Computer Science and Application 计算机科学与应用, 2017, 7(10), 1036-1044
Published Online October 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.710117

Hans汉斯

Research of Slope One Cooperative Filtering Algorithm Based on Cosine Similarity Weighting

Hua Zhou, Zhi Yuan*

Network Department, South China Institute of Software Engineering, Guangzhou Guangdong
Email: macro_z.h@163.com, *yz@sise.com.cn

Received: Oct. 11th, 2017; accepted: Oct. 23rd, 2017; published: Oct. 30th, 2017

Abstract

In this paper, we propose a collaborative filtering algorithm based on cosine similarity weight (COSLOPE algorithm). The similarity between the users is calculated by the cosine algorithm; the weights are determined according to the similarity degree and the scoring matrix is filled in order to establish the nearest neighbor set with high similarity to the object user. The nearest neighbor set of the nearest neighbor set is to predict the target user's project grade and make recommendations. The algorithm is validated by the MovieLens dataset, and the values of MAE, RMSE and MSE are superior to the traditional Slope One algorithm. COSLOPE algorithm is not only in the effective solution of data sparseness, but also improve the accuracy of the traditional recommendation algorithm and reduce the algorithm response time.

Keywords

Cosine Similarity, Slope One Algorithm, Sparsity Problem, Collaborative Filtering

余弦相似度加权的Slope One协同过滤算法研究

周　化，袁　志*

广州大学华软软件学院网络技术系，广东　广州
Email: macro_z.h@163.com, *yz@sise.com.cn

收稿日期：2017年10月11日；录用日期：2017年10月23日；发布日期：2017年10月30日

*通讯作者。

文章引用: 周化, 袁志. 余弦相似度加权的 Slope One 协同过滤算法研究[J]. 计算机科学与应用, 2017, 7(10): 1036-1044. DOI: 10.12677/csa.2017.710117

摘　要

针对slope one协同过滤算法中存在的数据稀疏性问题展开研究。提出一种基于余弦相似度加权的协同过滤算法(COSLOPE算法)。用加权slope one算法填充稀疏的评分矩阵后利用cosine算法计算用户之间的相似度，得出目标用户的近邻矩阵。通过近邻矩阵中拥有评分记录的用户来预测目标用户的项目评分，并进行推荐。该算法通过MovieLens数据集验证，MAE、RMSE 和MSE的值均优于传统Slope One算法。COSLOPE算法在有效解决数据稀疏性的同时亦提高了传统推荐算法的准确度并降低了算法响应时间。

关键词

余弦相似度，Slope One算法，数据稀疏性，协同过滤

1. 引言

互联网已成为人们获取信息资源的一条重要渠道，但网络数据纷繁复杂，简化对网络的探索过程，提高网络信息的检索效率已逐渐成为诸多学者的研究热点[1]。协同过滤算法主要分为以下两种：基于项目的协同过滤推荐算法、基于用户的协同过滤推荐算法[2] [3]。在[4]中，研究者通过计算借阅用户的皮尔森相关系数，确定用户的最近邻聚类，从而形成用户与图书之间的推荐，该类推荐算法对缺乏用户评论数据的项目有效。赵文涛等从用户单属性相似性入手，计算权重并代入用户多属性维度中，用以增加评分矩阵的稠密性。在一定程度上解决了用户评分矩阵稀疏性的问题[5]。孔维良等人针对用户冷启动的问题，提出用加权贝叶斯算法对新用户进行扩展，再计算后验概率避免了推荐过程中先验概率计算可能带来的数据敏感性问题[6]。协同过滤推荐算法的核心是相似度计算，在实际应用中，由于用户和项目数量十分庞大，而用户往往仅对一小部分项目进行评分，这就导致用户项目评分矩阵的稀疏，使相似度计算的难度和不准确性大大增加[7]。

针对上述问题，为了改善数据稀疏性影响，提高相似度计算的准确度，将加权 slope one 算法引入到传统的协同过滤算法中：首先，用加权 slope one 算法对评分矩阵进行填充，接着利用 cosine 余弦算法计算用户之间的相似度；然后，选取与目标用户相似度较高的前 i 个用户作为其近邻用户集合，利用近邻用户已有的评分记录来预测目标用户对未评分项目的评分；最后，根据预测评分结果中预测评分较高的项目对目标用户进行个性化推荐。经过实验验证，基于余弦相似度的加权 slope one 算法对推荐结果的精度有一定的提高。

