

基于隐式评分和相似度传递的学习资源推荐*

付 芬¹, 豆育升¹, 韩 鹏^{1,2}, 李耀辉¹

(1. 重庆邮电大学 计算机科学与技术学院, 重庆 400065; 2. 重庆市科学技术研究院 信息与自动化技术研究中心, 重庆 401123)

摘 要: 协同过滤推荐算法存在数据稀疏的问题,这使得学习平台中,由于用户学习行为记录的稀疏而无法满足用户的学习需求。为此,提出了一种基于隐式评分和相似度传递的学习资源推荐算法。首先,收集用户的学习行为;其次,改进传统的相似度计算方法,并在此基础上引入相似度传递策略;最终应用并实现 E-learning 平台中学习资源的推荐。实验表明,该算法能够在一定程度上解决相似度计算不准确以及数据稀疏问题,从而提高学习资源的推荐质量。

关键词: 协同过滤; 学习行为; 数据稀疏; 隐式评分; 相似度传递

中图分类号: TP181

文献标志码: A

文章编号: 1001-3695(2017)12-3725-05

doi:10.3969/j.issn.1001-3695.2017.12.047

Learning resource recommendation based on implicit scoring and similarity propagation

Fu Fen¹, Dou Yusheng¹, Han Peng^{1,2}, Li Yaohui¹

(1. College of Computer Science & Technology, Chongqing University of Posts & Telecommunications, Chongqing 400065, China; 2. Research Center for Information & Automation Technology, Chongqing Academy of Science & Technology, Chongqing 401123, China)

Abstract: Traditional collaborative filtering recommendation algorithm has the problem of sparse data, which makes the learning needs of users cannot be satisfied because of the sparsity of user learning behavior records. To address this issue, this paper proposed a learning resource recommendation algorithm based on implicit rating and similarity propagation. Firstly, it collected the user's learning behavior. Secondly, it improved the calculation method of similarity. On the basis of this, it introduced the similarity propagation strategy. Finally, it applied and implemented the collaborative filtering algorithm based on personalized learning resources in E-learning. Experiments show that the proposed algorithm can solve the problem of inaccurate and sparse data, and improves the quality of learning resources.

Key words: collaborative filtering; learning behavior; data sparsity; implicit rating; similarity propagation

0 引言

个性化推荐技术在电子商务、电影、音乐等领域得到了广泛应用并取得了不错的成绩,如今随着网络在线教育的发展,个性化推荐技术也逐渐引起了教育领域的重视及需求。面对海量的学习资源,学习者往往会陷入学习迷茫之中,面对复杂多样的网络学习资源,如何根据用户的历史行为轨迹,分析用户兴趣,推荐符合用户学习需求的学习资源变得越来越重要^[1,2]。协同过滤是目前应用最广泛且最成功的推荐算法^[3]。在 E-learning 平台中,大部分学习者都不愿意提供显式的评分,并且随着用户和学习资源数量的日益增长,用户对学习资源的评分密度日趋稀疏化和分离化,当评分密度较大时,使用传统的协同过滤推荐算法可以得到较可靠的结果,当评分密度过小时,就会使得传统的计算方法缺乏计算依据,使得计算结果不可靠,从而对推荐结果的准确性产生很大的影响^[4]。

针对协同过滤算法面临的数据稀疏问题,研究者大多从以

下方法来提高推荐系统的效果:a)通过对原始的数据评分矩阵进行预测填充,以此来增加用户—项目评分矩阵稠密度;b)在没有对评分矩阵进行预测填充情况下,采用经过改进的计算相似度的算法来提高其计算精确性,另外是结合其他的一些推荐算法^[5]。目前常见的数据稀疏问题解决办法有固定值法、平均值推荐法。固定值法的优点是简单,能在一定程度上减缓数据稀疏性带来的准确性不高问题,缺点是该方法采用的是固定值填充评分矩阵,没有考虑各用户之间或者项目之间的特征属性的差别^[6]。平均值法是获取现有用户的评分平均值作为未评分的资源的预测分值,但用户的偏好与平均值恰好一致的可能性太小,并未考虑用户的个性化差异。

针对上述存在的问题,本文提出了一种基于隐式评分和相似度传递的学习资源推荐算法。首先,通过收集用户的学习行为(评论、点赞、收藏、下载)记录,将行为(隐式评分)加权,转换为显式评分,构成用户—学习资源评分矩阵;然后通过本文提出的一种新的相似性度量方法计算用户之间的相似度,引入相似度传递策略,将用户对学习资源的相似度传播到全局,为

