

doi:10.14132/j.cnki.1673-5439.2017.03.014

融合用户相似度与评分信息的协同过滤算法

乔 雨,李玲娟

(南京邮电大学 计算机学院,江苏 南京 210023)

摘要:推荐系统利用机器学习的技术进行信息过滤,准确地定位用户需要的信息,并且能够预测用户对目标项目的喜好程度。但是由于新用户和新项目的存在,传统的协同过滤推荐系统面临着冷启动问题的挑战。为了解决协同过滤推荐系统中用户冷启动问题,设计了融合用户相似度与评分信息的协同过滤算法(SR-CF)。该算法用基于人口统计学的推荐算法找出用户基本信息之间的相似度,再根据最速下降法对用户评分矩阵进行更新,从而产生对目标用户的推荐。基于 MoiveLens 公开数据集的实验结果表明,所设计的算法在保证推荐准确率的同时提高了推荐的覆盖率,能有效解决用户冷启动问题。

关键词:推荐系统;用户冷启动;人口统计学;评分信息

中图分类号:TP311 **文献标志码:**A **文章编号:**1673-5439(2017)03-0100-06

Collaborative filtering algorithm based on user similarity and rating information

QIAO Yu, LI Lingjuan

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Recommendation systems by using machine learning techniques to filter information can accurately locate user information, and predict whether a user would like a given resource. As traditional collaborative filtering systems have to deal with the cold-start problems caused by new users and items, a collaborative filtering algorithm based on user similarity and rating information(SR-CF) is designed to solve the new-user cold-start problem. The algorithm can calculate similarities between users' basic information by the recommendation algorithm based on demography, and then update the matrix of ratings information with the steepest descent method to generate recommendations for users. The experimental results on MoiveLens dataset demonstrate that SR-CF can improve the accuracy of the recommendations and has higher coverage. Thus it can effectively deal with the cold-start problems caused by new-users.

Keywords: recommendation systems; new user cold-start problems; demography; rating information

目前,推荐系统已成为处理信息超载的不可或缺的手段,其中以协同过滤推荐算法应用最为广泛,该算法是利用用户对项目的评分数据构建用户信息表,通过查询用户信息表,找出与当前用户兴趣最相似的用户邻居,并根据该用户邻居的兴趣偏好为当前用户推荐其最可能感兴趣的项目^[1]。但是,当系

统中出现新用户时,由于新用户行为信息稀少、评分数据不足,系统很难为其产生合适的推荐,这就是所谓的用户冷启动问题。

冷启动问题一直是传统协同过滤推荐系统发展过程中的难点,它严重影响了推荐系统的推荐质量,并在一定程度上影响了用户对推荐系统的信任

度^[2]。目前针对冷启动问题已提出的解决方法有随机推荐法、平均值法、相似度度量法等,也有 Ahn^[3]提出对相似度进行改进的推荐算法,以及在协同过滤算法中引入用户的人口统计学信息以及社会网络信息的方法^[4]。

本文在基于用户评分模型的协同过滤推荐算法基础上,结合人口统计学的方法,改进设计了一种融合用户相似度与评分信息的混合推荐算法(A Collaborative Filtering Algorithm Based on User Similarity and Rating Information, SR-CF), 并利用 MoiveLens 数据集进行算法性能验证,得出所设计的算法能够在用户冷启动情况下进行准确推荐的结论。

1 相关工作

1.1 基于人口统计学的推荐算法

推荐系统除了利用用户-项目评分信息之外,还记录了注册用户和项目的基本信息。美国网络对话网站(www. cyberdialogue. com) 的研究结果显示 80% 的用户为获得网站定制的个性化信息不介意向 web 站点提供自己的性别、教育背景和兴趣等不太敏感的信息,并且越来越多的网站可以提供 Facebook、QQ 等社交网络的账号直接登录的功能,这种方式可以向推荐系统提供更多的完善的用户信息。而基于人口信息统计学的推荐算法正是将用户的性别、年龄、职业、爱好等信息融入到对用户相似度的检测中^[5-6],建立用户特征模型,根据这些特征计算用户之间的相似度,从而寻找到相似邻居,基本流程如图 1 所示。



图 1 基于人口统计学的推荐算法流程

此算法的优势是不需要历史数据,也不依赖物品的属性,没有用户的冷启动问题;但是算法比较粗糙,信息过滤不能达到很高的精确度,推荐效果很难令人满意,所以只适用于简单的推荐。因此需要对其进行改进或与其它算法混合使用,从而达到扬长避短的目的。

