ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31, No. 6, Nov. 2016, pp. 1234－1241
DOI:10.16337/j.1004-9037.2016.06.021

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于评分预测的协同过滤推荐算法

周海平[1,2]　黄凌英[3]　刘　妮[2]　周洪波[2]

(1.绍兴文理学院计算机科学与工程系,绍兴,312000; 2.贵阳学院数学与信息科学学院,贵阳,550005;3.湖南大学信息科学与工程学院,长沙,410082)

摘　要:传统的基于评分预测的协同过滤算法在计算用户之间相似性时只针对用户共同评过分的物品计算评分差异,然而由于不同用户共同评分的物品数目不同,使得计算标准不统一,从而导致推荐质量不理想。本文在传统算法的基础上进行改进,新算法在计算相似性的时候一方面考虑了用户共同评分的物品数,另一方面还考虑了物品的热门程度对用户相似性计算的影响。实验结果表明,新算法在推荐准确率和召回率上都比传统算法提高了1倍以上。研究还发现在算法中使用Pearson相关系数明显好于使用欧氏距离作相似性度量标准得到的推荐效果。

关键词:推荐系统;协同过滤;相似性;召回率;准确率

中图分类号:TP391　　**文献标志码**:A

Collaborative Filtering Recommendation Algorithm Based on Rating Prediction

Zhou Haiping[1,2], Huang Couying[3], Liu Ni[2], Zhou Hongbo[2]

(1. Department of Computer Science and Engineering, Shaoxing University, Shaoxing, 312000, China; 2. Department of Mathematic and Information Science, Guiyang College, Guiyang, 550005, China; 3. College of Information Science and Engineering, Hunan University, Changsha, 410082, China)

Abstract: Traditional collaborative filtering algorithm calculates the difference of scores only for the common items of users while calculating the similarity of users. Owing that the numbers of common items of different users is not the same, the recommendation quality is not reliable. We proposed a new algorithm, taking both the number of common items and the popularity of goods into consideration while calculating the similarity of users. Experimental results show that, the recommendation quality of new algorithm is improved by more than one time than traditional algorithm in both precision and recall. In addition, results also show that using Pearson correlation as similarity metric obtained higher recommendation quality than Euclidean distance.

Key words: recommendation system; collaborative filtering(CF); similarity; recall; precision

引　　言

随着互联网技术的进步,过去很多只能在线下发生的行为都可以在线上进行,一旦这些行为在线上

基金项目:国家自然科学基金(11247286)资助项目;贵州省自然科学基金(黔科合J字LGK[2013]53号,黔科合LH字[2014]7210号,黔科合LH字[2015]7294号)资助项目;贵州省教育厅"网络信息科学"创新团队资助项目。
收稿日期:2015-05-05;**修订日期**:2016-01-29

发生,便被系统记录下来,这使得互联网收集了大量人类活动的历史数据[1-2]。合理地利用这些历史数据可以给人们的生活和工作带来极大的便利,推荐系统就是在这个背景下诞生的。当前很多网站都提供了推荐功能,例如,Amazon、当当网、淘宝网和京东等大型购物网站都提供了各式各样的推荐服务,推荐质量的好坏决定着一个网站能否吸引并留住更多的用户[3-6]。为了提高推荐质量,每个网站都会根据不同用户的需求采用多种方法进行商品推荐。传统的推荐算法包括基于内容的推荐[7]、基于关联规则的推荐、协同过滤推荐和混合推荐算法等[8-9]。近年来,国内外又出现了各种新的算法,如文献[10,11]将物质扩散和热传导方法用于推荐系统,文献[12]提出了基于网络拓扑特性的推荐算法,这些算法从技术上来看各有特色,都能有效地提高推荐质量。从历史演进的角度看,推荐系统所使用的算法越来越多,也越来越复杂,优秀的推荐系统一般都综合使用了多种算法,但总体来看,当前推荐系统从技术角度上使用得最多的算法有基于内容的推荐算法和协同过滤推荐算法。基于内容的推荐算法主要是对用户曾经选择或购买过的产品进行分析,提取这些产品的特征,然后将具有相似特征的产品推荐给用户,这种算法比较好理解,也很容易实现,尽管如此,这种算法仍然存在许多缺点:(1)这种推荐算法对结构化数据的推荐比较有效,而对于非结构化数据,如图片、音视频等流媒体数据,该技术并不大适用;(2)该技术提取出的用户特征无法量化,例如,系统可以通过历史数据断定某个用户喜欢摇滚音乐,但是无法得知其喜欢的程度。(3)该技术只针对已有的历史数据刻画用户的特征,却不能挖掘用户的潜在兴趣。为了实现对非结构化数据的推荐,Goldberg 等[8]提出了协同过滤推荐算法,该算法通过某个用户对购买过的物品的评分值,找出与该用户具有相似购买记录或评分值的其他用户,这些用户被称为邻居用户,然后对邻居用户的购买记录进行分析,从中找出他们共同购买过的物品并把它推荐给原始用户。