收稿日期: 2016-08-10; 修回日期: 2016-10-08 基金项目: 重庆市科技研发基地能力提升项目(cstc2014pt-gc40004)

作者简介: 付芬(1991-),女,硕士研究生,主要研究方向为数据挖掘(692094632@qq.com);豆育升(1953-),男,博导,主要研究方向为计算机模拟、分子动力学;韩鹏(1977-),男,研究员,主要研究方向为计算机应用项目开发、互联网+;李耀辉(1989-),男,硕士研究生,主要研究方向为移动互联网数据应用。

目标用户寻找潜在的推荐者,扩大目标学习者的推荐范围。该算法不仅能够有效地缓解相似度计算不准确和数据稀疏性问题,而且在一定程度上提高了协同过滤推荐算法的准确性。

1 基于协同过滤推荐算法的学习资源推荐

1.1 协同过滤推荐算法主要思想

协同过滤推荐算法应用于个性化学习资源推荐的基本原理是:根据学习者的历史学习记录形成学习者—学习资源评分矩阵,利用相似性度量方法(如余弦相似度、皮尔森相关系数等),通过计算学习者或学习资源之间的相似度,挖掘出与目标学习者相似的学习者集合或与学习者历史偏好相似的学习资源集合,形成邻居,然后基于邻居学习者对学习资源的评分信息,预测目标学习者对学习资源的预测评分值,并根据预测评分值得大小为目标学习者进行个性化推荐^[7-9]。在协同过滤中,项目可以看做一个黑盒子,本文无法看到里面的内容,那么与该项目相关的一些用户交互行为(如评分、收藏等)就被应用于推荐过程中^[10]。图1为协同过滤推荐算法原理图,与用户 U 具有相似评分的用户,对包含于资源集中的资源 R 进行过评分,因此将资源 R 推荐给用户 U 。

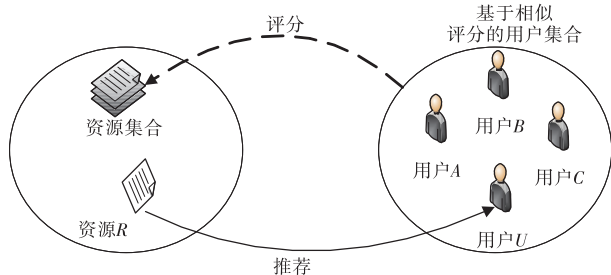


图1 协同过滤推荐算法原理图

依据上述原理,目前将协同过滤推荐技术应用于学习资源推荐主要有两种方法:

a) 基于学习者的协同过滤推荐算法。基于学习者的协同过滤推荐算法是基于学习者对共同评分资源的评分计算学习者之间的相似度,然后根据相似度大小为目标学习者找到相似邻居,将邻居集合中目标学习者感兴趣却未学习过的学习资源推荐给目标学习者。

b) 基于学习资源的协同过滤推荐算法。不同于基于学习者的协同过滤推荐,基于学习资源的协同过滤推荐算法是在计算学习资源的相似度的基础上,向目标用户推荐与其历史偏好相似的学习资源。

两种方法虽已应用于不同场合并取得了一定效果,但是仍然存在数据稀疏的问题。由于学习者评分数据的稀疏性,学习者之间的共同评分资源数很少,使得用户相似度的计算不够准确,从而导致推荐的准确度不高。

1.2 协同过滤推荐流程

传统的协同过滤推荐算法是基于系统中其他用户对某项目资源的评分向目标用户进行推荐,其推荐的流程主要包括构建用户—项目评分矩阵、寻找与目标用户相似的最近邻居集、根据最近邻居集对目标用户产生推荐^[11,12]。

1.2.1 构建用户—项目评分矩阵

在推荐系统中,通常需要收集系统中用户对所有项目的评分,将这些数据表示为一个 $M \times N$ 的用户—项目评分矩阵 R ,

如表1所示。其中, M 表示推荐系统中的用户数, N 表示推荐系统中的项目数, R_{ij} 表示第 i 个用户对第 j 个项目的评分, $U = \{user_1, user_2, \dots, user_m\}$ 表示用户集, $I = \{item_1, item_2, \dots, item_n\}$ 表示项目集, $R(i, j) \{i, j | i \in (1, m), j \in (1, n)\}$ 为评分矩阵 R 中的评分元素之一。本文中评分分值 $R(i, j)$ 为数值类型,设置为1~5。假如某用户对其中一项资源的评分为5,则表示用户对该资源非常感兴趣;假如评分为1,则表示用户对该资源不感兴趣。

