Neural Networks 58 (2014) 111–121

Contents lists available at ScienceDirect

Neural Networks

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

2014 Special Issue

Community-aware user profile enrichment in folksonomy

Haoran Xie [a], Qing Li [a,b], Xudong Mao [a], Xiaodong Li [a], Yi Cai [c,*], Yanghui Rao [a]

[a] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong Special Administrative Region
[b] Multimedia Software Engineering Research Centre, City University of Hong Kong, Kowloon, Hong Kong Special Administrative Region
[c] School of Software Engineering, South China University of Technology, Guangzhou 510006, China

ARTICLE INFO

Article history:
Available online 27 May 2014

Keywords:
Social media
Community
User profiling
Folksonomy
Big data

ABSTRACT

In the era of big data, collaborative tagging (a.k.a. folksonomy) systems have proliferated as a consequence of the growth of Web 2.0 communities. Constructing user profiles from folksonomy systems is useful for many applications such as personalized search and recommender systems. The identification of latent user communities is one way to better understand and meet user needs. The behavior of users is highly influenced by the behavior of their neighbors or community members, and this can be utilized in constructing user profiles. However, conventional user profiling techniques often encounter data sparsity problems as data from a single user is insufficient to build a powerful profile. Hence, in this paper we propose a method of enriching user profiles based on latent user communities in folksonomy data. Specifically, the proposed approach contains four sub-processes: (i) tag-based user profiles are extracted from a folksonomy tripartite graph; (ii) a multi-faceted folksonomy graph is constructed by integrating tag and image affinity subgraphs with the folksonomy tripartite graph; (iii) random walk distance is used to unify various relationships and measure user similarities; (iv) a novel prototype-based clustering method based on user similarities is used to identify user communities, which are further used to enrich the extracted user profiles. To evaluate the proposed method, we conducted experiments using a public dataset, the results of which show that our approach outperforms previous ones in user profile enrichment.

1. Introduction

There has been a proliferation of collaborative tagging (a.k.a. folksonomy) systems in Web 2.0 communities in recent years. For example, Delicious[1] allows users to tag bookmarks according to their own interests, and Flickr[2] users are able to annotate their own uploaded photos. The images (or bookmarks) and tags posted by Web users in such systems are indicative of user interests, and the tags given by users provide a rich information source for building accurate and specific user profiles (Cai & Li, 2010). Tags given by different users to a single image provide a collaborative description of the image, which, from the users' perspective, makes the image more meaningful and useful.

Collaborative tagging systems also allow for the construction of user profiles that are important in a variety of computational endeavors such as personalized search and recommender systems. Several studies, for example Noll and Meinel (2007), Xie, Li, and Cai (2012) and Xu, Bao, Fei, Su, and Yu (2008), have examined the construction of user profiles from tags in collaborative tagging systems. However, current methods of constructing a user profile by extracting features from user-generated tags often result in "sparse" user profiles which do not sufficiently reflect user preferences. The primary reason for this is that there are always some users that are relatively inactive and only use a small number of tags, or only annotate a limited number of images. Sparse user profiles inevitably restrict the utilization of user profiles in applications such as personalized search and recommender systems. Hence, there is a need to find suitable methods for enriching user profiles so that they reflect user preferences more thoroughly.

On the Web, the behavior of users is highly influenced by the behavior of their neighbors or community members (Teevan, Morris, & Bush, 2009). Users with similar tastes share common or close views on images and hence form a latent user community

* Corresponding author. Tel.: +86 13902333439.
E-mail addresses: yicai.scut@gmail.com, ycai@scut.edu.cn (Y. Cai).

[1] http://delicious.com.
[2] http://www.flickr.com.

http://dx.doi.org/10.1016/j.neunet.2014.05.009
0893-6080/

(Fischer, 2001). These are called 'latent' user communities because there is no explicit data identification for these communities, as opposed to other user communities which may be easily identifiable (e.g., they exist within a single domain). The identification of these latent user communities is one way to better understand and meet user needs. In systems that employ collaborative tagging, the tagging data generated by the user community forms a type of folksonomy (a 'folk taxonomy') attached to the images (or bookmarks) used by the members of the community. Importantly, the data found in these latent user communities can facilitate the construction of user profiles. Inspired by this, in this paper we focus on the issue of enriching user profiles based on latent user communities in folksonomies. To achieve this enrichment, we first propose a method to discover latent user communities based on user-generated tags and the correlation of image contents, and then devise a method to enrich user profiles in folksonomy based on the communities discovered. The contributions of this paper are listed as follows:

1. We propose a facility named the multi-faceted folksonomy graph (MFG) which incorporates both image content similarity and semantic tag similarity to represent the relationships among users, tags and images.
2. We propose a novel method to measure the similarity between any pair of users in an MFG by random walk distance.
3. We propose a prototype-based clustering method to discover latent user communities by exploiting the combination of contents and tags. Based on the extracted latent user community, we propose a novel folksonomy-based user profile enrichment method.
4. To evaluate the proposed method, we conduct experiments by using a public dataset, the results of which show that our proposed approach outperforms previous ones in user profile enrichment.

