Computer Communications 41 (2014) 22–30

Contents lists available at ScienceDirect

Computer Communications

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

Identifying website communities in mobile internet based on affinity measurement

Jun Liu [a,*], Nirwan Ansari [b]

[a] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[b] Advanced Networking Lab., Electrical and Computer Engineering Department, New Jersey Institute of Technology, NJ, USA

ARTICLE INFO

Article history:
Received 20 August 2013
Received in revised form 26 November 2013
Accepted 30 December 2013
Available online 22 January 2014

Keywords:
Website community
Graph theory
Affinity measurement
Degree distribution

ABSTRACT

With the rapid development of mobile devices and wireless technologies, mobile internet websites play an essential role for delivering networked services in our daily life. Thus, identifying website communities in mobile internet is of theoretical and practical significance in optimizing network resource and improving user experience. Existing solutions are, however, limited to retrieve website communities based on hyperlink structure and content similarities. The relationships between user behaviors and community structures are far from being understood. In this paper, we develop a three-step algorithm to extract communities by affinity measurement derived from user accessing information. Through experimental evaluation with massive detailed HTTP traffic records captured from a cellular core network by high performance monitoring devices, we show that our affinity measurement based method is effective in identifying hidden website communities in mobile internet, which have evaded previous link-based and content-based approaches.

1. Introduction

Mobile internet refers to accessing the internet websites from mobile devices via cellular networks. With the rapid increase of powerful mobile devices, innovative mobile applications and increased cellular spectrum allocation, the traffic volume of mobile internet has been growing continuously [1,2]. Video traffic was 57% of all consumer internet traffic in 2012 and is growing [1], and there have been constant efforts in speeding up content delivery [3,4] and enhancing user experience [5–7]. To cope with the explosive growth of data volume and best serve their customers, mobile network operators need to design and manage their network resources from the overall perspective of the entire mobile internet. Basically, mobile internet services are delivered to users via cooperation among mobile devices, cellular networks and websites. From the viewpoint of service providers, user activities are centered around websites. Therefore, the community structure of websites can be a rich source of information about the circumstance of the mobile internet. In this paper, we propose a new algorithm to extract the community structures of websites in mobile internet and demonstrate its effectiveness by experiments with real-world traffic data. In particular, we focus on how to measure the affinity relationships between websites and use them for identifying important web communities in the mobile internet environment.

Despite the decentralized, unorganized and heterogeneous nature of world wide web, some works have shown that the complex web system can be described by graphs or networks which capture the connections between the entities they are made of, such as clients and websites [8]. In a graph, some nodes maintain closer relationship with each other than the rest of the graph. The set of such nodes is usually referred to as community, cluster or module. It is important and interesting to discover these *a priori* unknown community structures in graphs. The purposes of initial works on investigating the community structures are for structure visualizing [9,10] and content searching [11,12] in the web environment. Subsequently, there have been a growing number of works directed at revealing community structures based on various context information. In general, these works can be categorized into two types. The first type is the link-based approach which extracts community information from the link structure of the hyperlinked web environments. Another type is the content-based approach which defines the relationships between websites in terms of similarities between their contents, such as title, keywords, description, and words in the web pages. Most of content-based works use a vectorial representation of web pages to cluster web sites by related topics. Owing to the semantic relevance between the definition of relationships between websites in the above two types [13], some works have been proposed to combine link and content information for identifying web communities [14].

* Corresponding author. Tel.: +86 10 62283742.
E-mail addresses: liujun@bupt.edu.cn (J. Liu), nirwan.ansari@njit.edu (N. Ansari).

0140-3664/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comcom.2013.12.008

Although there has been tremendous interest in identifying website communities, earlier studies mainly focused on the original information of web pages collected by various crawlers or through web data mining. To a certain extent, the existing relations between website communities and user behaviors have been overlooked and are far from being well understood. In this paper, we carry out the mining of website communities from the user perspective by examining detailed HTTP traffic records in mobile internet. Our traffic records were captured by powerful line-speed monitoring devices which tracked a 10Gbps trunk link in the backbone of a cellular data network. The records consist of complete information of accessing users, used devices and serving websites. The rich information allows us to establish the relations between website communities and accessing users. Based on the examination of these records, we propose a shared user based affinity measurement between websites and build an affinity graph to represent the structure of observed mobile websites. However, the original dense affinity graph may poll too large a set of relations between nodes that will incur high computational workload as well as decrease the fidelity of the mining task. So, we modify a scale-free topology criterion, originally designed for undirected graph, to transform the full directed affinity graph into a sparse one by choosing a threshold parameter value that leads to a graph whose in-degree distribution follows a power law. Then, the influence score, which represents the importance of each website in the sparsified affinity graph, can be calculated based on the out-degree values and weights of its neighbors. At last, all websites are ranked by the calculated influence scores. The k websites with top scores and their neighboring websites in the sparsified affinity graph are identified as top-k website communities. Convincing results based upon real-word traffic data have substantiated the effectiveness of our systematic method.