协同过滤算法一方面能够帮助用户发现与其兴趣一致的邻居用户,另一方面能够帮助用户发掘其感兴趣的潜在物品。虽然协同过滤算法在某种程度上弥补了基于内容推荐算法的缺陷,但它仍然面临一些需要解决的问题,例如,在计算用户之间的相似性时,传统算法必须首先提取他们共同评价过的物品,再对这些物品的评分进行比较,如果共同评价过的物品数太少,计算得到的相似性就不准确。此外,传统的协调过滤算法在计算用户相似性时没有考虑物品热门程度不同对相似性指标的影响。基于以上原因,本文对传统评分预测的协同过滤推荐算法存在的不足进行分析,然后对相似性计算模型进行改进,以提高系统的推荐质量。

1 传统基于评分的协同过滤算法

协同过滤算法的原理为如果两个用户同时喜欢某些产品,则说明他们对物品的偏好是相似的,并且这种相似性可以扩展到他们对其他产品的态度。因此在使用协同过滤算法时,首先需要根据历史数据计算用户之间的相似性,然后利用相似性找到目标用户的邻居,并根据邻居用户的评分记录预测目标用户对未购买物品的兴趣,最后对预测结果排序并给出推荐列表。

1.1 相似性度量方法

要找出目标用户的邻居就必须计算用户之间的相似性,然后将相似性最大的用户当做目标用户的邻居,因此相似性的计算过程显得非常重要,该过程直接影响整个推荐系统的质量。整个系统的评分数据集用一个 $m\times n$ 的矩阵表示,m 为用户数,n 为物品数,第 i 行第 k 列的元素 $r_{i,k}$ 代表用户 i 对物品 k 的评分,用户评分数据矩阵如表 1 所示。要度量用户 i 和用户 j 之间的相似性,首先需要找出用户 i 和用户 j 共同评过分的物品,然后利用相似性度量方法计算用户 i 和用户 j 之间的相似性,记为 $\mathrm{sim}(i,j)$。常见的相似性度量方法有欧式距离和 Pearson 相似度等。

表 1 用户评分数据矩阵

Tab. 1 Rating matrix of user

	物品 1	…	物品 k	…	物品 n
用户 i	$r_{i,1}$	…	$r_{i,k}$	…	$r_{i,n}$
⋮	⋮	⋮	⋮	⋮	⋮
用户 j	$r_{j,1}$	…	$r_{j,k}$	…	$r_{j,n}$
⋮	⋮	⋮	⋮	⋮	⋮
用户 m	$r_{m,1}$	…	$r_{m,k}$	…	$r_{m,n}$

(1)欧式距离。欧式距离通过计算欧式空间中两个点之间的距离来衡量两个用户之间的相似性。假设用户 i 和用户 j 共同评分的项目集合用 $I_{i,j}$ 表示,则用户 i 和用户 j 之间的欧式距离 $\mathrm{dist}(i,j)$ 为

$$\mathrm{dist}(i,j)=\sqrt{\sum_{c\in I_{i,j}}(r_{i,c}-r_{j,c})^2} \tag{1}$$

显然,欧式距离的值越小表示两个用户越相似,当欧式距离的值为 0 时,相似度的值应该为 1,因此在计算相似性时,常用式(2)代替式(1),则

$$\mathrm{sim}(i,j)=\frac{1}{1+\mathrm{dist}(i,j)} \tag{2}$$

(2)Pearson 相关系数。Pearson 系数即相关分析中的相关系数,该计算方法能反映两个用户对共同物品评分的相关性,则

$$\mathrm{sim}(i,j)=\frac{\sum_{c\in I_{i,j}}(r_{i,c}-\overline{r_i})(r_{j,c}-\overline{r_j})}{\sqrt{\sum_{c\in I_{i,j}}(r_{i,c}-\overline{r_i})^2}\cdot\sqrt{\sum_{c\in I_{i,j}}(r_{j,c}-\overline{r_j})^2}} \tag{3}$$