1.2 基于用户评分模型的协同过滤算法

矩阵分解是协同过滤算法中一种十分有效的方法,它利用数学或机器学习的方法从用户-项目的评分矩阵中分解出潜在特征来进行评分预测,从而实现高维数据的降维,并且在缓解数据稀疏性造成的推荐结果准确度下降问题方面有非常好的效果^[7]。

表 1 为用户-项目评分示例,评分范围为 1 ~ 5

分,“—”表示没有评分。该表描述了 6 个用户(User1, User2, User3, User4, User5, User6)对 4 个物品(Item1, Item2, Item3, Item4)的评分信息;可将表 1 抽象成矩阵 R_u ,利用矩阵分解的方法对 R_u 进行更新迭代,并对未评分的项进行预测,按预测结果产生推荐。

表 1 用户对项目的评分示例

用户	Item1	Item2	Item3	Item4
User1	3	5	—	3
User2	—	4	—	4
User3	2	3	1	—
User4	—	—	5	4
User5	4	—	4	2
User6	5	3	—	—

$$R_u = \begin{bmatrix} 3 & 5 & 0 & 3 \\ 0 & 4 & 0 & 4 \\ 2 & 3 & 1 & 0 \\ 0 & 0 & 5 & 4 \\ 4 & 0 & 4 & 2 \\ 5 & 3 & 0 & 0 \end{bmatrix}$$

其中对于缺失评分的预测为该方法的关键,这个问题可以转化为基于机器学习的回归问题,也就是连续值的预测。矩阵分解如式(1)所示,其中 R_p 为矩阵分解得出的预测评分, R_u 是原始评分。矩阵分解的方法是将原始评分矩阵 R_u 分解为 P 、 Q 两个子矩阵,其中 M 为用户数, N 为项目数; p_{ik} 是用户 i 对属性 k 的喜欢程度, q_{kj} 是项目 j 对于属性 k 的偏移程度; k 是算法中的一个参数,通过调节来确定取值,通常在 1 ~ 100 之间; $K \ll M, N, i \in [1, M], j \in [1, N]$ 。经过多次迭代更新后,通过计算 R_u 与 R_p 之间的误差大小来进一步判断分解的效果,误差越小,效果越好,这就需要采用一定的方法调整参数使得误差在一定的范围内。

$$R_p \approx R_u = P_{M \times K} \times Q_{N \times K}^T = \sum_{k=1}^K p_{ik} q_{kj} \quad (1)$$

2 SR-CF 算法设计

2.1 用户相似度的计算方法

2.1.1 用户相似度的定义

本文对基于人口统计学的推荐算法进行改进,为不同的用户属性赋予不同的权值。比如存在这样的情况,用户 s 和 t 的年龄差距较大,但是他们都喜

爱推理类的电影,因此我们认为个人的兴趣爱好这一属性的权值应大于年龄属性。用 s 和 t 表示两个不同的用户,用 A_1 到 A_n 表示用户的 n 个属性,用户之间的相似度可以用式(2)计算

$$\text{sim}(s, t)_u = w_1 A_1(s, t) + w_2 A_2(s, t) + \cdots + w_n A_n(s, t),$$
$$\sum_i^n w_i = 1 \tag{2}$$

其中, $\text{sim}(s, t)_u$ 表示用户 s 和 t 之间的信息相似度,被本文定义为用户相似度; $A_i(s, t)$ 表示用户 s 和 t 之间属性 i 的相似度; w_i 表示属性 i 所占的权值。

2.1.2 属性相似度的计算方法

本文选取年龄、性别、职业和爱好4种特征作为考虑的范围,将各属性信息预处理成数字型表示法的形式。其中年龄属性,我们以5岁为增量,比如年龄32岁可记为7,年龄56岁记为12。利用欧几里得距离计算用户 s 与用户 t 之间年龄的相似度,如式(3)所示。其中 $A(s, t)$ 表示用户 s 与用户 t 在年龄上的相似度, x_s 、 x_t 分别为用户 s 和 t 的年龄分段值。

$$A(s, t) = e^{-|x_s - x_t|} \tag{3}$$

性别是一种对称的二元属性^[8],即两个状态都是同等重要的。男性记为1,女性记为0,这样根据用户性别的取值可以得到一个两行两列的列表,如表2所示。这里我们用 $S(s, t)$ 表示用户 s 与用户 t 在性别上的相似度,如式(4)所示,其中 a 是对象 s 和 t 都取1情况的属性值, b 是对象 t 取1和对象 s 取0情况的属性值, c 是对象 t 中取0和对象 s 取1情况的属性值, d 是对象 s 和 t 都取0情况的属性值。

