Physica A 413 (2014) 125–133

Contents lists available at ScienceDirect

Physica A

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

Modeling social tagging using latent interaction potential☆

Zhenyu Wu *, Ming Zou

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

HIGHLIGHTS

- We examine the dynamic patterns in social tagging systems.
- We define latent interaction to present common interests among users.
- We propose a method to identify the importance of users.
- We propose a model to simulate social tagging.
- The proposed model is appropriate to explain the social tagging process.

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form 2 June 2014
Available online 5 July 2014

Keywords:
Social tagging system
Latent interaction potential
Collective intelligence

ABSTRACT

Modeling social tagging plays a critical role in identifying statistical regularities and structural principles common to social tagging systems. Existing modeling approaches only consider imitations or background knowledge of users. However, common interests among users are ignored. In this paper, latent interactions are applied to present the common interests, and dynamic patterns in empirical data are investigated. Furthermore, the latent interaction driven model (LIDM) is proposed to model social tagging. Experimental results show that the tag frequency distribution generated by LIDM is consistent with that in real-world data. Moreover, the latent interaction graph generated by LIDM has a higher average clustering coefficient and lower average shortest path compared with that generated by preferential attachment methods. This demonstrates that LIDM outperforms traditional methods.

1. Introduction

Users can discover, organize, and share resources using tags (words or phrases) in social tagging systems which are generated from decentralized local interactions among a group of users [1]. However, decentralized interactions do not necessarily imply that the systems remain unstructured over time. Instead, a highly stable structure emerges, and the stabilization can be described by power law distribution. So far, it is assumed that the emergence of power law distribution is mainly driven by imitation behavior of users [1–5].

Modeling social tagging is important to understand human behavior, and study collective intelligence. Moreover, it is helpful to make a deeper understanding of user interests, so that personalized information can be provided to them.

Although social tagging models are extensively studied recently, underlying interactions and dynamics among users are still unclear. Meanwhile, the following aspects are not considered. For one, many of these studies consider individual tag usage, while ignoring interactions among users [6]. Moreover, the majority models do not consider selection of similar users

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

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

http://dx.doi.org/10.1016/j.physa.2014.06.048
0378-4371/

Table 1
Dataset statistics.

Datasets	Users	Tags	Resources	Resource description	Tag assignments
Del.icio.us	1867	53 388	69 226	URLs	437 593
LastFM	1892	11 946	17 632	Artists	186 479
MovieLens	2113	13 222	10 197	Movies	47 957

when simulating imitation behavior. In addition, changes of user interests are ignored. Therefore, cognitive behavior of users should be considered, and a model that simply generates a power law distribution is not enough [7].

Because users tend to interact with others who have similar interests, latent interaction is defined to present common interests among users. Two users have latent interactions if they annotate the same resource. Recently, Monteiro et al. simulate the evolution of species based on similarity between individuals [8]. Thus, user profile vector is used to calculate the similarity between two users in this paper. Dynamic patterns in social tagging systems are investigated. Furthermore, the latent interaction driven model (LIDM) is proposed based on latent interaction and dynamic patterns. First, latent interaction graph is constructed. Second, latent interaction potential is defined to identify target interacting users.

To the best of our knowledge, our work is the first to model the social tagging considering dynamic patterns and latent interactions.

2. Background

2.1. Related work

2.1.1. Explanations of emergent patterns

Golder and Huberman determined that the proportion of tag frequencies stabilizes over time. They assumed that imitations among users lead to the stabilization, and applied the Polya Urn model to present imitation behavior [1]. However, the model cannot explain the popularity of tags, and cannot describe the addition of new tags. Halpin et al. modeled the imitation by the preferential attachment mechanism: relative frequency of a used tag determines probability of its reuse. They determined that power law distribution can be used to describe the stabilization [2]. Cattuto et al. proposed the memory-based Yule–Simon model to explain addition of new tags [4]. A user selects a new tag with probability p, and copies one from existing tags with probability $1 - p$. The probability of copying a tag decays with time, and the decay function follows a power law distribution. Results show that the generated tag frequency follows a power law distribution. In addition to imitations, Dellschaft et al. incorporated user background knowledge in their model [5]. One of the existing tags is imitated with probability I, and a tag is selected according to his or her background knowledge with probability BK.

Rather than imitating users at tag level, other researches focused on semantic imitations. Fu et al. adopted a conceptually different approach by explicitly applying models of cognitive psychology to explain semantic imitations [3]. In their model, users perceive tags as important cues, and process them intensively to infer topics contained in resources. The consensual usage of tags leads to tag frequency following power law distribution. Paul et al. investigated tag imitations, and found evidence for semantic reconstructive process, which may be the basis of semantic imitations [9].

2.1.2. Generative models

Numerous connectivity driven models were proposed to explain the emergence of power law distributions, such as preferential-attachment-based [10–12], optimization-strategy-based [13,14], and local-rule-based [15]. However, these models necessarily provide a time-aggregated representation that may fail to describe instantaneous and fluctuating dynamics of many networks [16].

Chuang et al. determined that imitations among individuals result in a self-organized state in social tagging systems [6], and they proposed an evolutionary hypergraph model for explaining the emerging statistical properties [17]. In addition, a variety of models on the basis of the latent Dirichlet allocation (LDA) were proposed for modeling generation of social annotations [18–21].

2.2. Datasets and dynamic patterns

2.2.1. Datasets

Detailed dataset statistics are listed in Table 1. The first dataset was extracted from Del.icio.us, which supports bookmark annotating; the second dataset was extracted from LastFM, which supports artist annotating; the third dataset was extracted from MovieLens, which supports movie annotating [22]. The assignment of a tag to a resource is referred to as a tag assignment, which can be presented as $\langle u, r, t, timestamp\rangle$ meaning user u annotates resource r using tag t at *timestamp*.

2.2.2. Dynamic patterns

In this section, dynamic patterns in three datasets are investigated. Because linear least squares fitting technique provides a solution to find the best fitting straight line, it is used to make clear dynamic patterns in different systems. The fitting parameters γ represents intercept, and τ represents slope.