表1 用户—项目评分矩阵

item ₁	item ₂	...	item _{n-1}	item _n
user ₁	4	3	$R(1, n-1)$	$R(1, n)$
user ₂	2	5	$R(2, n-1)$	$R(2, n)$
...	...	...	...	...
user _{m-1}	$R(m-1, 1)$	$R(m-1, 2)$	$R(m-1, n-1)$	$R(m-1, n)$
user _m	$R(m, 1)$	$R(m, 2)$	$R(m, n-1)$	$R(m, n)$

1.2.2 寻找与目标用户相似的最近邻居集

在传统的协同过滤推荐算法中,为需要推荐服务的某用户,寻找最相似的最近邻居,然后根据邻居用户的兴趣偏好产生对目标用户的推荐。例如,用户 U 需要产生根据相似度 $\text{sim}(U, X_i)$ 大小排序的邻居集合 $X = \{X_1, X_2, \dots, X_n\}$, $U \notin X$ 。 $X_1 \sim X_n$ 的相似度按照从小到大顺序排列。用户相似度计算是协同过滤推荐系统中决定推荐结果好坏的关键步骤^[13]。

首先对本部分使用的变量进行定义,定义用户 U 和用户 V 在 n 维项目空间对项目 i 的评分分别为 $r_{U,i}$ 、 $r_{V,i}$;用户 U 和用户 V 的共同评分项目集合为 $R_{U,V}$; RU 和 RV 分别表示用户 U 和用户 V 各自评分项目的集合; $\bar{r}_U$ 和 $\bar{r}_V$ 则分别代表了用户 U 和用户 V 对项目的平均评分。下面介绍常用的相似性度量方法。

a) 余弦相似性。用户间的相似度通过向量间的余弦夹角度量,将用户对项目的评分看做 n 维项目空间的向量。则用户 U 和用户 V 之间的相似性计算公式如下:

$$\text{sim}(U, V) = \cos(U, V) = \frac{\sum_{i \in R_{U,V}} r_{U,i} \times r_{V,i}}{\sqrt{\sum_{i \in R_{U,V}} r_{U,i}^2} \sqrt{\sum_{i \in R_{U,V}} r_{V,i}^2}} \quad (1)$$

b) 修正的余弦相似性。在余弦相似性度量方法中忽略了不同用户的评价尺度问题。为了改善余弦相似性度量方法出现的问题,在修正的余弦相似性度量方法中通过减去用户对项目的平均评分来改善上述缺陷^[14],则用户 U 和 V 之间的相似度计算方式如下:

$$\text{sim}(U, V) = \frac{\sum_{i \in R_{U,V}} (r_{U,i} - \bar{r}_U)(r_{V,i} - \bar{r}_V)}{\sqrt{\sum_{i \in R_{U,V}} (r_{U,i} - \bar{r}_U)^2} \sqrt{\sum_{i \in R_{U,V}} (r_{V,i} - \bar{r}_V)^2}} \quad (2)$$

c) 相关相似性。相关相似性也称皮尔森(Pearson)相关系数,通过皮尔森相关系数来度量用户间的相似性。则用户 U 和 V 之间的相关相似性计算方式如下:

$$\text{Sim}(U, V) = \frac{\sum_{i \in R_{U,V}} (r_{U,i} - \bar{r}_U)(r_{V,i} - \bar{r}_V)}{\sqrt{\sum_{i \in R_{U,V}} (r_{U,i} - \bar{r}_U)^2} \sqrt{\sum_{i \in R_{U,V}} (r_{V,i} - \bar{r}_V)^2}} \quad (3)$$

1.2.3 根据最近邻居集对目标用户产生推荐

根据相似度计算方法得到最近邻居集合后,可以预测目标用户对任意项的兴趣度的评分值。假设, $P_{U,i}$ 为用户 U 对项目 N 的预测评分; $\bar{R}_U$ 为用户 U 所有已经评分分值的平均值; $\text{sim}(U, X)$ 为用户 U 与最近邻居集中某用户的相似度; N 为最近邻居集中某用户对项目 N 的评分; N 为最近邻居集中某用户对项目的平均评分; N 为用户的最近邻居集合; $item_u$ 为最近邻居

集 $item_u$ 中的用户之一。则目标用户 U 对项目 N 的预测评分为

$$P_{U,i} = \bar{R}_U + \frac{\sum_{X \in N} \text{Sim}(U, X) (R_{X,i} - \bar{R}_X)}{\sum_{X \in N} |\text{Sim}(U, X)|}$$