表2 二元属性取值

对象 s	对象 t		Sum
	0	1	
0	a	b	$a + b$
1	c	d	$c + d$
Sum	$a + c$	$b + d$	f

$$S(s, t) = \frac{a + d}{a + b + c + d} \tag{4}$$

而职业、爱好属于标签类型的属性,即用某些字段来描述这一属性,对这类属性,我们利用式(5)来计算用户 s 与 t 之间的相似度。式(5)中 k 表示用户 s 与 t 共同拥有的同一属性的标签个数, n 表示该属性可供选择的标签个数。

$$\text{sim}(s, t) = \frac{k}{n} \tag{5}$$

2.1.3 用户相似度的计算

对用户 u 及其 k 个邻居集合 U_k ,用式(2)计算用户的信息相似度,并且保存 Top-N 个与目标用户 u 相似的用户集合 U_k 。在用户人口统计特征方法下,特征越多,越能准确预测用户兴趣。

2.2 用户评分信息的处理

当把用户对物品的评分信息抽象成矩阵后,本文采用基于矩阵奇异值分解(Singular Value Decomposition, SVD)方法对评分矩阵进行处理,并以减少时间复杂度为目的,对此加以改进。

奇异值分解是矩阵分解方法的一种,不仅具有可扩展性,还具有良好的预测准确度;它的不足之处在于时间复杂度非常高,不适合数据量非常大的情况。因此出现了多种针对奇异值矩阵分解改进的方法及应用,比如 Hofmann^[9]提出的概率潜在语义分析(PLSA), Ruslan 等^[10]提出的奇异值在神经网络上的应用等。本文对该方法的改进则是通过最速下降法进行近似计算,从而减少其所需的时间。

我们利用式(6)来表示计算出的预测分数^[11],其中向量 p_u 表示用户 u 潜在的 n 维特征,向量 q_m 表示物品 m 的潜在的 n 维特征,将用户-项目评分矩阵 R_u 分解为矩阵 P 和 Q^T 的乘积^[12],使用的近似计算方法是对式(7)进行最小化。

$$r_{um} = p_u \times q_m^T \tag{6}$$

$$\min_{p_u \in P^{n \times r}, q_m \in Q^{r \times s}} \sum (r_{um} - p_u \times q_m^T)^2 + \theta_1 (|p_u|^2 + |q_m|^2) \tag{7}$$

其中, r_{um} 表示矩阵 A 中第 u 行第 m 列的已知评分值, p_u 为用户矩阵 P 中的第 u 行, q_m 为物品矩阵 Q 中的第 m 列, $|p_u|^2$ 表示矩阵 P 中所有元素的平方和, $|q_m|^2$ 表示矩阵 Q 中所有元素的平方和; $\theta_1 (|p_u|^2 + |q_m|^2)$ 是为避免相对数据量比较小的项目和数据过拟合现象而添加的惩罚因子, θ_1 为自定义的实数,通常情况下取值在0到50之间,且不为0。这里我们用最速下降法来达到最小化的目的,求出局部最小值。

针对上述最小化公式,矩阵 P 和 Q 中的元素的偏导数公式如式(8)与式(9)所示,其中 μ 为迭代步长,这里通过设置阈值 ε 作为迭代更新停止条件,当公式(7)的值小于 ε 时停止迭代。在完成迭代后得到的 P 、 Q 两个矩阵即为目标分解矩阵, $p_u \times q_m^T$ 即为目标用户对特定项目的预测评分。

$$p_{uk+1} \leftarrow p_{uk} - \mu \frac{\partial f}{\partial p_u} \tag{8}$$

$$q_{mk+1} \leftarrow q_{mk} - \mu \frac{\partial f}{\partial q_m} \quad (9)$$

将预测的评分填充到用户-项目的原始评分矩阵中,从而缓解原始矩阵的稀疏性。本文在填充评分的同时加入判断,即检测用户对该物品是否存在评分记录,如果存在就保留实际评分;如果不存在则利用预测评分进行填充^[12],如式(10)所示。最后,将评分较高的物品推荐给目标用户。

$$R_{ai} = \begin{cases} r_{ai}, & r_{ai} \neq 0 \\ r_{pi}, & r_{ai} = 0 \end{cases} \quad (10)$$

