Physica A 391 (2012) 1822–1830

Contents lists available at SciVerse ScienceDirect

Physica A

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

Behavior patterns of online users and the effect on information filtering

Cheng-Jun Zhang, An Zeng *

Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland

ARTICLE INFO

Article history:
Received 23 July 2011
Received in revised form 27 September 2011
Available online 13 October 2011

Keywords:
Bipartite network
Behavior pattern
Recommender system
Information filtering

ABSTRACT

Understanding the structure and evolution of web-based user-item bipartite networks is an important task since they play a fundamental role in online information filtering. In this paper, we focus on investigating the patterns of online users' behavior and the effect on recommendation process. Empirical analysis on the e-commercial systems show that users' taste preferences are heterogeneous in general but their interests for niche items are highly clustered. Additionally, recommendation processes are investigated on both the real networks and the reshuffled networks in which real users' behavior patterns can be gradually destroyed. We find that the performance of personalized recommendation methods is strongly related to the real network structure. Detailed study on each item shows that most hot items are accurately recommended and their recommendation accuracy is robust to the reshuffling process. However, the accuracy for niche items is relatively low and drops significantly after removing users' behavior patterns. Our work is also meaningful in practical sense since it reveals an effective direction to improve the accuracy and the robustness of the existing recommender systems.

1. Introduction

Complex networks have been studied intensively for more than a decade. The rapid development of network science has greatly helped us to understand and model real systems [1]. So far, many systems have been described by networks, such as the transportation system where nodes are airports and links are airlines [2–5], the neural system where nodes are neurons and links are synapses [6], the social system where nodes are people and links are the social interactions [7,8], the power grid where nodes are power plants and links are power cables [9]. Some other systems coupled by two different elements are modeled by the bipartite networks. For example, the e-commercial systems consisting of online users and items [10], the scientific collaboration system consisting of authors and papers [11], family name inheritance system consisting of babies and names [12] are naturally described by such networks.

The e-commercial systems have brought giant benefit to our daily lives. Nowadays, we can simply order books, movies, clothes etc. from the online retailer even at home. However, like a coin has both sides, internet also brings us overabundant information so that we always have too many candidate products to compare. In order to solve the problem, many recommendation algorithms such as collaborative filtering [13,14], content-based analysis [15], spectral analysis [16] and iterative self-consistent refinement [17] were developed to filter irrelevant information. Some physical dynamics on the bipartite networks, including mass diffusion [18] and heat conduction process [19], have also been applied to design recommendation algorithms. The hybrid algorithm combining mass diffusion and heat conduction is shown to obtain significant improvement in both recommendation accuracy and item diversity [20]. Very recently, the performance of diffusion-based recommendation methods has been further enhanced by the preferential diffusion process [21] and modified heat conduction [22].

* Corresponding author. Tel.: +41 026 3009275.
E-mail address: an.zeng@unifr.ch (A. Zeng).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.09.038

Table 1
Properties of the used datasets.

Network	Users	Items	Links	Sparsity
Movielens	943	1682	82,520	$5.20 \cdot 10^{-2}$
Netflix	3000	3000	197,248	$2.19 \cdot 10^{-2}$
Delicious	10,000	232,657	1,233,997	$5.30 \cdot 10^{-4}$
Amazon	99,622	645,056	2,036,091	$3.17 \cdot 10^{-5}$

Previous works have found that people's behavior is far different from random and obeys certain predictable rules [23,24]. From the network point of view, the users' online behavior will emerge some typical statistical patterns on the network structure [25,26]. Besides, the network structure properties have been shown to affect the recommendation process [27–29]. Therefore, users' behavior patterns will inevitably influence the recommendation result. Actually, users' behavior patterns are commonly believed to be beneficial for recommendation since related algorithms generally filter relevant information by cooperating the history of similar users. If users randomly choose items, the recommender systems will not have valuable information to refer to. Hence, how much the recommendation performance relies on these behavior patterns of users can be an interesting problem. It is not only helpful for understanding the effect of network structure on recommendation process, but also meaningful in improving the robustness of the existing recommender systems.

