

结合商圈位置区域模型的商品推荐算法^①



陈思亦, 何利力, 郑军红

(浙江理工大学 信息学院, 杭州 310018)
通讯作者: 陈思亦, E-mail: 878752790@qq.com

摘要: 为解决电子商务冲击下线下销售萎靡及互联网“信息爆炸”下用户挑选商品时耗时耗力等问题, 本文引入商圈位置区域模型, 即基于圆形过滤方法与改进的基于分区的 DBSCAN 密度聚类算法对浙江省某一行业全域范围内 25 万商户订单数据进行地理位置特征分析, 并结合时间衰减参数进行传统推荐算法改进, 提出了面向商圈流行度的商品推荐算法与面向商圈相似度的协同过滤算法. 实验结果表明, 算法在推荐准确率上明显优于传统推荐算法, 且一定程度上缓解了冷启动和推荐商品惊喜度不足的问题, 有其实用价值与研究意义.

关键词: 商圈位置区域模型; 时间衰减参数; 流行度; 相似度; 推荐算法

引用格式: 陈思亦, 何利力, 郑军红. 结合商圈位置区域模型的商品推荐算法. 计算机系统应用, 2019, 28(8): 136-141. <http://www.c-s-a.org.cn/1003-3254/6966.html>

Commodity Recommendation Algorithm Based on Location Area Model of Business-circle

CHEN Si-Yi, HE Li-Li, ZHENG Jun-Hong

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In order to solve the problems of embarrassing sales under the impact of e-commerce and time consuming and exhausting effort of users when choosing the commodity under the "information explosion" of the Internet, this paper introduces the location model of the business-circle based on the circular filtering method and the improved partition-based DBSCAN density clustering algorithm for Zhejiang Province. The geographic location characteristics of the 250 000 merchants' order data in a certain industry of Zhejiang Province are analyzed, and the traditional recommendation algorithm is improved by combining the time decay parameters. The commodity recommendation algorithm for the popularity of the business-circle and the collaborative filtering algorithm for the similarity of the business-circle are proposed. The experimental results show that the algorithm is superior to the traditional recommendation algorithm in terms of recommendation accuracy rate, and to some extent, it alleviates the problem of insufficient cold start and recommended product surprise, which has its practical value and research significance.

Key words: business-circle location area model; time decay parameter; popularity; similarity; recommendation algorithm

引言

近年来, 随着电子商务的飞速发展, 人们的传统购物消费习惯受到了巨大冲击, 市场环境也因此变得更加错综复杂, 线下销售呈现出日渐萎靡的趋势. 而伴随着“智慧门店”、“智慧快闪”、“智慧商圈”等全渠道新

零售模式的出现, 零售商户的线下业务开始融合数字化技术, 加速了零售模式的商业转型, 出现回暖. 由于“智慧商圈”、手机天猫中“逛商圈”等的营销模式及商业活动的推出, 商圈作为企业营销过程中的基础环节, 发挥了重要作用. 因此面对海量数据下的企业信息处

① 基金项目: 浙江省科技厅 (重大) 项目 (2015C03001)

Foundation item: (Major) Project of Science and Technology Bureau, Zhejiang Province (2015C03001)

收稿时间: 2018-11-07; 修改时间: 2018-12-04, 2019-01-15; 采用时间: 2019-01-31; csa 在线出版时间: 2019-08-08

理, 本文提出将商圈与传统个性化推荐相结合, 以满足互联网用户和大数据环境下个人用户精准营销的要求, 挖掘用户潜在需求, 进行面向用户的商品智能个性化推荐。

目前商品推荐领域的研究已经较为成熟, 并能得到一定的推荐效果。如在个性化推荐中基于用户协同过滤算法的优化, 通过预先计算用户相似度函数, 取出匹配最高的 n 个用户的数据生成推荐, 在推荐的计算速度上有着显著提高^[1]。对于用户消费行为信息的分析, 采用用户聚类的方法能有效解决协同过滤推荐的“稀疏性”问题。而在缺少用户消费行为信息的情况下, 基于商品领域知识的交互式推荐系统的研究, 能一定程度上解决推荐系统的冷启动问题和隐私保存问题^[2]。除了协同过滤算法与基于知识的商品推荐, 常见的个性化推荐算法还有基于内容的推荐^[3]、基于关联规则的推荐和混合推荐等^[4]。而在目前云计算与人工智能的发展应用下, 以深度学习和神经网络为基础的基于模型的推荐算法^[5-7]逐渐成为了研究的热点, 但目前尚未成熟。

