Physics Letters A 379 (2015) 2839–2844

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

The effect of heterogeneous dynamics of online users on information filtering

Bo-Lun Chen [a,b,c], An Zeng [d,*], Ling Chen [a,b]

[a] Department of Computer Science, Yangzhou University of China, Yangzhou 225127, China
[b] Department of Computer Science, Nanjing University of Aeronautics and Astronautics of China, Nanjing 210016, China
[c] Department of Physics, University of Fribourg, Chemin du Musee 3, CH-1700 Fribourg, Switzerland
[d] School of Systems Science, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form 16 August 2015
Accepted 5 September 2015
Available online 16 September 2015
Communicated by C.R. Doering

Keywords:
Recommendation
Heterogeneous dynamics
Bipartite networks
Data division

ABSTRACT

The rapid expansion of the Internet requires effective information filtering techniques to extract the most essential and relevant information for online users. Many recommendation algorithms have been proposed to predict the future items that a given user might be interested in. However, there is an important issue that has always been ignored so far in related works, namely the heterogeneous dynamics of online users. The interest of active users changes more often than that of less active users, which asks for different update frequency of their recommendation lists. In this paper, we develop a framework to study the effect of heterogeneous dynamics of users on the recommendation performance. We find that the personalized application of recommendation algorithms results in remarkable improvement in the recommendation accuracy and diversity. Our findings may help online retailers make better use of the existing recommendation methods.

1. Introduction

With the fast development of the World Wide Web, our daily lives depend more and more on the Internet. However, how to find the information we need is not a simple problem [1]. The huge amount of online items such as the movies, books, bookmarks make it impossible for everyone to go over every item and find their favorite. Many approaches such as the collaborative filtering [2,3], matrix factorization [4–6], resource diffusion [7–9] have been intensively investigated recently. This so-called information filtering problem attracts researchers from computer science [10,11], physics [12–14], psychology [15,16], management [17] and so on. The research issues range from the recommendation accuracy [7] and diversity [9] to the sustainability of the whole system in evolution [18]. In this context, many recommendation algorithms have been proposed to help online users filter out irrelevant information and narrow down the search space [19,20].

In the literature, the studies on recommender systems overwhelmingly focus on the recommendation techniques while the effect of online users' features on the recommendation process has received far less attention. Research on human dynamics has shown clearly that the behavior of online users is heterogeneous [21]. At individual level, the inter-event time of item selections exhibits the burst property, i.e. in some period users select items frequently while in some other period the time between two selections of a user can be very long [22]. At system level, the distribution of users activity is very broad, indicating that some users are very active while many other users are much less active [23,24]. Moreover, it has been revealed that online users are driven by different network growth mechanism when they establish new links in the network [25]. In this paper, we focus on how the heterogeneous dynamics of online users affects the information filtering process.

To evaluate the accuracy and diversity of recommendation, usually the real data (i.e. links) are randomly divided into two sets: the training set represents the known historical information that can be used by the recommendation algorithm; the probe set represents the unknown future information that is used to check the quality of the recommendation [1]. This implies that the number of links that the recommendation algorithm tries to predict for each user is proportion the cumulative degree of the user. However, this assumption could be problematic from practical point of view. Since the behavior of online users is very heterogeneous, some users could be very active and requires the recommendation list to be updated very often. Some less active users, on the other hand, may require less frequent update of the recommendation list [26]. Therefore, it is necessary to take into account the heteroge-

* Corresponding author.
E-mail address: anzeng@bnu.edu.cn (A. Zeng).

http://dx.doi.org/10.1016/j.physleta.2015.09.019
0375-9601/

neous dynamics of online users in data division and examine the performance of recommendation algorithms when the amounts of links for prediction are unequal for different users.

In this paper, we propose a heterogeneous data division model for the recommendation process. Instead of randomly dividing the links into the training set and probe set, the number of links for each user in the probe set is determined by the user's degree. The bias towards different degree is controlled by a tunable parameter. By implementing some representative recommendation algorithms, we find that different data division indeed significantly influences the evaluation results of the existing recommendation algorithms. Interestingly, if the number of probe links for large degree users is smaller than that from the random division (i.e. update the recommendation list more frequently for large degree users), the overall recommendation accuracy and diversity are improved.