In this paper, we focus on understanding online users' statistical behavior pattern and the related effect on recommendation. We will compare the real bipartite networks with the randomized counterpart networks (i. e. the reshuffled networks) in which real users' behavior pattern can be gradually removed. Actually, some specific properties of the real networks have been discovered by the comparison to the reshuffled networks such as the loop distribution [30,31], rich club [32,33], community structure [34], assortative [35] and motifs [36]. Here, we find that users' taste preferences are well separated in general, which means that users tend to seek for different items from each other. Users' interests for niche items (i.e. items with small degree) are found to be highly clustered, which indicates that the selectors of niche items enjoy a high similarity. Furthermore, we investigate the recommendation processes on both the real networks and the reshuffled networks. The results show that the performance of popularity-based recommendation methods do not rely on the real network structure while the performance of personalized recommendation methods is strongly related to it. The recommendation accuracy on each item is studied in detail. We find that most personalized recommendation methods can accurately catch users' taste on *hot items* (i.e. items with large degree) and the reshuffling process does not influence the recommendation accuracy for hot items. However, the accuracy for niche items is relatively low and drops significantly after removing users' behavior patterns. Actually, a robust recommender system should enjoy a stable recommendation accuracy even the user-item networks are deliberately attacked with some random links, our finding suggests that preserving the recommendation accuracy for niche items is significant for enhancing the robustness of the recommender system. Since we reveal an effective direction to improve the accuracy and the robustness of the existing recommender systems, this work is meaningful from the practical point of view.

2. Statistical behavior patterns of online users

In this paper, the datasets that we will use are the subsets of data obtained from four online systems: Movielens (http://www.grouplens.com/), Netflix (http://www.netflixprize.com/), Delicious (http://www.delicious.com/) and Amazon (http://www.amazon.com/). These data are random samplings of the whole records of user activities in these websites, the descriptions of data are given in Table 1.

To investigate users' behavior pattern, we will compare the real bipartite networks with the reshuffled networks. In each step of the reshuffling process, we first randomly pick two links from the real network, for example, one is from user i to item α and the other is from user j to item β (throughout this paper we use Greek and Latin letters, respectively, for item- and user-related indices). Then we rewire these two links by i to β and j to α. Hence, the degree of the users and items would not be changed by this reshuffling process while the links in this reshuffled networks are randomized. Denoting T as the reshuffling times and L as the total links in the networks, we fix $T/L = 3$ in the following analysis.

After the reshuffling process, users' degree and items' degree are preserved while the correlation between users and items are destroyed. To begin our comparison, we focus on the average degree of users' selected items. Suppose a user i selects m items with degree k_α ($\alpha = 1, 2, \ldots, m$), we calculate the average degree of the items that he/she selected as $d_i = \frac{\sum_{\alpha=1}^{m} k_\alpha}{m}$. Actually, the distribution of d reflects the heterogeneity of users' preference. When all the users prefer the same items, users' d will be quite close to each other. Consequently, the distribution of d will be narrow. On the contrary, the distribution of d will be quite flat if all the users seek for different items. We then compare the distribution $P(d)$ in real networks with $P(d)$ in the reshuffled networks. As shown in Fig. 1, $P(d)$ in real networks indeed are much boarder than that in the reshuffled networks, which means the real users exhibit heterogeneous preference in choosing items.

Second, for each user we study the inter-similarity among all his/her selected items. Suppose a user i selects m items and the similarity between items α and β is denoted as $s_{\alpha\beta}$, the inter-similarity among all these m items can be obtained by $\widetilde{S}_i = \frac{2\sum_{\alpha=2}^{m}\sum_{\beta=1}^{\alpha} s_{\alpha\beta}}{m(m-1)}$. The similarity $s_{\alpha\beta}$ of two items is calculated by the common neighbor index which is simply the