式中:$r_{i,c}$ 和 $r_{j,c}$ 分别为用户 i 和用户 j 对物品 c 的评分,$I_{i,j}$ 为用户 i 和用户 j 共同评过分的物品集合。Pearson 相关系数的取值范围为[−1,1],相关系数的绝对值越大,则表明两个用户评分的相关度越高。

1.2 评分预测和推荐质量评价

目标用户的邻居选好之后,就可以利用邻居用户的评分数据预测目标用户对物品的评分。假设目标用户为 i,其邻居用户的集合用 N_i 表示,则可以利用式(4)预测用户 i 对物品 c 的评分,其中 $\delta_{j,c}$ 表示用户 j 是否对物品 c 评过分,如果用户 j 对物品 c 评过分,其值为 1,否则为 0。有

$$r_{i,c}=\frac{\sum_{j\in N_i}\mathrm{sim}(i,j)\cdot r_{j,c}}{\sum_{j\in N_i}\mathrm{sim}(i,j)\cdot\delta_{j,c}} \tag{4}$$

评价推荐质量好坏的的一个重要指标是命中率,它反映了推荐列表中的物品有多少是用户真正的需求。命中率通常包含准确率和召回率两个指标。令 $R(u)$ 为利用推荐算法对目标用户生成的推荐列表,而 $T(u)$ 为用户在测试集上实际评价过的物品列表,则召回率定义为

$$\mathrm{Recall}=\frac{\sum^{u\in U}|R(u)\cap T(u)|}{\sum^{u\in U}|T(u)|} \tag{5}$$

准确率定义为

$$\mathrm{Precision}=\frac{\sum^{u\in U}|R(u)\cap T(u)|}{\sum^{u\in U}|R(u)|} \tag{6}$$

2 算法改进

传统的基于评分的协同过滤推荐算法在计算两个用户的相似度时必须首先找出这两个用户共同评分的物品交集,然后在交集的基础上根据评分值的差异计算这两个用户的相似度。显然,只有当两个用户共同评分的物品数较多时,相似度的计算结果才比较可靠,而在大型的电子商务网站中,一个用户能够评分的物品数非常有限,一般情况下都不到1‰,而两个用户共同评价过的物品就更加稀少,假设用户 i 和用户 j 共同评价过的物品只有一个,就算他们的评分完全一样,也不能由此断定这两个用户具有很高的相似性,但是根据传统算法进行计算,这两个用户的相似度将为1。如果两个用户共同评价过10个物品,其中有5个物品的评分完全一样,根据传统的相似性计算方法,这两个用户的相似度要小于1,但是,这两个用户的相似度更可靠。因此,传统的基于评分的协同过滤推荐算法在共同评分物品不足的情况下难以保证推荐质量,为了弥补这个缺陷,需要对其相似性的计算方法进行修正。根据前面的分析,发现两个用户的相似性不仅与他们对物品评分值的差异程度有关,而且还与他们共同评价过的物品数目有关,如果两个用户在商务网站上购买了很多同样的物品,则意味着他们在品味上具有较高的相似性。因此,在计算相似性时需要考虑两个用户共同评价的物品数,改进后的相似性如式(7)所示,其中 $|I_{i,j}|$ 表示用户 i 和用户 j 共同评价过的物品数,而 $\mathrm{sim}(i,j)$ 为传统的相似性度量指标,则

$$\mathrm{sim}(i,j)' = \mathrm{sim}(i,j) \cdot |I_{i,j}| \tag{7}$$

式(7)虽然在计算两个用户的相似性时考虑了这两个用户共同评价过的物品数目,但该模型并没有对不同物品对相似性的贡献加以区分。例如,两个用户都对两个物品进行了评价,如果这两个物品其中一个是热门物品,另一个是冷门物品,显然,这两个物品对相似性的贡献不同。对于热门物品,由于其名气大、质量好,评价的人很多,而且基本上都是好评,如果两个用户对这样的物品同时给出好评很自然,并不代表这两个用户本质上有多相似,所以用户对热门物品的评价对相似性贡献并不大,相反,如果两个用户同时对某个冷门物品给出了一致的评价,则更能反映出这两个用户品味的一致性,因此,冷门物品比热门物品更能反映两个用户是否相似。因此,本文提出了第2种相似性的改进模型为

$$\mathrm{sim}(i,j)'' = \frac{\sum\limits_{c \in N_i \cap N_j} \frac{1}{1+\log(1+|r_{ic}-r_{jc}|)} \cdot \frac{1}{\log(1+|N_c|)}}{\sqrt{|N_i||N_j|}} \tag{8}$$