2. 协同过滤推荐算法

相比较其他推荐算法，协调过滤算法具有以下两点优势[5]：第一，对推荐目标无特殊要求，无论是复杂亦或抽象的目标都能够进行推荐；第二，只需要用户历史评分数据，而不需要用户本身的相关属性信息。

2.1. 基于项目的协同过滤推荐算法(IBCF)

基于项目的协同过滤方法，以计算项目相似度为核心，利用目标项目相似性较高的项目的评分信息来预测用户对目标项的评分结果。IBCF 基于如下假设——用户更倾向于选择自己喜欢的项目或类似的项目。基本思路是首先找到目标项目相似度较高的最近邻居并建立项目近邻矩阵，从而根据当前用户的未评分项目(目标推荐对象)的近邻矩阵来预测当前用户对目标推荐对象的评分；然后给当前用户推荐预测评分中排名靠前的若干个目标对象。由于项目相似度的计算可以以离线方式进行，因此基于项目的协同过滤方法就可以提前计算好所有项目之间的相似度，从而节省了推荐时间，以更高的效率得到推荐结果。

2.2. 基于用户的协同过滤推荐算法(UBCF)

UBCF 是当今最为流行的个性化推荐技术之一，该算法的基本思想是通过寻找相似度较高的用户，并将其感兴趣的项目推荐给目标用户[6]。整个推荐过程可具体划分为以下三点：首先，利用已有的用户对项目的评分记录来计算用户之间的相似度；然后，对得到的相似度结果进行降序排序，将相似度较高的用户作为目标用户的最近邻居，生成目标用户近邻矩阵，利用目标用户近邻矩阵已有的评分记录来预测目标用户对未评分项目的评分；最后，选取预测评分中排名靠前的若干个项目作为推荐结果反馈给用户。

相较于其他算法，UBCF 具有以下一些优点：1) 有推荐信息的能力，由于用户无法提前得到推荐信息的内容，因此可以挖掘用户的潜在需求(即隐性存在但自己尚未发现的兴趣偏好)并对其进行相应推荐；2) 能够排除机器难以识别的分析信息，如图片、音乐等；3) 借鉴他人的经历和实验，提高了内容分析的精度和完整性，并且能够过滤一些繁杂的、难以描述的概念(如信息质量、审美观念)。

3. Coslope 算法理论

3.1. Slope one 算法

Slope one 算法是一种基于线性回归模型假设的协同过滤算法，相较于其他复杂的协同过滤算法，在具有同等推荐精度的条件下，其花费更少，更加易于实现，且其预测和项目推荐速度较快，它的简洁高效使得采用 Slope one 算法的推荐系统更加易于维护[7]。当有新增的用户或新增项目的评分时，该算法能够动态更新模型，用户有较少的评分也可以进行推荐，因为可以通过其他用户来计算。

Slope one 算法采用 $f(x)=k-x$ 进行预测，其中 x 代表项目的已知评分，参数 k 是用户对两项目的平均评分偏差，$f(x)$是用户预测的评分。假设推荐系统中有 m 个用户和 n 个项目，分别建立两个集合 $U=\{u_1,u_2,u_3,\cdots,u_m\}$ 和 $I=\{i_1,i_2,i_3,\cdots,i_n\}$，$U$ 代表用户的集合，I 代表项目的集合。推荐算法的核心是用户项目评分矩阵，矩阵的行向量表示每个用户的评分，矩阵的列向量表示每个项目的得分。为了使计算更加明确，采用 $r_{i,j}\ (1\le i\le m,1\le a\le n)$ 表示用户 i 对项目 a 的评分。a，b，c……代表项目，S_{ab} 为项目 a，b 都评过分的用户集合，r 代表项目评分值，dev_{ab} 为项目 a，b 平均评分偏差，$Count(S_{ab})$ 为 S_{ab} 集合中有过评分记录的用户个数。Slope one 算法步骤为：