根据上述方法计算出目标用户对未评分项目的兴趣度预测值,取其中 k 个排在最前的项作为 top- N 推荐集结果。

2 基于隐式评分和相似度传递的协同过滤推荐算法 (CF-IRST)

部分研究者提出用户之间的兴趣可以通过用户间的信任关系来衡量,并提出通过信任关系的传递机制可以在一定程度上提高推荐质量^[15,16]。但本文考虑到即使存在信任关系也会产生不准确性,例如 A 信任 B, A 喜欢悬疑片, B 热衷于文艺片,此时将文艺片推荐给 A 显然是不准确的推荐。此外,大多数推荐系统并没有设置信任机制,同样会带来一定误差。由此想到将信任关系之间的传递机制移植到相似度上,使得相似度也能传递,从而缓解数据稀疏问题。

本章将详细描述本文提出的融合隐式评分和相似度传递的协同过滤推荐算法,算法流程如下:

- 收集用户学习行为,将收集到的用户学习行为转换为显式评分,构建评分模型,建立如表 1 所示的用户—学习资源评分矩阵。
- 寻找目标用户的最近邻居。加入相似因子以提高相似度计算的置信度,代替传统协同过滤算法中的相似度计算,并形成近邻。
- 预测目标用户的评分,产生推荐集合。根据本文提出的相似度传递策略,按照计算公式计算出新的相似因子,将传统协同过滤推荐算法中的相似度替代为上部分计算得出的相似因子。取前 k 个作为推荐集结果。算法流程图如图 2 所示。

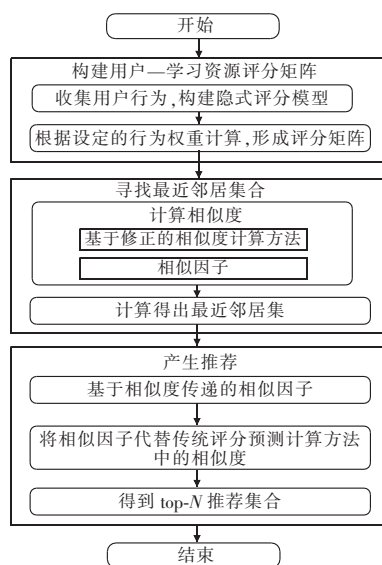


图 2 算法流程图

2.1 隐式评分模型

无论在电子商务推荐平台或者学习平台中,由于用户并不愿意花费时间和精力去主动提供信息,作出显式反馈,往往无法收集到足够的数据,因此用户对学习资源的显式评分通常比较稀疏甚至没有^[17,18]。本文从收集用户对学习资源的学习行

为出发,通过用户对学习资源的点赞、评论、收藏、下载四种行为进行研究,将用户的行为量化为用户的兴趣,实现隐式评分到显式评分的转换,最终构建用户—学习资源评分模型。

具体步骤如下:

a) 定义系统的评分为 $R_{\text{range}} \in [R_{\min}, R_{\max}] = [1, 5]$, $R_{\min}$ 和 $R_{\max}$ 分别为系统评分范围的最小值和最大值,评分范围依据系统情况而设定。

b) 定义 R_{down} 为用户下载行为的评分, Con_{down} 为下载行为的奖励因子。

$$R_{\text{down}} = \lambda \times R_{\max} \times Con_{\text{down}} \quad (5)$$

其中, λ 为 1 时表示用户触发该行为, λ 为 0 则表示未触发该行为。

c) 定义 R_{comment} 为用户评论行为的评分, Con_{comment} 为评论行为的奖励因子。

$$R_{\text{comment}} = \lambda \times R_{\max} \times Con_{\text{comment}} \quad (6)$$

d) 定义 R_{favorite} 为用户收藏行为的评分, Con_{favorite} 为收藏行为的奖励因子。

$$R_{\text{favorite}} = \lambda \times R_{\max} \times Con_{\text{favorite}} \quad (7)$$

e) 定义 R_{like} 为用户点赞行为的评分, Con_{like} 为点赞行为的奖励因子。

$$R_{\text{like}} = \lambda \times R_{\max} \times Con_{\text{like}} \quad (8)$$

f) 根据本系统情况,用户的四种行为都反映用户对学习资源的兴趣,但兴趣程度不同,本文假设用户的下载和评论行为反映出的对学习资源的兴趣比收藏和点赞行为要高。且本文所赋予的下载和评论行为的奖励因子相等,设为 0.3,收藏和点赞行为的奖励因子相等,设为 0.2。

$$Con_{\text{down}} + Con_{\text{comment}} + Con_{\text{favorite}} + Con_{\text{like}} = 1 \quad (9)$$