2.3 SR-CF 算法描述

在 SR-CF 算法中,令 u 表示活跃的用户, U_k 表示目标用户的相似邻居集合, U 表示用户基本信息矩阵, R_u 表示用户-项目评分信息矩阵, 矩阵 P 和 Q^T 为评分矩阵 R_u 分解后的子矩阵; $\text{sim}(s, t)_u$ 表示用户的信息相似性, R_{ai} 表示项目-评分矩阵更新后的评分, 推荐结果列表 L_r 。具体过程描述如下:

输入: 用户基本信息矩阵 U , 用户-项目评分信息矩阵 R_u ;

输出: 用户推荐列表 L_r 。

步骤:

Step1: 根据式(2)至式(5)计算用户 u 年龄、性别等属性的相似度, 并赋予不同的权值。

Step2: 调整用户各属性所占的权值, 求出用户之间的相似度 $\text{sim}(s, t)_u$, 以相似度误差最小的结果作为记录, 并保存用户 u 的相似邻居集合 U_k 。

Step3: 根据 2.2 小节中的增量奇异值矩阵分解的方法对用户-项目评分矩阵 R_u 进行迭代, 得出子矩阵 P 、 Q^T 。

Step4: 利用 $r_{um} = p_u \times q_m^T$ 来表示对未评分项目的预测分数。

Step5: 当式(7)的值小于阈值 ε 时, 停止迭代; 否则转到 Step2。

Step6: 将用户 u 的相似邻居集合 U_k 与用户-项目评分矩阵 R_u 进行映射, 并利用式(10)的方式更新用户原始评分矩阵 R_u 为 R_{ai} 。

Step7: 根据 Top-N 算法对计算出的预测评分进行排序, 形成推荐列表 L_r 。

Step8: 将推荐列表中的物品展示给用户 u 。

SR-CF 算法流程如图 2 所示。

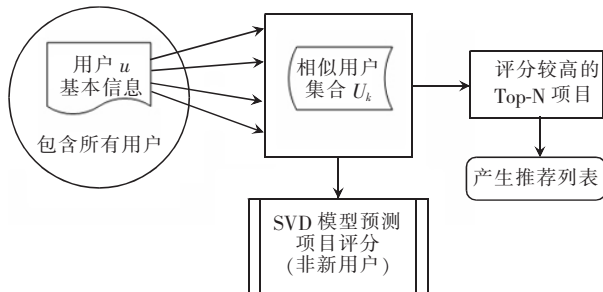


图2 SR-CF 算法推荐流程

2.4 SR-CF 算法时间复杂度分析

在算法实现推荐列表输出的过程中, 记 m 为用

户的特征数, n 为项目的特征数。在 Step1 中, 计算用户之间特征相似度的时间复杂度为 $O((m^2 - m)/2)$; 在 Step2 中, 计算用户之间的相似度 $\text{sim}(s, t)_u$, 受系统所需参考的用户集合中的数量 c (c 为常数) 影响, 时间复杂度为 $O(c(m^2 - m)/2)$; 在 Step3 中最好的情况是迭代一次后, 式(7)的值就已经小于阈值 ε , 最坏的情况是循环计算了矩阵的所有行列才使得式(7)在阈值之内, 此时的时间复杂度为 $O(mn)$ 。由于 $c \ll n$, $c \ll m$, 所以 SR-CF 算法总时间复杂度为 $O(m^2 + mn)$ 。

3 实验及结果分析

3.1 实验环境与数据集

实验环境是: 英特尔酷睿 i7 处理器, 8 G 内存, Windows 8.1 的 64 位操作系统中通过虚拟机 VM-Ware 10 安装 Ubuntu 32 位操作系统, 以 hadoop 2.7.2 为基础搭建 Spark 1.2.0 的环境, 利用 scala 编程语言对算法进行实现, 并利用了 MLlib 库。

本文采用 MoiveLens 公开数据集验证算法的有效性, MoiveLens 是一个基于 Web 的研究型推荐系统, 它接受用户对电影的评价, 并提供相应的电影推荐列表^[13]。目前该系统的用户已超过 240 000 人, 用户评分记录超过 22 000 000 条。本文采用 ml-1m 的数据集, 它包含了 943 位用户对 1 682 部电影做出的 100 000 条评分记录, 评分等级为 1(差)至 5(非常好), 数据的稀疏性为 93.7%。