首先计算项目 i_a 与其他项目 i_b 之间的平均评分偏差 dev_{ab}，

$$dev_{ab}=\sum_{U_a\in S_{ab}}\frac{r_{i.b}-r_{i,a}}{count(S_{ab})} \tag{1}$$

然后预测当前活跃用户 u 对目标项目 a 的可能评分 $Prediction_{u,a}$，

$$Prediction_{u,a}=\frac{\sum_{b\in R_a}(r_{u,b}-dev_{ab})}{Count(R_a)} \tag{2}$$

3.2. Coslope 算法的加权建模

数据稀疏性问题是当前推荐系统所面临的主要问题之一。Slope one 算法的实现简单高效，而且精确度也有不错表现。例如：如表 1 有 User 1、User 2 和 User 3 三个用户对 Item 1 进行了评分，同时 User 1 和 User 2 对 Item 2 进行了评分，由此可以计算出 User 3 对 Item 3 的评分。

代入公式(1)、公式(2)可得，User 3 对 Item 2 的评分为 $3-\left[(3-2)+(5-1)\right]/2=0.5$。

然而 slope one 算法在计算不同项目之间的评分差异时，未考虑参与评分的用户数量不一致。例如：如果现有 50 个用户对 Item 1 和 Item 2 都有评分记录，有 500 个用户对 Item 2 和 Item 3 进行了评分，显然这两个评分记录差的权重是不相等的。此外，预测评分的过程中会用到与目标项目完全不同的对象，而产生较大误差。在数据极度稀疏的情况下，数据的缺失会导致推荐精度急剧下降。

因此为解决以上问题，本文在 Slope one 算法的基础上提出了一种基于余弦加权的 COSLOPE 算法。COSLOPE 算法，是针对数据稀疏性的 Slope one 算法优化。其中，改进后的数学模型如下所示：

$$Prediction_{u,a}^{w}=\frac{\sum_{b\in R_a}\left(dev_{ab}+r_{u,b}\right)S_{ab}}{\sum_{b\in R_a}S_{ab}} \tag{3}$$

在协同过滤推荐算法中，相似度计算的准确性是至关重要的。根据众多文献中的相似度实验结果表明，在数据稀疏性的情况下，大多数相似度计算只考虑了两个用户共同评分的项目，但是当两个用户他们共同评分的项目数量较少时，这会降低相似度的准确性导致推荐质量下降。除此之外，在计算相似度时可能会出现 sim 值为正、负、0 以及无法计算的情况。其中，导致相似度无法计算的情况主要有两种：一种是两个用户之间没有共同评分项目；另一种是在相似度算法的计算公式中，存在分母为零的情况。在周张兰的个性化推荐研究论文中提到，相比其他相似度算法，余弦相似度在数据稀疏的情况下的相似度较高，同时，针对第二种情况，余弦相似度可以有效地计算出相似度的值[8]。对此，本文采用基于余弦相似度的加权 slope one 算法来构建推荐模型，以此提高算法的推荐精度。

在余弦相似度计算中，用户对项目的评分为 n 维项目空间上的向量，在用户项目评分矩阵中，用户对项目已有的评分为实际的评分值，用户未进行评分的项目用 0 表示，通过向量间的余弦夹角来度量用户间的相似性。在基于用户的协同过滤推荐(UBCF)中，相似度的计算是指用户与用户之间的相似度，即用户项目评分矩阵中行向量之间(用户之间)的相似度。假设 j、k 表示用户，h，i 表示项目，其他符号说明如下：

R_{jh}：用户 j 对项目 h 的评分。

I_j：用户 j 评分过的项目集。

I_{jk}：用户 j、k 一起评分过的项目集 $I_{jk}=I_j\cap I_k$.

I_h：所有对项目 h 评分的用户集合。

Table 1. User rating table
表 1. 用户评分表

User / User Rating	Item 1	Item 2
User 1	3	2
User 2	5	1
User 3	3	?

I_{hi}：项目 h、i 一起评分的用户集合 $I_{hi}=I_h\cap I_i$.