The main contributions of our work are twofold. First, we have proposed a new algorithm for identifying top-k website communities in mobile internet. Unlike existing solutions, our method can identify the hidden community structure from the user behavior perspective, which cannot be revealed by existing link-based and content-based community identification approaches. Second, we have conducted experiments on a large cellular data network with massive HTTP traffic data to evaluate the effectiveness of our algorithm. Experiments on real-world data show that our algorithm can effectively identify communities in which websites have strong affinity relationships. Two novel types of website communities are identified and explained in the evaluation.

The rest of this paper is organized as follows. Section 2 provides a brief review of the background and related works. Section 3 details our proposed method for identifying website communities. We then present how we evaluate the effectiveness of our method and discuss the results in Section 4. Finally, we conclude the paper in Section 5.

2. Background and related works

Website community identification in the web environment has attracted ample attention in recent years. The most basic and straightforward approach for identifying communities is grouping web objects according to their natural hyperlink structure characteristics. A web environment can be modeled as a graph in which a node represents a website or web page on various contexts and an edge represents a link from one node to another. After abstracting the web as a graph, a set of graph theory based mathematical tools can be applied to detect communities for different purposes. For example, Gibson et al. [15] defined a hyperlinked community on web which contains a core of authoritative central node and linked hub pages. Based on their proposed HITS algorithm [11], community structures can be derived through the link topology. Another graph theory tool applied in identifying web communities is the maximum flow and minimal cut theorem. Flake et al. [16] defined a community as a set of web pages that link to more pages in the community than to pages outside the community. Under this definition, the problem of identifying communities is mapped into a family of graph partitioning problems. The identification procedure is carried out as a loop for a given number of iterations using the maximum flow and minimum cut algorithm. Also, Merelo-Guervs et al. [17] proposed a Self-Organizing Map (SOM) based method to divide a set of blog websites into communities and produce a community navigation map. Besides the graph model, vectors are also used to represent the web objects for both content similarity and hyperlink structure based considerations [18–20]. The advantage of vector representation is that it is suitable for applying clustering technologies based on well studied vector oriented distance measuring algorithms like k-means and k-nearest neighbors clustering. For example, works in [21,22] clustered websites by computed distances between vectors, which represent the topic features and link weights respectively, and both yielded reasonable results.

Owing to the high computational complexity of crawling and parsing contents and links in web pages, the above link-based and content-based methods do not scale well, i.e., computationally expensive in a large scale environment. In addition, absence of user behavior information in these procedures renders the identified communities useless in some situations. We present two motivating examples to understand the importance of identifying user behavior related website communities in mobile internet. The first example deals with the replication of hotspot web contents to reduce cost and improve user experience. With the increasing bandwidth of cellular data networks, users are consuming more resource-hungry and quality sensitive services on their mobile devices, such as video and online gaming applications. To reduce network traffic from the third party for cost saving and to shorten response time for improving user experience, mobile operators tend to build self-owned web servers for replicating hotspot web contents. Because of the copyright issue, they have to negotiate with the owners of such websites for importing hotspot contents one by one. Hence, choosing appropriate websites to replicate in a limited resource condition including space, power and bandwidth is becoming critical. A straightforward method is to replicate web objects based on the rank of accessing popularity [23]. This simple approach, however, requires a long time to accumulate enough request statistics for each object. During the collection time, the hot object may not be popular anymore. Therefore, mobile operators need a more intelligent method to find correlated popular websites to be replicated. The second example is the requirement of identifying the latent service dependency between websites. With advances of web services technologies, some important websites which provide critical web services are hidden behind other websites, such as authentication server for single sign-on, advertising content server for advertising display websites, and streaming servers. The relations between these service providing websites and supporting websites may not be obvious in the hyperlink structure because most of them are implemented by dynamic programming interface rather than static texture link. The user access logs are the only information that can be relied onto infer the relationships between websites.

In above two examples, the common goal is to find some communities in which websites are correlated by shared users rather than hyperlink structure or content similarity. We call such a set of websites as an *affinity community*. In this paper, we propose a means to identify such communities by: (a) considering the relationships between websites based on the user behavior information, (b) quantifying the top-k important communities, and (c)