Physica A 402 (2014) 134–140

Contents lists available at ScienceDirect

Physica A

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

Ceiling effect of online user interests for the movies

Jing Ni, Yi-Lu Zhang, Zhao-Long Hu, Wen-Jun Song, Lei Hou, Qiang Guo, Jian-Guo Liu *

Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, PR China

HIGHLIGHTS

- The information entropy is introduced to measure the diversity of the user interest.
- The interests of the small- and large-degree users are centralized, while others' interests are diverse.
- Small-degree users' interests are far easier to predict than the other users, which may shed some light for the cold-start problem.

ARTICLE INFO

Article history:
Received 9 August 2013
Received in revised form 22 December 2013
Available online 28 January 2014

Keywords:
Online social networks
User interests
Information entropy
Recommendation

ABSTRACT

Online users' collective interests play an important role for analyzing the online social networks and personalized recommendations. In this paper, we introduce the information entropy to measure the diversity of the user interests. We empirically analyze the information entropy of the objects selected by the users with the same degree in both the MovieLens and Netflix datasets. The results show that as the user degree increases, the entropy increases from the lowest value at first to the highest value and then begins to fall, which indicates that the interests of the small-degree and large-degree users are more centralized, while the interests of normal users are more diverse. Furthermore, a null model is proposed to compare with the empirical results. In a null model, we keep the number of users and objects as well as the user degrees unchangeable, but the selection behaviors are totally random in both datasets. Results show that the diversity of the majority of users in the real datasets is higher than that the random case, with the exception of the diversity of only a fraction of small-degree users. That may because new users just like popular objects, while with the increase of the user experiences, they quickly become users of broad interests. Therefore, small-degree users' interests are much easier to predict than the other users', which may shed some light for the cold-start problem.

1. Introduction

The growth and maturity of online social networks have provided good conditions and opportunities to analyze the collective behaviors of online users [1,2]. Especially in online e-commerce systems, the study of user collective behaviors could help sellers understand the selections or purchasing behaviors of online users and design highly efficient personalized recommendation algorithms [3–5]. Barabási [6] investigated the online network science and emphasized the great significance of understanding the Web's structure and human dynamics. Onnela et al. [7] analyzed the Facebook applications to study the role of social influence on the patterns of user behaviors and found that there was an on–off pattern of user selections, which was also named common and specific interests [8,9]. Ye et al. [10] analyzed the group purchasing behaviors of daily deals and formulated a predictive dynamic model of the collective attention for group buying behaviors.

* Corresponding author. Tel.: +86 15202121269.
E-mail address: liujg004@ustc.edu.cn (J.-G. Liu).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2014.01.046

Actually, the user interest plays an important role for analyzing the user collective behaviors since one's selection or rating behaviors are always influenced by his own interest [11]. A better understanding of user interests has the potential to significantly improve user experience [12]. Qiu et al. [13] and Speretta et al. [14] investigated user interests by search engines to provide personalized search results. Huang et al. [11] applied a random graph modeling methodology to studying the consumer purchase behaviors, and found that the topological characteristics of several consumer-product graphs were significantly different with the random bipartite graph, which suggested that consumers' product choices were not random. In addition, the online users' collective behaviors are very important for personalized recommendations [15–19]. For instance, Ye et al. [16] introduced a point-of-interest (POI) recommendation service for the local-based social networks. Oh et al. [17] took users' preferences and demographics into account when presenting a user-selectable recommendation system that reflected a user interest group. Overall, it is necessary and essential to recognize a user's interest more accurately and efficiently.

Now the question is what kind of metrics would be better for measuring the user interests. In the previous studies, the clustering coefficient of the bipartite network C_4 was introduced to measure the diversity of the user interests [20]. The clustering coefficient C_4 with cycle of size equal to 4 proposed by Lind et al. [21,22] was used to measure the contribution to the cliquishness of the second neighbors. Based on C_4, Shang et al. [4] introduced the collaborative similarity into the user-object bipartite network to measure the user interest. Besides, Liu et al. argued that the measurement of the user tastes [12] were very important for personalized recommendations, and they also empirically analyzed the clustering coefficient C_4 in the user-object bipartite networks [20]. The empirical results implied that C_4 could be used to describe the diversity of the user interest and that the higher the C_4 was, the more centralized interest the user had. However, C_4 just focuses on the neighbors' closeness of a target node, which means that such user interest is the correlation among their neighbors, therefore, we need more specific index to measure online users' collective behaviors for movies.

The information entropy [23–25], a concept of measuring the amount of the information [26], is an effective way to measure the information embedded into the users' collective behaviors. The greater the entropy is, the more information is contained in the user behaviors. Inspired by this idea, we argue that the information entropy of the objects selected by users with the same degree could indicate the diversity of their interests. That is, the higher the entropy is, the more diverse the user interest would be. In this paper, we select both the MovieLens and Netflix datasets and classify the data according to the user degree. Then we calculate the information entropy of the objects selected by each kind of users respectively, and the average degree of the objects as well. The empirical results show that there is a ceiling effect of the user collective interests for the movies, which means that both these datasets have similar trends, which is that the entropy of users with small-degree as well as with large-degree is relatively lower than that of the normal users. Besides, the average degree of the objects selected by each kind of users decreases as the user degree increases. We find from the results that small-degree users tend to select popular objects while large-degree users like unpopular objects, so their interests are both less diverse. Therefore, the information entropy should be taken into account to measure the user interests. Furthermore, a null model is proposed to compare with the empirical results, which indicates that user collective behaviors are far more different with the random selection mechanism.

2. Information entropy

The information entropy proposed by Shannon [23] quantifies the expected value of the information contained in a message. Actually, it is a method of measuring the amount of the information, or the amount of uncertainty of a system. That is, the greater the uncertainty of a variable is, the higher the entropy would be, so that greater amount of information is needed. For a random variable P with n outcomes $\{p_1, p_2, \ldots, p_n\}$, the information entropy, denoted by H, is expressed as

$$H = H(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}, \tag{1}$$

where $p_1, p_2, \ldots, p_n$ is a finite probability distribution, that is, suppose $p_i \geq 0$ $(i = 1, 2, \ldots, n)$ and $\sum_{k=1}^{n} p_i = 1$.

It should be noted that the entropy has two characterizations, which is called the characterization of maximum. They will be used to standardize the entropy hereinafter. The first one is that,

$$H(p_1, p_2, \ldots, p_n) \leq H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_2 n. \tag{2}$$

The above inequality indicates that the measure should be maximal if all the outcomes are equally likely, which is that the uncertainty is the highest when all the possible events are equiprobable.

The other feature is defined by,

$$H\left(\frac{1}{n-1}, \frac{1}{n-1}, \ldots, \frac{1}{n-1}\right) < H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right), \tag{3}$$

For equiprobable events, the entropy should increase with the number of outcomes, which means that as the n becomes larger, the uncertainty of the events would be greater.