2. Related work

In the literature, many researchers have considered the heterogeneity of online users when designing recommendation algorithms. Motivated by the observed significant difference between users' structural properties in the network, Zhang et al. proposed to remove redundant links for each user to extract the so-called information backbone [19]. Guan et al. observed that large degree users tend to select niche objects while small degree users tend to select popular objects. They thus proposed a personalized hybrid algorithm in which each user is assigned with a parameter to adjust the popularity of the objects shown in his/her personal recommendation list [27]. Zeng et al. argued that due to users' heterogeneity, they are carrying different amount of information for the recommendation algorithms. Accordingly, Zeng et al. identified some core users in the network and achieve 90% of the accuracy by taking only 20% of the core users' data into account [20]. Similar ideas have been extended to the study of online search engine. Sugiyama et al. proposed a personalized web search engine according to each user's need for relevant information without any user effort [28]. User heterogeneity will result in different information needs for each user's query. Therefore, the search results should be adapted to users with different information needs.

Compared to the recommendation algorithm design, user heterogeneity has less impact on the research on the data division. Most of the recommendation algorithms were validated based on the random data division, which obviously neglects the user heterogeneity in selecting items. In a recent review [1], it is mentioned that recommendation should be done with the data divided into the training set and probe set based on the time stamps on links. Focusing on the over-fitting problems for recommendation algorithms, Zeng et al. proposed a triple data division model in which the real data is divided into a training set, a learning set and a probe set [13]. The basic idea is to estimate users' parameters with the learning set and then applied the learned parameters to actually predict users' future objects in the probe set. Since the purchase behaviors which happened long time ago could not truly reflect the current interests of the target user, Guo et al. investigated the impact of the time window on the training set on the recommender algorithms [29]. In order to improve the diversity and accuracy of the recommender system, Song et al presented an improved hybrid information filtering of adopting the partial recent information in terms of the face that the recent behaviors are more effective to capture the users' potential interests, they also generated a series of training sets, each of which is treated as known information to predict the future links proven by the probe set [23].

Taken together, the research on users' heterogeneity mainly focuses on new algorithm design and the modification on the training set. One crucial direction has not been investigated so far, that is the effect of users' heterogeneity on the probe set. In this paper, we take into account the heterogeneous dynamics of online users and associate it with the length of the probe set (i.e. the number of items they will connect to in the future). We argue that active users and inactive users should have different amount of links in the probe set. This assumption not only helps us to understand better the recommendation process in real system but also provides us with a better implementation of the recommendation algorithms (i.e. update users' recommendation lists with different frequency).

3. Data and model

In this paper, we use two standard data sets which have been widely used to examine the performance of recommendation algorithms [29–31]. The first one is the Movielens data with 1682 movies (items) and 943 users (http://www.grouplens.org/). Users rate movies from 1 (worst) to 5 (best). Consistent with the literature, we consider the ratings higher than 2 as a link. Finally, 82520 links remain in the network. The second one is the Netflix data which is a random sample of the whole records of users ratings in Netflix.com (http://www.netflixprize.com/). It consists of 2294 users, 1891 movies, and 71074 links. Like Movielens, Netflix is also based on a 5-star rating system. With the same rating filtering process, we obtain 59464 links in Netflix data. Throughout this paper, we mainly present the results on Netflix data by figures and the results of both data sets are reported by tables.

In order to model the prediction process of the recommender systems, the above data (i.e. links) are divided into two parts: the training set E^T represents the known information while the probe set E^P represents the unknown information for prediction. Considering the heterogeneous dynamics of users, the division of links into these two sets are not completely random. As active and inactive users spend different amount of time online, their recommendation needs to be updated with different frequency. Some users are very active, their recommendation lists should be updated often. For the less active users, the generated recommendation lists could be used for a relatively longer time. Accordingly, we propose a data division model in which the amount of data in the probe set for each user is tunable. In each step, we randomly pick up a user i with the probability $p_i = k_i^\theta / \sum_j k_j^\theta$. Then one of his/her links is randomly moved to the probe set. The process is terminated when the total amount of links in the probe set reaches 10% of the links in the original network. Here, θ is a tunable parameter. When $\theta = 1$, the data division process reduces to the traditional random data division. When $\theta < 1$, the links connecting to small degree users are more likely to be moved to the probe set, and vice versa.

4. Recommendation algorithms

In this paper, we consider the hybrid recommendation algorithm which combines the Mass diffusion and Heat conduction methods [9]. The user-item bipartite network is characterized by an adjacency matrix A where the element $a_{i\alpha}$ equals to 1 if user i has collected object α, and 0 otherwise. The number of users and items is denoted as N and M, respectively. Consistent with the literature, we use Latin and Greek letters, respectively, for user- and item-related indices. To generate the recommendation list for a specific user i, the Hybrid method starts by assigning each item selected by user i one unit of resource. This resource assignment can be represented by a vector f_i. The resources of these selected items then diffuse in the bipartite network for two steps with the transition matrix W. Each component in this matrix can be computed as