ARTICLE IN PRESS
Physica A ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Physica A

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

Solving the stability–accuracy–diversity dilemma of recommender systems

Lei Hou[a,*], Kecheng Liu[a,b,**], Jianguo Liu[b], Runtong Zhang[c]

[a] *Informatics Research Center, University of Reading, Reading RG6 6UD, United Kingdom*
[b] *Data Science and Cloud Service Research Centre, Shanghai University of Finance and Economics, Shanghai 200433, China*
[c] *School of Economics and Management, Beijing Jiaotong University, Beijing, 100044, China*

HIGHLIGHTS

- The stability–accuracy–diversity dilemma of recommender systems is addressed.
- By considering only the stable similarities, a top-*n*-stability method is proposed.
- The proposed algorithm is proved to be efficient for solving the dilemma.

ARTICLE INFO

Article history:
Received 16 May 2016
Received in revised form 2 September 2016
Available online xxxx

Keywords:
Recommender system
Recommendation stability
Stability–accuracy–diversity dilemma
Heat Conduction algorithm

ABSTRACT

Recommender systems are of great significance in predicting the potential interesting items based on the target user's historical selections. However, the recommendation list for a specific user has been found changing vastly when the system changes, due to the unstable quantification of item similarities, which is defined as the recommendation stability problem. To improve the similarity stability and recommendation stability is crucial for the user experience enhancement and the better understanding of user interests. While the stability as well as accuracy of recommendation could be guaranteed by recommending only popular items, studies have been addressing the necessity of diversity which requires the system to recommend unpopular items. By ranking the similarities in terms of stability and considering only the most stable ones, we present a top-*n*-stability method based on the Heat Conduction algorithm (denoted as TNS-HC henceforth) for solving the stability–accuracy–diversity dilemma. Experiments on four benchmark data sets indicate that the TNS-HC algorithm could significantly improve the recommendation stability and accuracy simultaneously and still retain the high-diversity nature of the Heat Conduction algorithm. Furthermore, we compare the performance of the TNS-HC algorithm with a number of benchmark recommendation algorithms. The result suggests that the TNS-HC algorithm is more efficient in solving the stability–accuracy–diversity triple dilemma of recommender systems.

1. Introduction

Recent decades witnessed the explosive growth of online information, which brings a great deal of information to fit people's preferences. However, the volume of online information is considerably more than any person can possibly pro-

* Corresponding author.
** Corresponding author at: Informatics Research Center, University of Reading, Reading RG6 6UD, United Kingdom.
E-mail addresses: lhou4397@gmail.com (L. Hou), k.liu@henley.ac.uk (K. Liu), liujg004@ustc.edu.cn (J. Liu), rtzhang@bjtu.edu.cn (R. Zhang).

http://dx.doi.org/10.1016/j.physa.2016.10.083
0378-4371/© 2016 Published by Elsevier B.V.

cess [1] which has been characterised as the information overloading problem. Regarding this dilemma, the recommender system [2–4] is one of the powerful tools to offer a solution, by predicting online users' interests in terms of their historical rating or selecting behaviours.

A fundamental idea of the recommender system is to find items that are similar to the target user's historical selections. Therefore, how to properly quantify the similarity between items is of great significance for the recommendation. One of the classical techniques is the content-based method [5,6], which detects similarity by examining whether two items share the same attributes or features. In collaborative tagging systems [7,8] which allow users to freely assign tags on items, similarities can also be quantified using those user-created tags and this method is known as the folksonomy-based similarity [9,10]. However, those external attributes are not always available and, in many scenarios, the data is just binary records of links from users to items. For this kind of systems, using what is normally called association rules, one can mine the co-selecting patterns between items. The fundamental assumption is that two items should be similar if they are selected by the same users. For example, the Common Neighbour index defines the similarity between two items exactly as the number of users who have rated or selected both of them. By considering the popularity information of the items or the activity level of users, some variations of Common Neighbour index have been developed, such as the Salton index, the Jaccard index, the Hub-Promoted index [11], and the Leicht–Holme–Newman index [12]. Given the rapid development of network science [13], the binary relations between users and items could be modelled as the user-item bipartite networks [14,15]. By regarding the users and items as nodes, and the accessing records as the links, the user-item bipartite network therefore simplifies the recommendation problem as predicting possible links between unconnected user-item pairs. Based on the bipartite network, physical spreading processes have been applied to quantify the similarities between nodes (users or items), such as the Resource Allocation index and Mass Diffusion index [15].