为避免目前常规推荐算法推荐质量差、推荐效率低等问题, 本文提出基于商圈位置区域模型的商品推荐算法, 主要包含 4 方面的研究。1) 通过对全域数据采集下的商户订单数据的分析, 以商圈这一商业集群模式进行商品推荐领域划分, 基于地理位置特征建立商

圈位置区域模型。2) 针对现有推荐算法, 结合位置区域特征, 提出引入了商圈位置区域模型的面向流行度的推荐算法与面向相似度的改进协同过滤算法, 建立起浙江范围内某一行业的商圈商品推荐模型。3) 进行模型参数调整, 通过实验结果分析与对比, 验证了基于商圈位置区域模型的推荐算法相比传统推荐算法在推荐质量与推荐效率上的优化。4) 应用商圈位置模型, 提出以商圈相似性协同过滤算法为基础的新品投放策略, 在一定程度上加强企业对市场环境的预判与掌控。

1 数据采集与预处理

原始数据: 本文所用的实验数据采集于某行业营销业务系统, 通过相关 ETL 过程从系统数据库采集得到 2017 年所有商户基本数据、订单交易数据及商品基本数据, 并部署到数据库内进行后续的分析研究。

实验数据: 选取交易数据中浙江地区的 25 万户商户的交易记录作为实验主要数据集, 以 1-10 月的交易数据作为训练集, 11-12 月数据作为测试集。对数据集中商户位置数据及交易位置数据中的文地址, 采用百度地图 API 进行地理位置文本与经纬度的转换, 便于后续的位置区域划分。实验用到的商圈数据来源于美团全国城市商圈数据中的浙江省商圈数据 (含 777 个商圈), 相关样本数据示例如表 1 所示。

表 1 商圈数据示例

id	province_name	city_name	area_name	trade_name	lon	lat
3022	浙江省	杭州市	西湖区	天目山	120.097 482	30.263 725
3029	浙江省	杭州市	西湖区	灵隐	120.106 286	30.237 053
3034	浙江省	杭州市	西湖区	西湖	120.130 203	30.259 324

2 基于商圈位置区域的商品推荐算法

2.1 商圈位置区域模型

商圈作为一种特殊的商业集群模式, 与零售商户相辅相成, 一个合适的商圈, 能够增强商店吸引力, 扩大客户群, 同时强化规模优势, 提高经济效益。因此, 如何建立一个合理的商圈位置模型直接关系到推荐结果的好坏。早期商圈测评模型主要以区域 (或城市) 为单元, 结合人口、距离、零售面积、经济环境、消费比例等因素, 并进行实地考察来划分区域边界, 但具有较高的实施成本, 且考虑因素众多, 划分结果较差。近几年, 有不少学者考虑将空间聚类算法加入到商圈划分中去, 然而, 常见的几种传统聚类算法在一定程度上均

存在聚类结果不够准确、时间复杂度较大的问题。因此, 有学者考虑将空间自相关算法、核密度估计算法等方法纳入聚类算法中, 以实现商圈的划分。其中, 较为典型的有郝斌等人^[8]提出的多维特征融合的商圈划分方法, 将 *K-means* 空间聚类算法、空间自相关算法和核密度估计等算法融入商圈模型中, 但由于 *K-means* 算法的聚类结果主要取决于聚类中心的选取, 无法保证其聚类个数能收敛至全局最优, 使得在商圈区域划分时需要多次设定该值, 并从中选取最好结果, 具有一定的时间复杂度且随机性较大, 在很大程度上影响商圈划分结果。

为减少初始聚类中心对聚类结果带来的影响, 本

文主要采用基于分区的 DBSCAN 聚类算法, 算法主要思想是: 只要聚类空间中相邻区域数据样本点的密度不小于一定阈值就继续聚类, 可将密度足够大的相邻区域进行连接, 具有聚类速度快、有效处理噪声点和发现任意形状的空间聚类的优点. 由于传统 DBSCAN 算法在参数 Eps 和 $MinPts$ 的选择上存在不足^[9], 本文将结合核密度估计方法^[10]进行改进, 以实现参数的自适应选取.

