of users (Backstrom, Kumar, & Marlow, 2008). Menascé et al. consider classes of users to optimize the computational resources and the e-business itself (Menascé, Almeida, & Fonseca, 1999).

Differs from the works aforementioned, our study focuses directly on the social features formed behavior pattern rather than analyzing the user's individual information. We also present a methodology to identify characteristics that define user's role in the network.

2.3. Clustering techniques

There exist many different kinds of clustering algorithms for unsupervised learning. There are three widely used types of clustering, which are called K-means, spectral clustering, mixture models (Xiang & Gong, 2008). The performance of these clustering algorithms strongly depends on the clustering domains and dataset.

2.3.1. K-means clustering

K-means is a prototype-based, simple partition clustering technique which attempt to find an user-specified k number of clusters (Xiong, Wu, & Chen, 2006). These clusters are represented by their centroids, which is typically the mean of the points in the cluster. The procedure of K-means clustering is partition into three steps. First, k initial centroids are selected randomly, where k is the initial number of clusters. Second, assigning every point in the data to the closest centroid and each collection of points assigned to a centroid form a cluster. Third, update the centroid of each cluster based on the new cluster. This process is repeated until no point changes clusters. The whole procedure is repeated until no point switches cluster assignment or a number of iterations are performed.

2.3.2. Hidden Markov model based clustering

A HMM is a non-deterministic stochastic finite state automata (FSA). The basic structure of a HMM consists of a connected set of hidden states. HMM models can be described by the following three sets of probabilities: (i) the initial state probabilities, π, which defines the probability of each state being the starting state for any value sequence, (ii) the transition probability matrix, A, which defines the probability of going from one state to another, and (iii) the emission probability matrix, B, which defines the probability of generating a value at any given state.

2.3.3. Spectral clustering

Spectral clustering, for which an eigenvector or a combination of several eigenvectors is used a vertex similarity measure for computing the clusters. Many matrices as well as the widely used Laplacian can be used to compute such spectral measures; if instead of the adjacency matrix of a simple graph, the input is some kind of a similarity matrix for a complete graph, similar computations still may yield good results. Spectral clustering is typically based on computing the eigenvectors corresponding to the second smallest eigenvalue of the normalized Laplacian or some eigenvector of some other matrix representing the graph structure.

3. Behavior extraction

3.1. Data collection

We have consider three different telecommunication networks, built from one synthetic data and two real data from two different operators in China.

- *Synthetic data:* VAST 2008 call data. The cell phone call records of this dataset are synthetic data provided from the 2008 VAST