$\overline{R}_j$：用户 j 的平均评分。

$\overline{R}_h$：项目 h 的平均评分。

基于用户的协同过滤推荐中用户 j 和用户 k 余弦相似性计算公式如下：

$$sim(j,k)=\frac{\sum_{h\in I_{jk}}\left(R_{j,i}\times R_{k,i}\right)}{\sqrt{\sum_{h\in I_j}\left(R_{j,h}\right)^2}\sqrt{\sum_{h\in I_k}\left(R_{k,h}\right)^2}} \tag{4}$$

其中分子为两个用户评分向量的内积；分母为两个向量模的乘积，夹角越小，表示相似度越高。

4. 实验结果与分析

4.1. 算法思路

由图 1 可知，COSLOPE 算法的核心是对已建的项目评分矩阵内的空值进行加权填充，再采用余弦算法计算出评分用户的近邻矩阵，最后通过个性化推荐形成推荐结果。

实验数据集及检验标准

电影评分数据集 Movielens 是本次实验所采用的数据集，该数据包含了 943 个用户对 1682 部电影的 100,000 条评分，其中每位用户至少对 20 部电影拥有评分记录。评分值范围是 1 到 5 分，代表用户对电影的评价越好分值越高。该数据稀疏度为 6.30%。为了对比 COSLOPE 算法，实验时从该数据集中随机抽取 70%的数据作为训练集，其余 30%作为测试集。评价推荐算法效用的指标有很多，本文选择算法的精确度作为主要评价指标。其中精确度的度量方法有平均绝对误差 MAE、均方根误差 RMSE 和平均平方误差 MSE 是衡量精确度的最典型指标，MAE、RMSE 和 MSE 的值越小，代表推荐精度越高[9]。

1) 平均绝对误差(MAE)：是对实际和预测的差异的绝对值取平均得到的。它是最基础，也是应用最为广泛的评价标准之一。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|p_{ua}-r_{ua}\right| \tag{5}$$

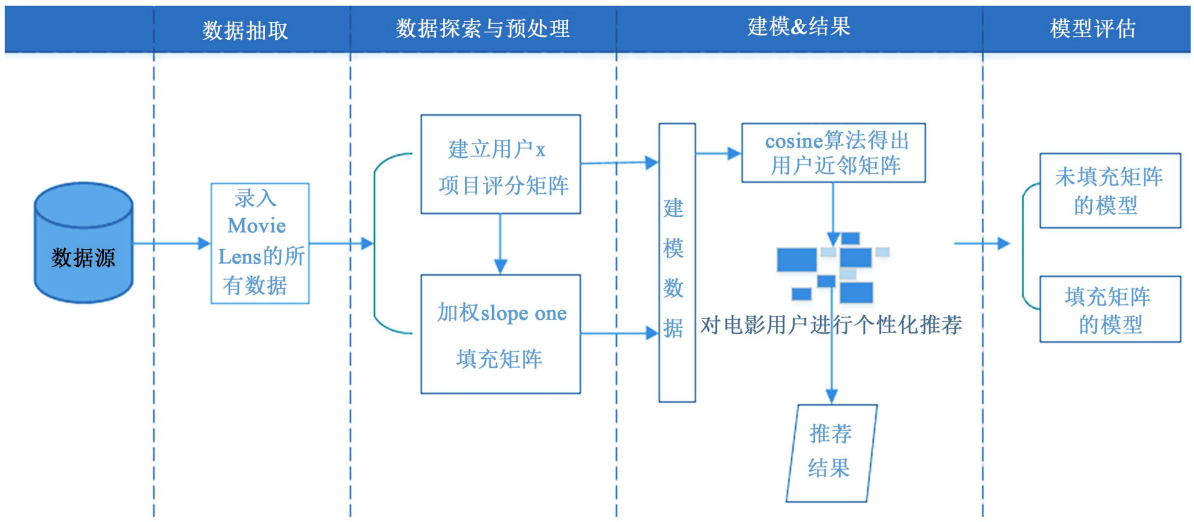

Figure 1. Algorithm flowchart of COSLOPE
图 1. COSLOPE 算法流程

2) 均方根误差(RMSE)：是指预测值与实际值偏差的平方与观测次数 n 的比值的平方根，在实际预测中，观测次数 n 总是有限的，实际值只能用最佳值来代替。在一组测量中，均方根误差对特大或特小的误差反映特别敏感，因此，RMSE 能够很好地反映出预测的精确度。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{a=1}^{n}\left|p_{ia}-r_{ia}\right|^{2}} \tag{6}$$