根据上述步骤,则可以通过量化用户的行为从而计算出用户 u 对资源 i 的评分 $R_{u,i}$ 。

$$R_{u,i} = R_{\text{down}} + R_{\text{comment}} + R_{\text{favorite}} + R_{\text{like}} \quad (10)$$

2.2 基于修正的相似性计算 (RS)

无论是电子商务平台或是学习平台中,随着系统资源的不断增加,用户提供的主动评分会越来越稀疏,本文将用户的学习行为量化为显式的评分,在建立用户评分矩阵的基础上,传统的协同过滤通常选用皮尔森相关系数来计算用户相似度。如表 2 所示,当计算用户间相似性时,若出现下表中的评分现象,当使用第 2 章所述的相似度计算方法计算,则用户 1 与用户 2 之间的相似性为 0 或者出现较大误差。

表 2 用户对资源的评分信息

	resource ₁	resource ₃	resource ₃
user ₁	1	1	1
user ₂	5	5	5
user ₃	4	0	0

本文针对上述存在的问题提出了一种基于修正的相似性计算方法 (RS),将皮尔森 (PC) 与 RS 搭配使用,当 PC 不适合使用时,选择使用本文提出的 RS 计算相似性。本文定义用户 U 和用户 V 共同评分的学习资源集合用 $R_{U,V}$ 表示, $R_{u,i}$ 和 $R_{v,i}$ 分别表示用户 U 和用户 V 对资源 i 的评分, $r_{u,i}$ 和 $r_{v,i}$ 分别代表了用户 U 对项目 i 的评分以及用户 V 对项目 i 的评分, $\bar{r}_U$ 和 $\bar{r}_V$ 则分别代表了用户 U 和 V 对项目的平均评分,则用户 U 和用户 V 的相似度 SU, V 可按照如下公式计算。

$$SU,V = \begin{cases} PC = \frac{\sum_{i \in R_{U,V}} (r_{U,i} - \bar{r}_U)(r_{V,i} - \bar{r}_V)}{\sqrt{\sum_{i \in R_{U,V}} (r_{U,i} - \bar{r}_U)^2} \sqrt{\sum_{i \in R_{U,V}} (r_{V,i} - \bar{r}_V)^2}} & \text{当 } |R_{U,V}| \geq D \\ RS = \frac{\sum_{i \in R_{U,V}} F(R_{U,i}, R_{V,i})}{|R_{U,V}|} & \text{当 } |R_{U,V}| < D \end{cases} \quad (11)$$

$$F(R_{U,i}, R_{V,i}) = \frac{1}{d_i + 1} \quad (12)$$

其中, d_i 表示用户 U 和 V 对学习资源 i 评分的绝对距离, 两个用户间共同评分项目数为 $|R_{U,V}|$, 共同评分项目门限值 D 。当用户之间共同评分项目数大于门限值时, 使用 PC 计算, 当不满足门限值要求时则使用本文提出的 RS。通过本文提出的修正的相似度计算不仅可以控制相似度在 $[0, 1]$ 而且还在一定程度上减少了相似度计算误差的出现。

2.3 基于相似因子的相似度传递策略

2.3.1 相似因子

协同过滤推荐算法最关键的部分在于相似度的计算^[19], 因此本文将相似因子作为相似度计算准确性的评价标准。本文的相似因子定义如式 (13) 所示。

$$SF_{U,V} = \begin{cases} S(U,V) \times |R_{U,V}| & \text{当使用 PC 时} \\ S(U,V) \times |R_{U,V} - 0.1| & \text{当使用 RS 时} \end{cases} \quad (13)$$

其中, $SF_{U,V}$ 表示用户 U 和 V 之间的相似因子; $S(U,V)$ 表示用户 U 和 V 之间的相似度, $R_{U,V}$ 表示用户 U, V 之间共同评分的资源数量。显然, 用户间的共同评分项目越多, 用户间的相似因子越大, 相似度计算的准确性越高。

2.3.2 基于相似因子的相似度传递策略

在学习资源推荐系统中, 以用户为节点, 用户之间的相似度为边, 可构成一个相似度网络图。由于系统中用户对学习资源评价的不对称性, 所以相似度网络图中的边都是有向边。相似度网络图的构成步骤如下所述: a) 以用户为节点, 若节点间拥有共同评分资源, 则连接两节点构成一条边; b) 根据前文中计算得出的用户间的相似度和共同评分资源数为每条边分别赋予特征值。以用户 A 和 B 为例, 则 $\text{Sim}(A, B)$ 表示用户 A 和 B 之间的相似度, $R_{A,B}$ 表示用户 A 和 B 之间拥有的共同评分资源数量。

