Physica A 388 (2009) 462–468

Contents lists available at ScienceDirect

Physica A

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

Personal recommendation via modified collaborative filtering

Run-Ran Liu [a], Chun-Xiao Jia [a], Tao Zhou [a,b,*], Duo Sun [a], Bing-Hong Wang [a,c]

[a] Department of Modern Physics and Nonlinear Science Center, University of Science and Technology of China, Hefei Anhui, 230026, PR China
[b] Department of Physics, University of Fribourg, Chemin du Muse 3, CH-1700 Fribourg, Switzerland
[c] Institute of Complex Adaptive System, Shanghai Academy of System Science, Shanghai, PR China

A R T I C L E I N F O

Article history:
Received 8 January 2008
Received in revised form 27 September 2008
Available online 17 October 2008

PACS:
89.75.Hc
87.23.Ge
05.70.Ln

Keywords:
Recommendation system
Bipartite network
Similarity
Collaborative filtering
Infophysics

A B S T R A C T

In this paper, we propose a novel method to compute the similarity between congeneric nodes in bipartite networks. Different from the standard cosine similarity, we take into account the influence of a node's degree. Substituting this new definition of similarity for the standard cosine similarity, we propose a modified collaborative filtering (MCF). Based on a benchmark database, we demonstrate the great improvement of algorithmic accuracy for both user-based MCF and object-based MCF.

1. Introduction

Recently, recommendation systems are attracting more and more attention, because it can help users to deal with information overload, which is a great challenge in modern society, especially under the exponential growth of the Internet [1] and the World-Wide-Web [2]. Recommendation algorithms have been used to recommend books and CDs at Amazon.com, movies at Netflix.com, and news at VERSIFI Technologies (formerly AdaptiveInfo.com) [3]. The simplest algorithm we can use in these systems is the global ranking method (GRM) [4], which sorts all the objects in descending order of degree and recommends those with the highest degrees. GRM is not a personal algorithm and its accuracy is not very high because it does not take personal preferences into account. Accordingly, various kinds of personal recommendation algorithms are proposed, for example, collaborative filtering (CF) [5,6], content-based methods [7,8], spectral analysis [9,10], principal component analysis [11], the diffusion approach [4,12–14], and so on. However, the current generation of recommendation systems still requires further improvements to make recommendation methods more effective [3]. For example, content analysis is practical only if the items have well-defined attributes and those attributes can be extracted automatically; for some multimedia data, such as audio/video streams and graphical images, the content analysis is hard to apply. Collaborative filtering usually provides very bad predictions/recommendations to new users having very few collections. Spectral analysis has high computational complexity and is thus infeasible to deal with huge-size systems.

* Corresponding author at: Department of Modern Physics and Nonlinear Science Center, University of Science and Technology of China, Hefei Anhui, 230026, PR China.
E-mail address: zhutou@ustc.edu (T. Zhou).

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.10.010

Thus far, the widest applied personal recommendation algorithm is CF [3,15]. CF has two categories in general, one is user-based (U-CF), which recommends the target user the objects collected by the users sharing similar tastes; the other is object-based (O-CF), which recommends those objects similar to the ones the target user preferred in the past. In this paper, we introduce a modified collaborative filtering (MCF), which can be implemented for both object-based and user-based cases and achieve much higher accuracy of recommendation.

2. Method

We assume that there is a recommendation system which consists of m users and n objects, and each user has collected some objects. The relationship between users and objects can be described by a bipartite network. Bipartite network is a particular class of networks [4,16], whose nodes are divided into two sets, and connections among one set are not allowed. We use one set to represent users, and the other represents objects: if an object o_i is collected by a user u_j, there is an edge between o_i and u_j, and the corresponding element a_{ij} in the adjacent matrix A is set as 1, otherwise it is 0.

In U-CF, the predicted score v_{ij} (to what extent u_j likes o_i), is given as:

$$v_{ij} = \sum_{l=1, l\neq i}^{m} s_{il} a_{jl}, \tag{1}$$

where s_{il} denotes the similarity between u_i and u_l. For any user u_i, all v_{ij} are ranked by values from high to low, objects on the top and have not been collected by u_i are recommended.

How to determine the similarity between users? The most common approach taken in previous works focuses on the so-called structural equivalence. Two congeneric nodes (i.e. in the same set of a bipartite network) are considered structurally equivalent if they share many common neighbors. The number of common objects shared by users u_i and u_j is

$$c_{ij} = \sum_{l=1}^{n} a_{li} a_{lj}, \tag{2}$$

which can be regarded as a rudimentary measure of s_{il}. Generally, the similarity between u_i and u_j should be somewhat relative to their degrees [17]. There are at least three ways previously proposed to measure similarity, as:

$$s_{ij} = \frac{2c_{ij}}{k(u_i) + k(u_j)}, \tag{3}$$

$$s_{ij} = \frac{c_{ij}}{\sqrt{k(u_i)k(u_j)}}, \tag{4}$$

$$s_{ij} = \frac{c_{ij}}{\min(k(u_i), k(u_j))}. \tag{5}$$

The Eq. (3) is called Sorensen's index of similarity (SI) [18], which was proposed by Sorensen in 1948; the Eq. (4), called the cosine similarity, was proposed by Salton in 1983 and has a long history of the study on citation networks [17]. Both the Eqs. (4) and (5) are widely used in recommendation systems [3,4].

A common problem of Eqs. (3)–(5) is that they have not taken into account the influence of an object's degree, so that objects with different degrees have the same contribution to the similarity. If user u_i and u_j both have selected object o_l, that is to say, they have a similar taste for the object o_l. Provided that object o_l is very popular (the degree of o_l is very large), this taste (the favor for o_l) is a very ordinary taste and it does not mean u_i and u_j are very similar. Therefore, its contribution to s_{ij} should be small. On the other hand, provided that object o_l is very unpopular (the degree of o_l is very small), this taste is a peculiar taste, so its contribution to s_{ij} should be large. In other words, it is not very meaningful if two users both select a popular object, while if a very unpopular object is simultaneously selected by two users, there must be some common tastes shared by these two users. Accordingly, the contribution of object o_l to the similarity s_{ij} (if u_i and u_j both collected o_l) should be negatively correlated with its degree $k(o_l)$. We suppose the object o_l's contribution to s_{ij} being inversely proportional to $k^\alpha(o_l)$, with α a freely tunable parameter. The s_{ij}, consisted of all the contributions of commonly collected objects, is measured by the cosine similarity as shown in Eq. (4). Therefore, the proposed similarity reads:

$$s_{ij} = \frac{1}{\sqrt{k(u_i)k(u_j)}} \sum_{l=1}^{n} \frac{a_{li} a_{lj}}{k^\alpha(o_l)}. \tag{6}$$

Note that, the influence of an object's degree can also be embedded into the other two forms, shown in Eqs. (3) and (5), and the corresponding algorithmic accuracies will be improved too. Here in this paper, we only show the numerical results on cosine similarity as a typical example.

For any user-object pair u_i-o_j, if u_i has not yet collected o_j, the predicted score can be obtained by using Eq. (1). Here we do not normalize Eq. (1), because it will not affect the recommendation list, since for a given target user, we need sort all