3) 平均平方误差(MSE)：先对系统预测评分与用户实际评分偏差进行平方，然后求和，因此评分误差越大对平均平方误差的影响比对平均值对误差的影响越大。

$$\mathrm{MSE}=\frac{1}{n}\sum_{a=1}^{n}\left|p_{ua}-r_{ua}\right|^{2} \tag{7}$$

假设 p 和 r 分别代表预测评分集合和实际评分集合，其中，用户 u 已评分项目的个数用 n 来表示，评分的预测值和真实值用 p_{ua} 和 r_{ua} 表示。

4.2. 加权填充代码

```
for (tar_ID in 1:nrow(mydata))
 {
 tar_0_col=which(mydata[tar_ID,]==0)
 tar_n0_col=which(mydata[tar_ID,]!=0)
for (tar_0 in tar_0_col)
 {
for (tar_n0 in tar_n0_col)
 {
ints_rating_row=intersect(which(mydata[,tar_0]!=0),which(mydata[,tar_n0]!=0))
for(m in ints_rating_row)
 {
sum_pair_diff=sum_pair_diff+(mydata[m,tar_n0]-mydata[m,tar_0])
 } sum=(mydata[tar_ID,tar_n0]-sum_pair_diff/length(ints_rating_row))*length(ints_rating_row)
sums=sums+sum
len=len+length(ints_rating_row)
sum_pair_diff=0
 }
```

实验结果的分析

对两种算法进行对比，得到 MAE、RMSE、MSE 值如下(图 2~4 和表 2~4)：

由以上分析结果可知，通过矩阵填充进行改进的基于用户的协同过滤的推荐结果具有较小的平均误差，改进后的基于用户的协同过滤算法总体推荐性能更好。使用加权 slope one 算法对矩阵进行填充，能够显著提高系统预测精度。

从 Movielens 实验数据集随机选取 3 个用户进行推荐，选取用户 ID 为 801、802、803。图 5 所示为用未填充的矩阵进行协调过滤推荐的结果，由于矩阵过于稀疏，3 个用户都没有得到推荐结果。图 6 为加权 slope one 算法填充矩阵后的推荐结果，801 号用户推荐电影 ID 为 12、271、427、223、316；802 号用户推荐电影 ID 为 474、408、272、316、641；803 号用户推荐电影 ID 为 427、484、408、170、513。由此可见填充后的推荐性能更高。