当用户 A 对学习资源 i 感兴趣时, 则首先遍历 A 的直接邻居, 如果 A 的直接邻居没有对学习资源 i 进行过评分, 那么就使用本文提出的相似度传递策略, 为 A 扩展相似用户。主要过程如下:

删除相似度小于 0 的边; 在相似度网络图中, 存在相似度小于 0 的情况, 在推荐系统中, 假设相似度为 S , 若 $S \geq 0$, 则推荐存在意义; 若 $S < 0$, 则推荐无意义。

为了保证只从目标节点出发可以访问每个节点, 优化相似度网络图为有向无环图。设置传递路径阈值 T , 保证相似度传递的路径长度不超过 T 。

按照下述公式, 计算基于相似度传递策略, 所有用户在传递阈值 T 约束下的相似因子。假设要计算基于相似度传递策略下用户 A 与 U 的相似因子 $ST_{A,U}$, 则计算公式如下:

$$ST_{A,U} = \frac{\sum_{x \in X} (SF_{A,x} \times SF_{x,U})}{\sum_{x \in X} SF_{A,x}} \quad (14)$$

其中: $SF_{A,x}$ 与 $SF_{x,U}$ 分别表示用户 A 与用户 x 之间的相似因子以及用户 x 与用户 U 之间的相似因子, 且 $x \in X$, X 表示在传递阈值 T 的约束下, 用户 U 的直接前驱用户节点集。

假设 $A \rightarrow G$ 中每个字母分别表示系统中的一个用户。以用户 A 为例, A 与 E, C, B 三个用户均有共同评分资源, 且数量分别为 12、4 和 4。因此分别连接 A 与 E, A 与 C, A 与 B 构成三条边, 每条边都具有两个属性, 代表相似度和共同评分资源数。其他用户使用相同的方法找到自己的邻居用户。最后构成以 $A \rightarrow G$ 为节点的相似度网络图, 如图 3 所示。

假设用户 G 对用户 A 感兴趣的学习资源进行了评分, 但用户 A 和 G 之间没有共同评分资源, 因此无法计算用户 A 与用户 G 的相似度, 从而也无法得知用户 G 对学习资源的评分是否对用户 A 有推荐意义。但通过用户 A, C 以及 G 之间的相似度传递, 可以预测用户 A 与 G 的相似度 $\text{sim}(A, G)$ 。

设传递路径阈值 $T=2$, 在相似度传递的路径长度不超过 T 的条件下, 节点 A 到 G 存在两条传递路径: $A \rightarrow E \rightarrow G$ 和 $A \rightarrow C \rightarrow G$ 。按照上述相似度传递过程优化后的相似度网络图如图 4 所示。

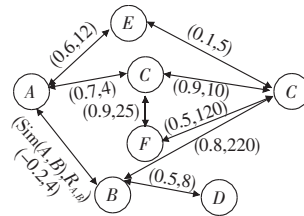


图3 相似度网络图

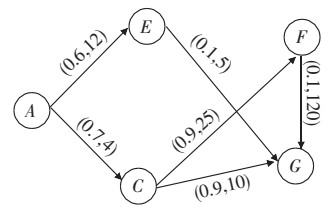


图4 优化后的相似度网络图

假设计算用户 A 在传递路径阈值 $S \leq T$ 约束下与用户 G 的相似因子 $ST_{A,G}$ 。按照式 (14) 计算结果如下: $ST_{A,G} = \frac{SF_{A,E} \times SF_{E,G} + SF_{A,C} \times SF_{C,G}}{SF_{A,E} + SF_{A,C}} = 2.88$ 。

2.4 基于相似度传递的评分预测

与传统算法中的评分预测计算方法相比, 本文基于相似度传递的评分预测方法用相似因子代替了传统计算方法中的相似度。通过本方法, 赋予与目标用户有高相似性且有较多共同评分资源数的用户更多权重。假设, $\bar{r}_U$ 表示用户 U 对项目的平均评分; $ST_{U,X}$ 是经过相似度传递后计算得出的相似因子; $r_{X,i}$ 表示用户 X 对项目 i 的评分; N_U 表示与用户 U 最相似的邻居集合; X 表示用户 U 最相似的邻居集合中的用户之一。基于相似度传递的评分预测计算方法如下:

$$P_{U,i} = \bar{r}_U + \frac{\sum_{x \in N_U} ST_{U,x} (r_{x,i} - \bar{r}_U)}{\sum_{x \in N_U} ST_{U,x}} \quad (15)$$