式中:N_i 和 N_j 分别为用户 i 和 j 评价过的物品集合;N_c 为评价过物品 c 的用户集合;$\frac{1}{1+\log(1+|r_{ic}-r_{jc}|)}$ 反映了两个用户对物品评分差异对相似性的贡献;而 $\frac{1}{\log(1+|N_c|)}$ 反映了物品热门程度对用户之间相似性的贡献;$|N_c|$ 为评价过物品 c 的人数;物品热门程度越高,$|N_c|$ 的值就越大,$\frac{1}{\log(1+|N_c|)}$ 的值也就越小,其对用户之间相似性的贡献也就越小;反之,物品越冷门,$|N_c|$ 的值就越小,$\frac{1}{\log(1+|N_c|)}$ 的值就越大,其对用户之间相似性的贡献也就越大。

3 实验结果及分析

本文采用MovieLens电影推荐系统中的数据集进行实验,该系统提供了有关影片信息及用户评分的数据集。数据集采用的评分制为5分制,取1~5的整数,打1分表示用户认为该部影片质量不好,打5分表示用户认为该部影片质量非常好。本实验选取的数据集包含943名用户对1 682部影片的10万

条评分记录，并且其中每位用户至少为 20 部影片做过评价。从评分数据集中随机抽取 80%的评分记录当做训练集，抽取 20%的评分记录当做测试集。利用训练集中的数据运行协同过滤算法，然后将得到的预测结果与测试集中的数据进行比较。分别用召回率和准确率来对算法的运行效果进行分析和比较，两种改进算法的效果分别见实验结果 1 和实验结果 2。

3.1 实验结果 1

首先将第 1 种改进算法与传统的基于评分预测的协同过滤推荐算法进行比较，分别以欧式距离和 Pearson 相关性作为相似性度量标准，邻居个数从 5 增加到 60，间隔为 5，为每个目标用户给出预测评分最高的 25 部影片，用于计算召回率和准确率。将计算结果与本文改进过的协同过滤推荐算法作比较，实验结果如图 1～4 所示。

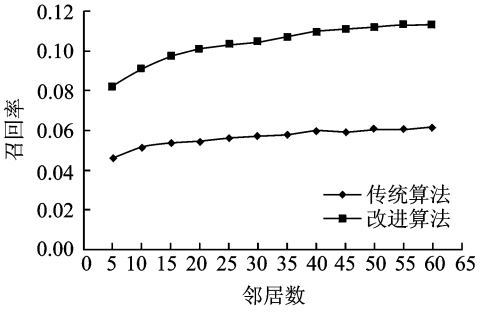

图 1 传统算法与改进算法 1 召回率的对比（采用欧氏距离相似性度量标准）

Fig. 1 Comparison in recall between traditional algorithm and first improved algorithm (Euclidean distance as similarity measure)

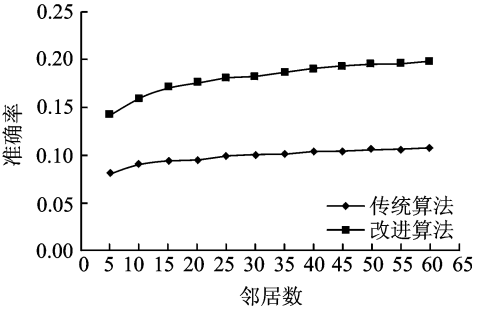

图 2 传统算法与改进算法 1 准确率的对比（采用欧氏距离相似性度量标准）

Fig. 2 Comparison in precision between traditional algorithm and first improved algorithm (Euclidean distance as similarity measure)

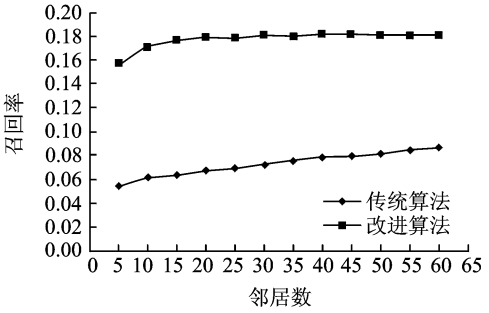

图 3 传统算法与改进算法 1 召回率的对比（采用 Pearson 相似性度量标准）

Fig. 3 Comparison in recall between traditional algorithm and first improved algorithm (Pearson correlation as similarity measure)