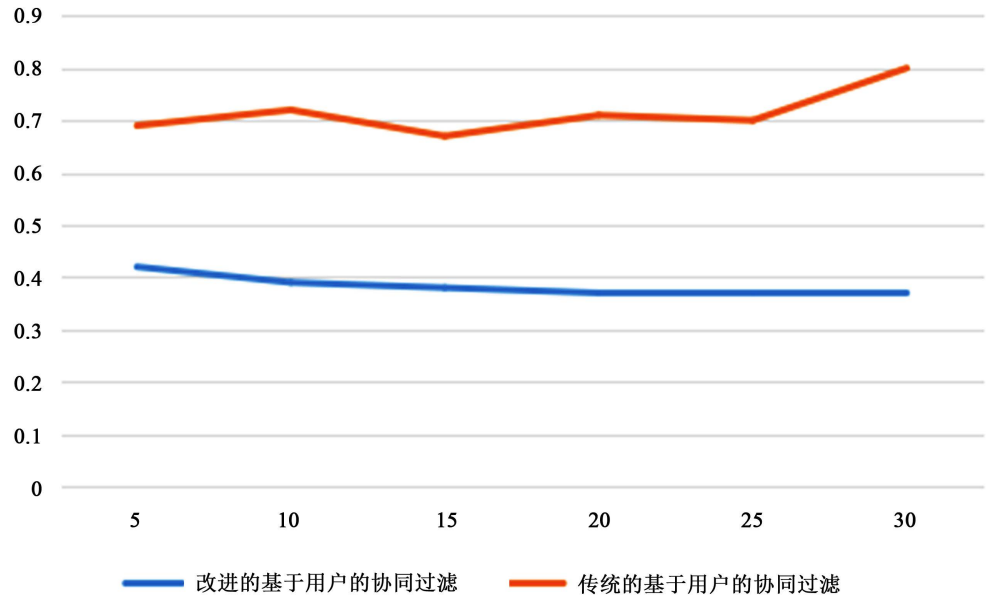

Figure 2. Recommended algorithm MAE comparison
图 2. 推荐算法 MAE 比较

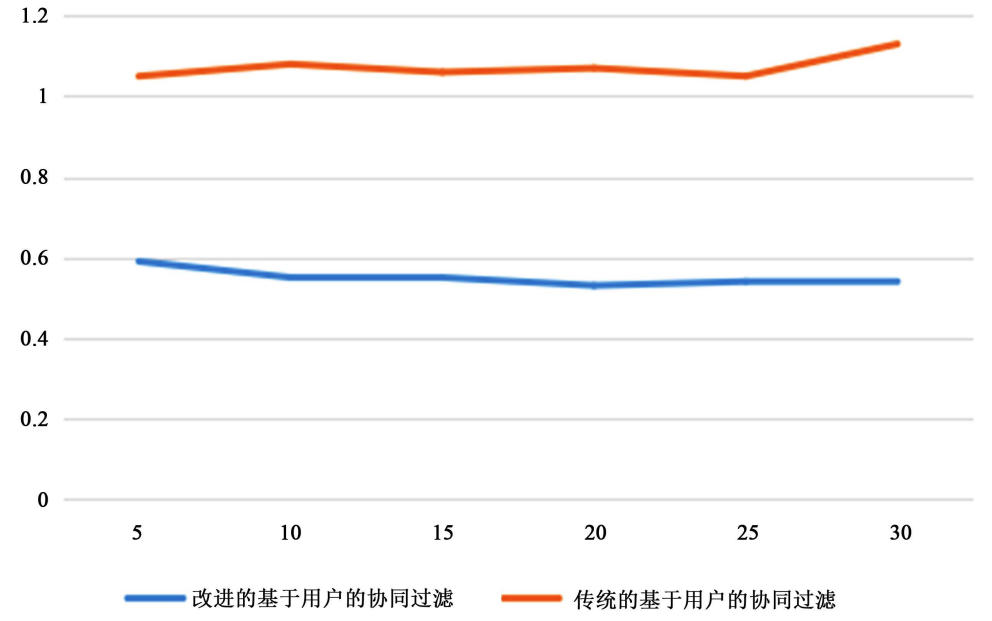

Figure 3. Recommended algorithm RMSE comparison
图 3. 推荐算法 RMSE 比较

Table 2. MAE comparison results
表 2. MAE 对比结果

推荐算法	最近邻居数目					
	5	10	15	20	25	30
COSLOPE 算法	0.35	0.31	0.30	0.28	0.3	0.29
SLOPE ONE 算法	1.11	1.16	1.12	1.15	1.1	1.27