3 实验分析

3.1 数据集

本文的实验数据来源于重庆市集成示范计划项目中学习资源应用平台—壹笔。该学习平台每一个注册用户都至少对 10 个学习资源进行过学习行为。目前该学习应用平台的用户已经超过 20 000 人, 用户评分过的学习资源超过 3 000 个。本文实验使用的数据集是该应用截止到 2015 年 12 月份的真实数据集。随机抽取 1 000 个用户数据进行离线实验。

对于文中所述隐式评分的获取, 首先, 从后台中导出日志文件, 分析用户上网日志, 根据本文实验所需的四种行为 (点赞、评论、收藏、下载) 从日志文件中进行筛选; 其次, 对各用户的不同行为信息进行数值量化, 若存在该行为, 则将该行为置为 1, 反之, 置为 0; 最后, 根据本文 2.1 所述的隐式评分模型构

建过程,化隐为显,得到用户对学习资源的评分。

3.2 实验设计及分析

本文针对相似度计算的缺陷提出了基于修正的相似性度量方法(RS),为了验证本文提出的RS的有效性,分别与余弦相似性(Cosine)和皮尔森相关系数(Pearson)进行对比,设置邻居个数按照步长为5从5增加至50,然后计算使用不同相似性度量方法时协同过滤推荐的MAE。实验结果如图5所示。

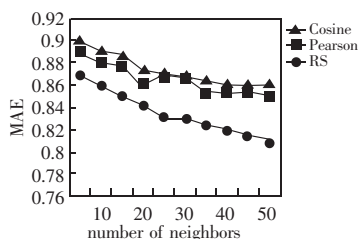


图5 RS与Cosine, Pearson对推荐结果的影响

由图5可以得出,在邻居个数不断增加的情况下,MAE的值都是呈递减趋势的。且本文提出的基于修正的相似性度量方法具有最小的MAE。由此可见,本文提出的RS方法相比传统的相似度计算方法在推荐质量上有一定的提高。

本文定义数据的稀疏等级为:在用户对学习资源的评分矩阵中,没有评分的矩阵元素占全部矩阵元素的比例。本文实验将分别设置稀疏等级0.2、0.4、0.6和0.8,则按照稀疏等级,随机的将评分矩阵中相应数量的数据置为0,表示未评分的资源。将实验数据分成训练集和测试集。将传递阈值 T 的值分别设置为1、2、3。结果如图6所示。

在图中,Spath1、Spath2、Spath3分别表示传递路径阈值 $T=1, 2, 3$ 的相似度传递策略。可以看出,当数据稀疏系数等级增大时,用户的相似用户个数在减少。而当使用本文提出的相似度传递策略时,当传递路径增大时,相似用户数的数量则比未传递时的相似用户数要大。在推荐系统中,相似用户的数量对预测评分的准确性有重要的影响,因此,本策略对传统算法的改进具有一定的意义。

设置改进后的算法传递路径阈值为2,图7展示了平均值推荐法、传统的协同过滤、基于社会信任的协同过滤以及本文提出的基于隐式评分和相似度传递的协同过滤算法的MAE,可以看出平均值推荐法的MAE由于未将个性化考虑其中,看起来并无太大波动,基于社会信任的以及本文的算法与传统的CF相比,MAE明显优于传统的CF。基于社会信任的CF与本文的算法MAE较为接近,实验中的基于社会信任的CF在缓解数据稀疏性问题中也有一定效果,但用户间的信任度往往缺乏一定可靠性,例如,用户A偏好文学类的学习资源,用户A信任B, B偏好计算机工程类学习资源,按照此信任方式推荐,也会产生一定误差,导致推荐不准确的问题。而本文算法,当相似度经过传递之后,扩大了用户的潜在推荐者,此外,经过相似度传递策略的推荐准确性在一定程度上也得到了提高。

4 结束语

针对学习平台中使用协同过滤推荐存在的数据稀疏和相似度计算问题,本文提出了一种隐式评分和相似度传递的学习资源推荐。通过获取用户行为来填充用户评分矩阵,以及使用相似度传递策略来扩大用户的潜在推荐者。实验结果表明,本文提出的推荐算法在数据稀疏的情况下具有较好的推荐质量。

