Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

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

Diffusion-like recommendation with enhanced similarity of objects

Ya-Hui An [a], Qiang Dong [a,b,*], Chong-Jing Sun [b], Da-Cheng Nie [a], Yan Fu [a,b]

[a] School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 611731, PR China

[b] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, PR China

HIGHLIGHTS

- The distributions of RA similarity follow almost power-law.
- The RA similarity is a key factor in measuring the resource transfer between objects.
- The distribution of ERA similarity is more even, thus the similarity of a large proportion pairs of unpopular objects is increased.
- The enhancement of similarity greatly improves both of accuracy and diversity of diffusion-like models.
- The proposed method can be analogously applied to most similarity-based recommendation models on user–object bipartite networks.

ARTICLE INFO

Article history:
Received 14 December 2015
Received in revised form 31 March 2016
Available online xxxx

Keywords:
Recommender systems
Bipartite networks
Resource-allocation similarity
Diffusion-like algorithms

ABSTRACT

In the last decade, diversity and accuracy have been regarded as two important measures in evaluating a recommendation model. However, a clear concern is that a model focusing excessively on one measure will put the other one at risk, thus it is not easy to greatly improve diversity and accuracy simultaneously. In this paper, we propose to enhance the Resource-Allocation (RA) similarity in resource transfer equations of diffusion-like models, by giving a tunable exponent to the RA similarity, and traversing the value of this exponent to achieve the optimal recommendation results. In this way, we can increase the recommendation scores (allocated resource) of many unpopular objects. Experiments on three benchmark data sets, MovieLens, Netflix and RateYourMusic show that the modified models can yield remarkable performance improvement compared with the original ones.

1. Introduction

Nowadays, the explosive growth of storage capability has allowed people to collect and store almost all the information generated every day. This so-called "*big data*" provides great benefit to our life, e.g., we can easily get the cast list of a niche film via search engines. However, we find that it becomes very difficult to find the relevant movie of our interest from countless candidates, if we cannot describe it by appropriate keywords. On this *information overload* occasion, *recommender systems* arise to help us make the right decisions.

Different from search engines requiring keywords, recommender systems are designed to uncover users' potential preferences and interests based on users' past activities and profile descriptions, and accordingly deliver a personalized list of

* Corresponding author at: School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 611731, PR China.
E-mail address: dongq@uestc.edu.cn (Q. Dong).

http://dx.doi.org/10.1016/j.physa.2016.06.027
0378-4371/

Table 1
The basic statistics of three data sets, where n, m and q denote the number of users, objects and edges, respectively; $\langle k_u \rangle$ and $\langle k_o \rangle$ are the average degrees of users and objects.

Data set	n	m	q	$\langle k_u \rangle$	$\langle k_o \rangle$
MovieLens	943	1682	100,000	106	59.5
Netflix	10,000	5640	701,947	70.2	124.5
RYM	33,762	5267	675,817	20	128.3

recommended objects to every user. In the last decade, recommender systems have become a significant issue in both academic and industrial communities. The early recommender models are based on a simple observation that similar users are likely to purchase the same items, or, the items collected by the same user are prone to be similar to each other, such as collaborate filtering [1] and content-based methods [2]. These methods are shown to give accurate recommendation results, but they confront a recommender system with the risk that more and more users will be exposed to a narrowing band of popular items, leading to poor diversity among users' recommendation lists [3,4].

Given this, many other recommendation models have been proposed in the literature, including dimensionality reduction techniques [5–8], diffusion-like methods [9–12], social filtering [13–15], and hybrid recommendation models [5,16–18]. However, people found that accuracy and diversity seem to be two sides of the seesaw: when one side rises, the other side falls. Examples are two primary diffusion-like methods, ProbS [9] and HeatS [19], which mimic two basic physical processes on user–object bipartite networks. ProbS is demonstrated to give recommendation results with good accuracy but poor diversity, while HeatS is found to be effective in providing a diverse recommendation lists at the cost of accuracy.

This diversity–accuracy dilemma has received considerable research attentions in the field of recommender systems. Zhou et al. [5] designed delicately a nonlinear hybrid model of HeatS and ProbS, called HHP, which achieves significant improvements in both accuracy and diversity of recommendation results. Another two effective methods modified respectively from original ProbS and HeatS, named Preferential Diffusion (PD) [20] and Biased Heat Conduction (BHC) [21], also make a good trade off on accuracy and diversity. Based on BHC and HHP, Qiu et al. [22] proposed the Heterogeneous Heat Conduction (HHC) model which takes into account the heterogeneity of the source objects. Nie et al. [11] investigated the optimal hybrid coefficients of HeatS and ProbS, and proposed accordingly the Balance Diffusion (BD) model, respectively.

In the first diffusion-like model ProbS, every user distributes the total resource he receives previously from objects, back averagely to his neighbor objects. The niche objects will receive lower final resources (recommendation scores) because they have fewer neighbor users (resource portals), thus rank in the bottom of the recommendation lists. That is why ProbS suffers from poor diversity. In view of this, the PD model proposed by Lv et al. [20] intentionally allocates more resource to small-degree objects, and less resource to large-degree objects. For the resource of a given user, every neighbor object receives the percentage approximately inversely proportional to its degree. Compared with ProbS, PD simultaneously improves the diversity and accuracy of recommendation results.

With the similar motivation, we proposed a method which can be used in many being diffusion-like algorithms, like ProbS, BHC and HHP, to get better recommendation results on user–object bipartite networks. We propose to enhance the RA similarity in the transfer equations of diffusion-like models, by giving a tunable exponent on the shoulder of RA similarity, and traverse this parameter to achieve the optimal recommendation results. Experiments on three benchmark data sets, MovieLens, Netflix, and RYM (Rate Your Music) show that our model can yield a great performance improvement compared with the original models.

2. Materials and methods

In this paper, a recommender system is represented by a bipartite network $G(U, O, E)$, where $U = \{u_1, u_2, \ldots, u_m\}$, $O = \{o_1, o_2, \ldots, o_n\}$ and $E = \{e_1, e_2, \ldots, e_q\}$ correspond to m users, n objects and q edges between users and objects, respectively. This bipartite network could be fully described by an adjacency matrix $A = \{a_{l\alpha}\}_{m\times n}$, where the element $a_{l\alpha} = 1$ if there exists an edge between user u_l and object o_α (user u_l *collects* object o_α), meaning that user u_l declared explicitly his preference on object o_α in the past, and $a_{l\alpha} = 0$ otherwise. For every target user, the essential task of a recommender system becomes to recommend him a sublist of uncollected objects of his potential interest.

2.1. Data set description

Three real-world data sets are adopted to test the recommendation result, namely, MovieLens, Netflix and RYM (Rate Your Music). Here we will briefly describe these three data sets. MovieLens, a movie rating data set, was collected by the GroupLens Research Project at the University of Minnesota and can be found at the website www.grouplens.org. Netflix, a randomly sampled subset of the huge data set provided by the Netflix company for the Netflix Prize (www.netflixprize.com) [23]. RYM, a music rating data set, is obtained by downloading publicly available data from the music ratings website www.RateYourMusic.com [5]. In this paper, we make use of nothing but the binary information whether there exists an interaction between a user and an object in the past. The basic statistics of the three data sets are presented in Table 1.

Please cite this article in press as: Y.-H. An, et al., Diffusion-like recommendation with enhanced similarity of objects, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.06.027