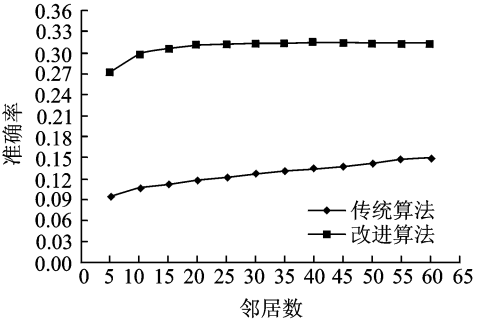

图 4 传统算法与改进算法 1 准确率的对比（采用 Pearson 相似性度量标准）

Fig. 4 Comparison in precision between traditional algorithm and first improved algorithm (Pearson correlation as similarity measure)

由图 1～4 可以看出,改进的算法不论是在召回率还是准确率上都比传统基于评分的协同过滤算法提升了 1 倍以上。另外将图 3,4 与图 1,2 进行对比可知,采用 Pearson 相似性度量标准比采用欧氏距离度量标准在准确率和召回率上表现要好。

3.2 实验结果 2

图 5,6 显示了改进算法 2 和改进算法 1 的对比结果,显然,不论是准确率还是召回率,改进算法 2 都比改进算法 1 表现好很多,这说明降低热门物品对用户相似性的贡献能够带来推荐质量的提高。

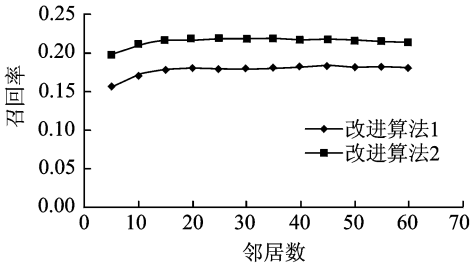

图 5 两种改进算法在召回率上的比较

Fig. 5 Comparison in recall between two improved algorithms

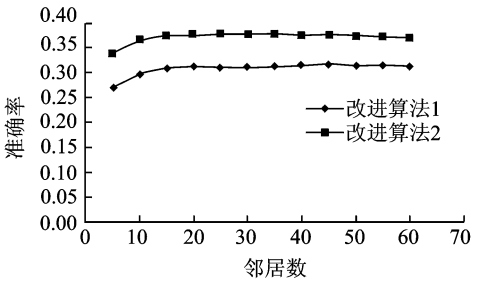

图 6 两种改进算法在准确率上的比较

Fig. 6 Comparison in precision between two improved algorithms

3.3 结果分析

传统的基于评分预测的协同过滤推荐算法在计算用户之间相似度时是基于共同评分的物品进行计算的,由于不同的用户之间共同评分的物品数不一样,所以进行比较的前提条件也就不一致,有的邻居用户与目标用户共同评过分的影片可能只有一部,因此只要他们对这一部影片的评分比较一致,就可以获得很高的相似度,而有的邻居用户可能与目标用户共同评过分的影片很多,但只要有少部分评分不一致,都可能获得不高的相似度,显然这种计算方法有失公允,它忽略了共同评过分的影片数量在相似性计算中的作用,因此改进后的算法更加合理,实验结果 1 也证实了这一点。另外,从实验结果 2 发现,在计算两个用户的相似性时不能只考虑这两个用户共同评价过的物品数,不同的物品如果其热门程度不一样,它对相似性的贡献不同,质量好的热门物品大家都喜欢,所以它对用户相似性的贡献小,而冷门物品更能够反映两个用户的相似性程度,因此在计算相似性时给冷门物品应赋予更高的权重,实验结果 2 充分验证了这一点。从实验结果 1 中还发现:Pearson 相似性在实验中比欧氏距离表现要好。主要原因是欧氏距离在处理评分这类主观性很强的问题时无法反映用户之间的相关性。例如,假设两个用户对同一件物品评价都很好,但是其中一个用户对物品的打分很宽松,习惯对评价好的物品打满分,而另一个用户对物品的打分很苛刻,对评价好的物品最多只打 4 分(满分为 5 分),这样就算这两个用户对很多物品的观点非常相似,但反映在欧氏距离的相似性指标上也不会很高。在这一点,Pearson 相关系数具有明显的优势,由于 Pearson 相关系数反映的是两个用户对物品评分的相关性,就算他们对物品的评分不完全一样,只要他们对物品在整体上的评分趋势保持一致,就认为这两个用户是相似用户,因此,在这种情况下采用 pearson 相关系数的推荐效果要好于欧氏距离,这一点在实验中得到了充分验证。最后,本文将改进算法与传统算法进行综合比较,表 2 展示了各种算法的推荐质量效果对比情况。从表 2 中的数据可以看出,本文的改进算法在推荐准确率和召回率上明显优于传统协同过滤(Collaboration filtering,CF)算法,在数据挖掘和信息推荐领域中,除了协同过滤算法之外,其他一些比较复杂的算法也经