3.2 推荐效果的度量标准

平均绝对误差。在 MoiveLens 数据集中, 我们将系统对电影预测的评分与用户对该电影的实际评分进行比较, 并采用平均绝对误差 MAE ^[14] 来计算推荐系统产生的预测与用户实际评分的误差, 以此来衡量推荐结果的精确性。平均绝对误差 MAE 的计算公式如下:

$$MAE = \frac{\sum_{i=1}^n |r_{pi} - r_{ai}|}{n} \quad (11)$$

其中, r_{pi} 、 r_{ai} 分别表示预测评分与实际评分, n 表示所有具有实际值的项目个数。MAE 越小, 推荐精度越高。

覆盖率。采用覆盖率^[15]来表示推荐算法的覆盖能力, 即所有能够预测的项目占项目总数的比例。本文设为用户 u 提供的预测值集合为 $P_u = \{p_{u,j} | j=1, 2, \dots, n\}$, 其中 $p_{u,j} \neq 0$ 个数为 k , n 为集合 P_u 中总的项数, 则用户 u 的覆盖率 C_u 为:

$$C_u = \frac{k}{n}$$

(12)

3.3 实验过程与结果分析

实验采用文献[16]的方式随机选择数据集的80%作为训练集,20%作为测试集,同时为了测量参数的敏感度,将训练集细分为测试集与训练集进行交叉验证,具体做法是通过每次随机选择不同的测试样本与训练样本,对结果采用平均值。将本文的SR-CF算法与基于用户的协同过滤算法(User-Based Collaborative Filtering, UB-CF)和基于物品的协同算法(Item-Based Collaborative Filtering, IB-CF)作比较,同时采用实验结果最好的参数值进行记录。

确定邻域的最优值 k 。实验从10到50,以5为一个梯度设置用户的邻居数,并计算相应预测评分的平均绝对误差 MAE,结果如图3所示。

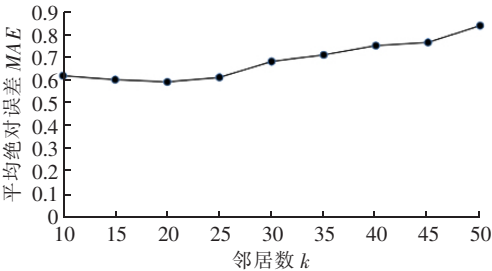


图3 SR-CF 算法不同邻居数 k 相应的 MAE

从图3中可以看出在 $k=20$ 时,SR-CF 算法的预测平均绝对误差值相对达到最小。因此我们选择邻居数为20,做更进一步的研究。

算法预测精确度分析。图4为SR-CF算法与其它两种算法的比较,我们通过计算不同算法的预测评分与实际评分的平均绝对误差,可以看出 UB-CF 算法的精度随着邻居数的增加而提高;而 IB-CF 算法在用户冷启动方面并没有突出的表现;而 SR-CF 算法在用户邻居数发生变化时候反应比较敏感,预测精确度方面也明显高于传统的协同过滤算法。

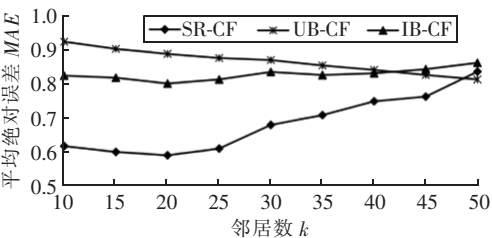


图4 三种算法的不同邻居数 k 取值相应的 MAE

算法覆盖率分析。算法对数据集中项目的覆盖率如表3所示,本文随机抽选200个用户作为新用户测试集,分别考察在不同的邻居数情况下算法的覆盖率。

表3 不同算法在不同邻居数下覆盖率 %

算法	邻居数				
	15	20	25	30	35
IB-CF	9.8	12.1	13.4	15.5	16.9
UB-CF	11.3	13.6	15.9	18.2	21.3
SR-CF	12.8	14.8	18.3	21.1	23.6