因此可以将本文提出的新算法应用到学习平台中,为学习者推荐符合其学习兴趣的学习资源。但本文在考虑学习者个性偏好与资源推荐时效性过程中稍有欠缺,下一步工作可能引入时效性,用马尔科夫来考量资源的时效性问题,并进一步完善算法,以达到更好的推荐效果。

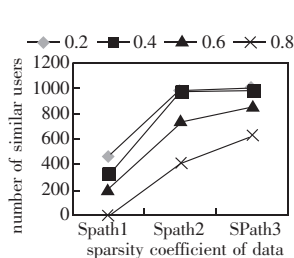


图6 传递路径阈值对相似用户数的影响

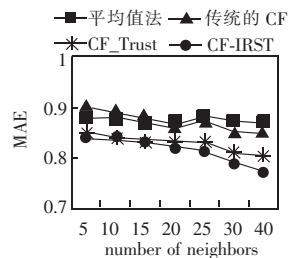


图7 不同协同过滤算法的MAE进行比较

参考文献:

- [1] 牛文娟. 基于协同过滤的学习资源个性化推荐研究[D]. 北京: 北京理工大学, 2015.
- [2] Yu Ruiguo, Cai Zhiyong, Du Xiuping, et al. The research of the recommendation algorithm in online learning[J]. *International Journal of Multimedia & Ubiquitous Engineering*, 2015, 10(4): 71-80.
- [3] 马宏伟, 张光卫, 李鹏. 协同过滤推荐算法综述[J]. *小型微型计算机系统*, 2009, 30(7): 1282-1288.
- [4] Cheng Weijie, Yin Guisheng, Dong Yuxin, et al. Improved collaborative filtering recommendation via non-commonly-rated items[C]//Proc of the 8th International Conference on Internet Computing for Science and Engineering. [S. l.]: IEEE Press, 2015: 56-60.
- [5] 代金龙. 协同过滤算法中数据稀疏性问题研究[D]. 重庆: 重庆大学, 2013.
- [6] 孙小华. 协同过滤系统的稀疏性与冷启动问题研究[D]. 杭州: 浙江大学, 2005.
- [7] 丁永刚, 张馨, 桑秋侠, 等. 融合学习者社交网络的协同过滤学习资源推荐[J]. *现代教育技术*, 2016, 26(2): 108-114.
- [8] Lu Zhongqi, Dou Zicheng, Lian Jianxun, et al. Content-based collaborative filtering for news topic recommendation[C]//Proc of AAAI Conference on Artificial Intelligence. 2015: 217-223.
- [9] Guo Guibing, Zhang Jie, Thalmann D. Merging trust in collaborative filtering to alleviate data sparsity and cold start[J]. *Knowledge-Based Systems*, 2014, 57(2): 57-68.
- [10] Satnam A. Collective intelligence in action[M]. [S. l.]: Manning Publications, 2009: 349-350.
- [11] 黄洋. 基于聚类和项目类别偏好的协同过滤推荐算法研究[D]. 杭州: 浙江理工大学, 2014.
- [12] 党博, 姜久雷. 基于评分差异度和用户偏好的协同过滤算法[J]. *计算机应用*, 2016, 36(4): 1050-1053.
- [13] 陈洪涛, 肖如良, 邓新国. 一种新的基于社交关系的相似度传播式推荐算法[J]. *小型微型计算机系统*, 2015, 36(5): 1073-1077.
- [14] 嵇晓声, 刘宴兵, 罗来明. 协同过滤中基于用户兴趣度的相似性度量方法[J]. *计算机应用*, 2010, 30(10): 2618-2620.
- [15] Guha R, Kumar R, Raghavan P, et al. Propagation of trust and distrust[C]//Proc of the 13th International Conference on World Wide Web. 2004: 403-412.
- [16] 刘慧婷, 熊瑞瑞, 赵鹏. 基于信任关系重建和社交网络传递的推荐算法[J]. *计算机工程*, 2016, 42(1): 174-179.
- [17] 王永固, 邱飞岳, 赵建龙, 等. 基于协同过滤技术的学习资源个性化推荐研究[J]. *远程教育杂志*, 2011(3): 66-71.
- [18] Niu Wenjuan, Niu Zhendong, Tang Shiping, et al. A learning resource recommendation method combining user sequential interaction with collaborative filtering[C]//Proc of Modelling, Identification and Control; Computational Intelligence Conference. 2015: 130-145.
- [19] 祝奇伟, 陈家琪. 一种改进相似度计算方法的协同过滤推荐算法[J]. *信息技术*, 2015(3): 13-16.