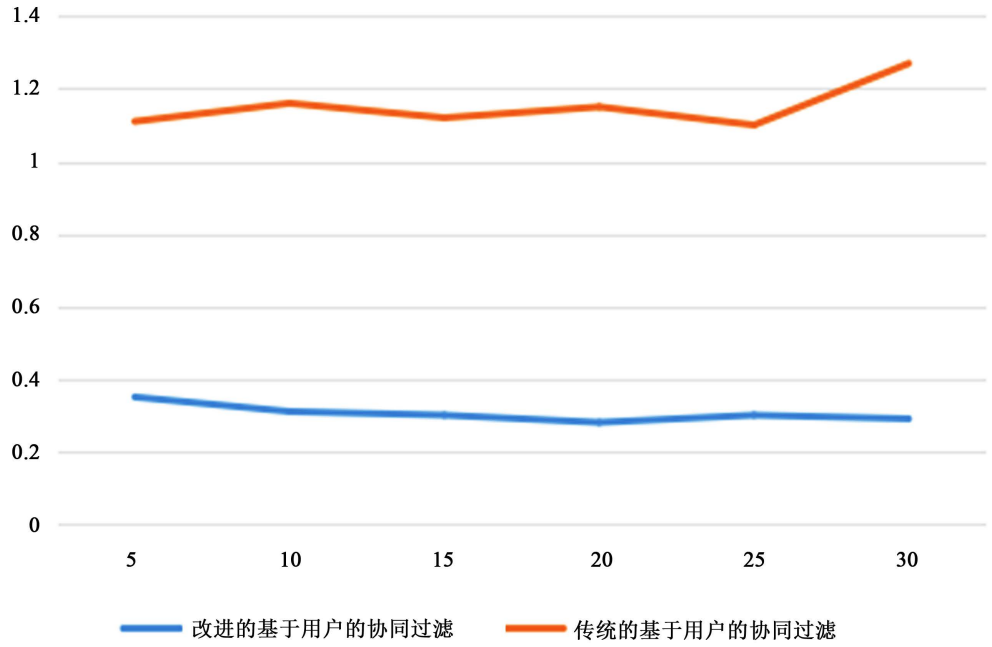

Figure 4. Recommended algorithm MSE comparison
图 4. 推荐算法 MSE 比较

```
>ml.predict3

[[1]]
character (θ)

[[2]]
character (θ)

[[3]]
character (θ)
```

Figure 5. The recommended result is not filled
图 5. 未进行填充的推荐结果

```
>ml.predict1

[[1]]
[1]  “12”  “272”  “427”  “223”  “316”

[[2]]
[1]  “474”  “408”  “272”  “316”  “641”

[[3]]
[1]  “427”  “484”  “408”  “170”  “513”
```

Figure 6. After filling the recommended results
图 6. 填充后的推荐结果

Table 3. RMSE comparison results
表 3. RMSE 对比结果

推荐算法	最近邻居数目					
	5	10	15	20	25	30
COSLOPE 算法	0.42	0.39	0.38	0.37	0.37	0.37
SLOPE ONE 算法	0.69	0.72	.067	0.71	0.7	0.8

Table 4. MSE comparison results
表 4. MSE 对比结果

推荐算法	最近邻居数目					
	5	10	15	20	25	30
COSLOPE 算法	0.59	0.55	0.55	0.53	0.54	0.54
SLOPE ONE 算法	1.05	1.08	1.06	1.07	1.05	1.13

5. 结论

互联网信息量不断膨胀，网络经济发展迅速，社会信息超载成为越来越严重的问题。而推荐系统是一种利用信息过滤来帮助人们在大量数据中找到对自己有用的信息的方法，是解决当前信息超载问题的十分有效地手段。虽然社会上存在各式各样大量的数据，但能用来有效的进行分析的数据并不多，从而导致推荐系统精度面临挑战。

本文从传统推荐算法对数据矩阵数据稀疏性的应对不足着手，提出基于余弦相似度计算和加权 slope one 填充矩阵的协同过滤算法，并通过反复多次的实验分析比较算法的精度。实验结果表明基于评分相似性和加权 slope one 填充矩阵的协同过滤算法明显地降低了数据的稀疏性、提高了信息推荐精度。

基金项目

湖南省自然科学基金(2016JJ4090)广东省攀登计划基金(PDJH2016a0991)。

参考文献 (References)

[1] 刘蓓琳. 电子商务个性化推荐研究[M]. 北京: 中国经济出版社, 2015.

[2] 王毅, 楼恒越. 一种改进的 Slope One 协同过滤算法[J]. 计算机科学, 2011, 38(10A): 192-194.

[3] 赵亮, 胡乃静, 张守志. 个性化推荐算法设计[J]. 计算机研究与发展, 2002, 39(8): 986-991.

[4] 刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展, 2009, 19(1): 1-15.

[5] 李伟霖, 王成良, 文俊浩. 基于评论与评分的协同过滤算法[J]. 计算机应用研究, 2016, 5(34): 37-40.

[6] 赵文涛, 王春春. 基于用户多属性与兴趣的协同过滤算法[J]. 计算机应用研究, 2016, 4(33): 29-32.

[7] 田松瑞. 基于用户相似度加权的 Slope One 算法[J]. 软件, 2016, 37(4): 56-59.

[8] 周张兰. 基于协同过滤的个性化推荐算法研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2009.

[9] Lv, L., Medo, M., Yeung, H.C., *et al*. (2012) Recommender Systems. *Physics Reports*, **519**, 1-49. https://doi.org/10.1016/j.physrep.2012.02.006

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org