从表3中可以看出 SR-CF 算法在用户数较少时与 UB-CF 算法的覆盖率较为相近,但是随着用户数量的增加,SR-CF 算法在覆盖率方面相较于 UB-CF 算法具有一定的优势;同样地,IB-CF 算法在以用户为中心的推荐场景中覆盖率也无特别的表现。由此我们可以认为本文的 SR-CF 算法因融合了人口统计学的用户相似度思想,在新用户的推荐覆盖率方面优于传统的协同过滤推荐算法。

4 结束语

本文以解决协同过滤推荐系统中用户冷启动问题为目标,设计了融合用户相似度与评分信息的协同过滤算法 SR-CF,该算法基于人口信息统计学的思想对新用户的基本信息进行相似度检测,计算得出新用户的相似邻居集合;同时对老用户-项目的评分矩阵进行奇异值矩阵分解,经过多次迭代处理后得出预测评分值,并采用一定的方法对用户的原始评分矩阵进行更新,由此产生相应的推荐列表。本文通过实验证明了 SR-CF 算法在保证推荐的准确性的同时,相较于传统的协同过滤算法提高了覆盖率,在一定程度上解决了新用户的冷启动问题,并且能够缓解数据的稀疏性问题。

参考文献:

[1] LIU N N, MENG X, LIU C, et al. Wisdom of the better few: cold start recommendation via representative based rating elicitation[C]// Proceedings of the 15th ACM Conference on Recommender Systems. 2011:37 – 44.

[2] CHANG S, HARPER F M, TERVEEN L. Using groups of items for preference elicitation in recommender systems [C]// Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. 2015:1258 – 1269.

[3] AHN H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1):37 – 51.

[4] SEDHAIN S, SANNER S, BRAZIUNAS D, et al. Social collaborative filtering for cold-start recommendations[C]// Proceedings of the 8th ACM Conference on Recommender

- Systems. 2014:345–348.
- [5] BURKE R. Hybrid recommender systems: survey and experiments[J]. User Modelling and User-Adapted Interaction, 2002, 12(4):331–370.
- [6] GOGNA A, MAJUMDAR A. A comprehensive recommender system model: improving accuracy for both warm and cold start users[J]. IEEE Access, 2015, 3:2803–2813.
- [7] BOBADILLA J, ORTEGA F, HERNANDO A, et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46(1):109–132.
- [8] 赵文涛,王春春,成亚飞,等. 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, 33(12):1–6.
ZHAO Wentao, WANG Chunchun, CHENG Yafei, et al. Collaborative filtering algorithm based on multiple attributes and interests of users[J]. Application Research of Computers, 2016, 33(12):1–6. (in Chinese)
- [9] HOFMANN T. Latent semantic models for collaborative filtering[J]. ACM Transactions on Information System, 2004, 22(1):89–115.
- [10] RUSLAN S, ANDRIY M, GEOFFREY H. Restricted Boltzmann machines for collaborative filtering[C]//Proceedings of the 24th International Conference on Machine Learning. 2007:791–798.
- [11] YEHUDA K. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008:426–434.
- [12] GHAZANFAR M A, PRUGELBENNETT A. A scalable, accurate hybrid recommender system[C]//Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. 2010:94–98.
- [13] KUPISZ B, UNOLD O. Collaborative filtering recommendation algorithm based on Hadoop and Spark[C]//IEEE International Conference on Industrial Technology(ICIT). 2015:1510–1514.
- [14] 孟祥武,刘树栋,张玉洁,等. 社会化推荐系统研究[J]. 软件学报, 2015, 26(6):1356–1372.
MENG Xiangwu, LIU Shudong, ZHANG Yujie, et al. Research on social recommender systems[J]. Journal of Software, 2015, 26(6):1356–1372. (in Chinese)
- [15] WU Q, LIU S, MIAO C, et al. A social curiosity inspired recommendation model to improve precision, coverage and diversity[C]//IEEE/WIC/ACM International Conference on Web Intelligence (WI). 2016:240–247.
- [16] 郭弘毅,刘功申,苏波,等. 融合社区结构和兴趣聚类的协同过滤推荐算法[J]. 计算机研究与发展, 2016, 53(8):1664–1672.
GUO Hongyi, LIU Gongshen, SU Bo, et al. Collaborative filtering recommendation algorithm combining community structure and interest clusters[J]. Journal of Computer Research and Development, 2016, 53(8):1664–1672. (in Chinese)

作者简介:



乔雨(1993 –),女,江苏扬州人。南京邮电大学计算机学院硕士研究生。主要研究方向为数据挖掘与推荐系统。