Given the enormous efforts on the investigations of similarity quantification, most of the well established recommender systems are based on similarities, such as the Amazon [16,17], the TiVo digital video system [18] and the YouTube [19]. Despite the wide applications and investigations, the similarity-based recommender system are still with challenges, such as the cold start problem [20] and how to evaluate the effect of time [21,22]. Notably, one of the most discussed challenges of the recommender system is how to improve the recommendations' diversity (also referred as the personalisation problem) [23,24]. Since the popular items are welcome by most users, the recommender system could achieve high accuracy by recommending popular items, which will lead to centralised interests. However, the users always have interests that are different with each others', and hence, the recommender systems also should present diverse recommendation lists by digging out dark information. The Heat Conduction algorithm [25,26] is one of the most discussed such methods, which has very high diversity but low accuracy. Regarding to the accuracy–diversity dilemma, numerous efforts have been devoted [27–29]. However, the performances of these algorithms highly depend on the similarity measures.

Recently, the stability property of similarity measures for user-item bipartite networks has been investigated [30]. While similarity measures are trying to reveal the real similarities in terms of the wiring patterns, they should stand in presence of noises such as random wiring. Additionally, although the real similarity between two specific items may evolve in time which leads to the instability, the similarity measures should at least be stable when evaluating with different samples of the network. However, the experimental results indicate that most of the measures may generate totally different evaluations of the similarity when using different samples even from the same period of data. Therefore, a serious question raises that, if the similarity measures are unstable evaluating the similarities, how could one be sure of that the measured similarity is the reflection of real similarity. Furthermore, given the recommender systems highly depending on the similarity quantifications, those unstable similarity measures offering inappropriate quantifications puts the system at risk, *i.e.* the recommendations will also be unstable. Recommendations to become unstable may cause risks such as (1) users finding recommendations unreasonable which leads to bad experiences, and (2) the uncontrollable performance of the recommendation algorithm in practical applications. The stability problem of similarity quantification and recommendation should be of both theoretical and practical concerns. From the theoretical perspective, if the extracted similarity is unstable, it would be hard to evaluate weather a user is interested in an item or not. From the practical perspective, the stability problem would be a gap between laboratory investigation and real-time application because practical systems are always vastly evolving.

Note that, there are some related researches on the stability problem of recommender systems, such as Adomavicius and Zhang (2012) [31,32]. They define the stability as the consistency between the original recommendations and the recommendations using the combination of the historical data and some of the original recommendations (assuming some of the original recommendations have been adopted by the users). The stability of the present paper is defined differently from their study [31,32]. While Adomavicius and Zhang used the output (commendations) of the first recommendation experiment as the input (historical records) of the second recommendation experiment to examine the consistency of the prediction, which could be regarded as the recommendation algorithm's self-consistency, we explore the influence of the users' real behaviour growth on the similarity quantification and recommendation change. Other studies also have discussed the systems' ability to remain stable under malicious attached (records faked for specific purpose) [33,34], which also been referred as the robustness of the recommender systems, while we study the systems' stability with its own natural evolution.

Similar to the accuracy problem, by only recommending popular items, the system could have very high stability. However, the recommender system then falls again into the dilemma that whether should the recommender system recommends popular items to achieve high stability and accuracy or recommends unpopular items to achieve high diversity. So, there rises the triple dilemma of stability–accuracy–diversity. By ranking the similarities in terms of stability and only