The remainder of this paper is structured as follows: In Section 2, we review some relevant studies. Section 3 describes our proposed method to enrich user profiles based on latent user communities by constructing a multi-faceted folksonomy graph (MFG). Section 4 details the experiment conducted on the NUS-wide dataset (Chua et al., 2009) to evaluate our proposed method. Finally, we conclude our work and give possible future research directions in Section 5.

2. Related work

Work related to the present study includes research on folksonomy, sentiment analysis, community discovery, and user profiling techniques for Web 2.0. We reviewed a selection of related studies and categorized them into the following four categories.

2.1. Folksonomy

Folksonomies resulting from collaborative tagging systems are a prevalent feature of prevailing the Web 2.0 era. These differ from other taxonomies or classification schemes in being user-generated rather than constructed by a third party. Two streams of research in this area are apparent. One focuses on intrinsic characteristics and pattern mining in tags or tagging behavior. For example, Bischoff, Firan, Nejdl, and Paiu (2008) analyzed tag datasets through statistical approaches in order to identify valuable tags for search purposes. Gupta, Li, Yin, and Han (2010) investigated and summarized various characteristics of tag data and various techniques of tag generation, tag analysis, and tag visualization. The focus of the other research stream is on potential applications of tagging techniques in Web applications. Bao et al. (2007) proposed two algorithms named the SocialSimRank (SSR) and SocialPageRank (SPR), which improved the popular PageRank algorithm by incorporating social tags so as to facilitate Web searches. Markines et al. (2009) investigated semantic content in tag-based profiles to identify appropriate metrics for similarity measures. To address the issue of tag refinement, Sang, Xu, and Liu (2012) devised ranking-based multicorrelation tensor factorization (RMTF) by integrating the semantic, visual, and social relations of tags, images, and users. Lau, Xia, and Ye (2014) proposed a probabilistic generative model to uncover hidden cybercriminal networks from online social media.

2.2. Sentiment analysis

Sentiment analysis techniques (e.g., sentiment dictionaries Godbole, Srinivasaiah, & Skiena, 2007, Kim & Hovy, 2004 and Mullen, Wiebe, & Hoffmann, 2009) can be used to exploit latent semantic content in tags. A trend in recent research (Cambria, Havasi, & Hussain, 2012; Cambria & Hussain, 2012; Cambria, Schuller, Liu, Wang, & Havasi, 2013a, 2013b; Cambria, Schuller, Xia, & Havasi, 2013; Cambria, Speer, Havasi, & Hussain, 2010) has been on concept-level sentiment mining, which can be used to identify sentimental relationships in tag corpora. Tsai, Tsai, and Hsu (2013) proposed a two-step method, which integrates iterative regression and random walk techniques to build a concept-level sentiment dictionary, by generating sentiment values using the ConceptNet semantic network. Xia, Zong, Hu, and Cambria (2013) addressed the domain adaptation problem of sentiment classification by proposing a feature ensemble plus sample selection (SS–FE) approach, which exploited both labeling adaptation and instance adaptation to improve overall performance. Cambria, Olsher, and Rajagopal (2014) developed SenticNet which makes use of sentic computing for concept-level analysis. SenticNet has been widely used in various social media applications such as information management (Grassi, Cambria, Hussain, & Piazza, 2011), social media marketing (Cambria, Grassi, Hussain, & Havasi, 2012), and big social data analysis (Cambria, Rajagopal, Olsher, & Das, 2013; Poria, Cambria, Winterstein, & Huang, 2014).

2.3. User community

In terms of user communities, there are two schools of research relevant to the present study. The first is community detection, which endeavors to uncover latent communities from the structures or features of Web applications. Typical approaches in folksonomic data for community discovery follow cluster-based paradigms such as the topic model (Tang, Jin, & Zhang, 2008), conceptual space (Wu, Zhang, & Yu, 2006), and probabilistic models (Zhou, Manavoglu, Li, Giles, & Zha, 2006). Lin et al. (2011) searched for user communities, that is, sets of users who share similar features in a social media context, using a metagraph factorization method, while Zhou, Cheng, and Yu (2009) used an adaptive clustering method to model both structure and attributes of a community into a unified graph. The second focus of research concerns the facilitation of Web searches on existing communities or communities that have already been discovered. Almeida and Almeida (2004) designed a community-aware search engine by leveraging the typical ranking method with additional information from the community. Similarly, Park and Ramamohanarao (2007) improved on the classic PageRank algorithm for Web searches with the concept of "community popularity". Other complex search tasks have been facilitated by the utilization of various information sources in user communities, such as collaborative searches (Smyth, 2007), expertise searches (Deng, King, & Lyu, 2009), groupized searches (Teevan et al., 2009), and community-aware searches (Xie et al., 2012).