常被使用,其中潜层语义模型(Probabilistic matrix factorization,PFM)是公认的推荐质量比较好的一种算法[13],从表 2 中可以看到,本文所提出的算法在召回率和准确率上仍然略优于 PFM 算法,考虑到 PFM 算法涉及到矩阵分解这种比较复杂的运算,当数据集比较大时,其运算速度下降很快,因此本文的改进算法相对来说更加经济实用。

表 2 本文算法与 PFM 算法的比较

Tab. 2 Comparison between PFM and proposed algorithm

算法	准确率	召回率
CF 算法($K=25$,采用欧式距离相似性度量标准)	0.098	0.056
CF 算法($K=25$,采用 Pearson 相似性度量标准)	0.120	0.069
PFM 算法($F=100$)	0.305	0.169
本文改进算法 1($K=25$)	0.312	0.178
本文改进算法 2($K=25$)	0.379	0.217

4 结束语

本文通过对传统的基于评分的协同过滤推荐算法进行分析,找出了其在计算相似性中存在的缺陷,然后提出了改进的协同过滤算法,新算法一方面考虑了用户的评分值和评分物品数,另一方面考虑了物品的热门程度对相似性的影响,实验结果显示,新算法在召回率和准确率这两个指标上都表现得比传统算法要好。在实际情况中,影响推荐质量的因素还有很多,如影片的风格、用户的年龄和性别等因素都是影片推荐系统需要关注的内容,而这些因素很难在一种算法中全部考虑,事实上,在真实的推荐系统中,研究人员一般会根据具体的应用场景融合各种不同的推荐算法以提高推荐质量。此外,推荐质量的评价指标除了本文使用的命中率之外,还有新颖性、流行度和覆盖率等多种指标,如何综合考虑这些因素提高推荐质量有待于进一步研究。

参考文献:

[1] Jones K, Leonard L. Trust in consumer to consumer electronic commerce[J]. Information Management, 2008, 45(2): 88-95.

[2] Borchers A, Herlocker J, Konstan J, et al. Ganging up on information overload computer,[J]. Computer, 1998, 31(4): 106-108.

[3] Resnick P, Varian H R. Recommender systems[J]. Communications of the ACM, 1997, 40(3): 56-58.

[4] Zenebe A, Norcio A F. Representation, similarity measures and aggregation methods using fuzzy sets for content based recommender systems[J]. Fuzzy Sets and Systems, 2009, 160(1): 76-94.

[5] Schafer J B, Konstan J A, Riedl J. E-commerce recommendation applications[J]. Data Mining and Knowledge Discovery, 2001, 5(1/2): 115-153.

[6] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state of the art and possible extensions[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[7] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval[M]. New York, USA: ACM Press, 1997.

[8] Goldberg D, Nichols D. Using collaborative filtering to weare an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

[9] 许海玲,吴潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报, 2009, 20(2): 350-362.

Xu Hailing, Wu Xiao, Li Xiaodong, et al. Comparison study of internet recommendation system[J]. Journal of Software, 2009,20(2):350-362.

[10] Zhang Y C, Blattner M, Yu Y K. Heat conduction process on community networks as a recommendation model[J]. Physical Review Letters, 2007,99(15):154301.

[11] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009,71(4):623-630.

[12] Liu J G, Hu Z, Guo Q. Effect of the social influence on topological properties of user-object bipartite networks[J]. The European Physical Journal B, 2013,86(11):1-11.

[13] 高新波, 王笛, 王秀美. 一种潜在信息约束的非负矩阵分解方法[J]. 数据采集与处理,2014,29(1):12-18.
Gao Xinbo, Wang Di, Wang Xiumei. Potential information restrained non-negative matrix factorization[J], Journal of Data Acquisition and Processing, 2014,29(1):12-18.

作者简介:

周海平(1978-),男,教授,研究方向:推荐系统复杂网络,E-mail: hpzhou2885 @163.com。

黄凌英(1979-),女,硕士研究生,研究方向:推荐系统。

刘妮(1983-),女,副教授,研究方向:数据挖掘。

周洪波(1977-),女,讲师,研究方向:推荐系统。