本文通过将交易数据进行地图投影, 研究地域偏好差异与消费水平、消费习惯方面的关系, 发现其能呈现明显的地理位置特征^[11], 以此对市场进行位置区域划分, 划定商圈范围, 建立商圈位置区域模型, 进行以商圈为主体的商品推荐.

将商户位置数据与某商品 A 的订单交易数据在地图上投影, 得到关于商品 A 的交易情况热力图, 如图 1 所示, 以红色气泡标记表示商户位置, 色彩渲染区域的颜色深浅表示商品 A 的交易分布情况, 其中颜色越深表示交易越多, 越浅表示越少.



图 1 某商品交易热力图示例

对商圈这一点、线、面三部分构成的地理区域^[12], 本文采用圆形过滤与 DBSCAN 密度聚类算法、核密度估计算法相结合的方法进行范围划定, 首先以预处理下的浙江省商圈数据中的经纬度数据为圆心, 以 3 km 为半径, 划定该圆的外切正方形为商圈范围, 搜索出初步商户集合, 划定商圈搜索范围. 并获得最小经度, 最大经度, 最小纬度, 最大纬度的四个顶点坐标, 用以判断商户位置与商圈的关系. 然后进行离圆心点距离大于 3 km 的商户的过滤, 得到圆形过滤下的商户集合, 即初始商圈.

根据订单中的商户位置数据和交易数据, 结合圆形过滤方式下的初始商圈, 并通过改进的 DBSCAN 密度聚类算法对商户位置数据集中每个商户点的 Eps 邻域进行簇搜索, 即利用核密度估计方法来确定每个初始商圈内的 Eps 和 $MinPts$ 参数, 导出最大密度相连的商户集合, 实现对圆形过滤方式下的商圈商户进行进

一步去噪, 保证同一商圈下商户的相似性. 对划分商圈过程中浙江省全域范围下 25 万商户中未被包含在任何所划定的商圈内的近 8 千户商户, 以距离该商户最近商圈的推荐列表进行推荐. 建立起商圈与商户间的商圈位置区域模型.

2.2 面向流行度的推荐算法

基于商圈位置区域模型, 结合商圈流行度, 以商圈为关键字进行对商户商品交易数据的聚类, 考虑到时间因素对商品流行度的影响, 采用以交易时间计算权值的方法来统计衡量每个商品在各商圈内的流行度, 距离现在越近的交易所统计出的权重越高, 越远的权重越低. 具体计算每次交易权值的公式如下:

$$weight(k) = \frac{1}{1 + \alpha(T - t_k)} \quad (1)$$

上述公式中的 α 为时间衰减参数, 用来调控时间参数对商户兴趣变化的影响; k 指的是某商品的一次交易记录; t_k 表示该交易记录的完成时间; T 为当前时间.

针对商圈流行度, 按商品对时间加权后的所有交易进行统计, 得到各商品的总权重, 即为当前商圈内各个商品的最终流行度, 按流行度排序计入各商圈的推荐列表. 具体计算商圈内某商品流行度的公式如下:

$$pop(i) = \sum_{i \in u, k \in i} weight(k) \quad (2)$$

其中, u 表示某个商圈; i 表示在商圈 u 内有交易记录的一个商品; k 表示商品 i 的一次交易记录.

通过上述时间衰减参数与商品权重的排序计量, 根据商品在商圈内部的权重随时间变化的趋势有效反馈商圈商品流行度变化, 建立起面向流行度的商品推荐算法 (Business-circle Popularity Based Algorithm, BPA).

2.3 面向商圈相似度的推荐算法

在面向商圈流行度的推荐算法中, 采用的是将流行商品在商圈内进行推荐的方法, 容易造成长尾效应, 使流行商品一直在推荐列表中, 而新商品难以得到推荐, 导致推荐的商品存在新颖性和惊喜度不足的问题. 为了解决这一困境, 考虑商圈相似度因素, 通过计算不同商圈之间的相似度, 为目标商圈即当前被推荐的商圈推荐相似商圈中较为流行的商品, 并过滤掉目标商圈中已经交易过的商品, 建立起面向商圈相似度的协同过滤算法 (Business-circle Similarity Based Collaborative Filtering, BSCF), 主要进行如下操作:

统计筛选每个商圈内的商品交易数据,考虑时间上的流行度,计算各商圈内商户同时对一些商品感兴趣的程度,即先通过商品交易记录结合其流行度来计算商圈内各商品之间的相似度.再由所交易过的商品的相似度来计算商圈之间的相似度,从而将被预测商品与商圈中共同交易过的商品之间的相似性作为权重,代入计算商圈相似度的公式中:

$$sim(u, v) = \frac{\sum_{j \in N(u) \cap N(v)} sim(i, j)}{\sqrt{|N(u)| |N(v)|}} \quad (3)$$

其中, u 为目标商圈, i 为被预测的商品, $sim(u, v)$ 表示商圈 u 和商圈 v 之间的相似性, $N(u)$ 和 $N(v)$ 分别表示商圈 u 和商圈 v 中有过交易记录的商品集合. 其中 $sim(i, j)$ 表示商品 i 与在商圈 u 和商圈 v 中均有交易记录的商品 j 之间的相似性, 计算公式如下:

$$sim(i, j) = \frac{\sum_{x \in N(i) \cap N(j)} |pop(i) - pop(j)|}{\sqrt{|N(i)| |N(j)|}} \quad (4)$$

其中, x 表示某个与商品 i 和商品 j 均有过交易的商圈, $N(i)$ 和 $N(j)$ 分别表示与商品 i 和商品 j 的有过交易的商圈集合, $pop(i)$ 和 $pop(j)$ 分别表示商圈 x 中商品 i 和商品 j 的流行度.

3 实验结果及分析

3.1 实验测评标准

对于推荐算法性能,一般采用准确率 P 、召回率 R 进行评价,但本文在实验过程中发现,准确率和召回率并不是完全正线性相关的,二者互有高低且表现在

不同方面,因此本文还将采用 Pazzani 提出的调和平均数 F 值 (F -measure)^[13].

其中准确率 P 是指推荐正确的商品数目在总的推荐数目中的占比^[12],对于以 $RecommendSet$ 表示推荐集合, $PurchaseSet$ 表示交易商品种类数目集合,准确率 P 的计算公式如下所示:

$$P = \frac{1}{777} \sum_{s=1}^{777} \frac{|RecommendSet \cap PurchaseSet|}{|RecommendSet|} \quad (5)$$

召回率 R 是指推荐正确数目占总的购买种类数目的百分比^[12],计算公式如下:

$$R = \frac{1}{777} \sum_{s=1}^{777} \frac{|RecommendSet \cap PurchaseSet|}{|PurchaseSet|} \quad (6)$$

结合准确率和召回率的 F 值,能进行更全面的评价,计算公式如下:

$$F - measure = \frac{2 \times P \times R}{P + R} \quad (7)$$

3.2 结果分析与对比

3.2.1 时间衰减参数调整

时间衰减参数 α 在交易权重 $weight$ 的计算中起关键作用,并直接影响商品在各商圈内的流行度,但由于 α 没有特定取值范围,不同取值在不同数据集、系统中都会产生不一样的推荐效果.因此,实际应用中需通过反复试验来得到最优值.通过对面向商圈流行度的推荐算法进行实验,设定推荐数量 $N=15$, α 在 $[0, 0.09]$ 之间,具体结果如表 2 所示.其中, $\alpha=0$ 时,算法将不考虑时间因素的影响.

表 2 推荐数量 $N=15$ 时,不同 α 下的准确率和召回率

α	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
准确率 P/%	39.72	50.32	51.98	52.13	51.55	50.87	50.12	49.74	49.18	48.88
召回率 R/%	12.37	15.04	15.42	15.62	15.49	15.24	15.18	15.03	15.16	15.09

由上表可知,面向商圈流行度的推荐算法在结合时间衰减参数的前提下会有更好的推荐准确率和召回率.同时,当 $\alpha=0.03$ 时,效果最佳.因此,本文将选取时间衰减参数 α 为 0.03 来进行后续的对比实验.

3.2.2 实验对比算法

在对比实验中,选择 3 种传统推荐算法与本文的两种基于商圈位置区域模型的推荐算法来进行性能上的比较.对比算法如下: 1) 基于用户的协同过滤算法^[14] (User Based Collaborative Filtering, UCF), 根据商户的

商品交易次数矩阵来计算商户之间兴趣相似度,找出目标商户的最邻近商户集合进行推荐,并通过单值分解和聚类实现矩阵的降维,来缓解稀疏性问题; 2) 基于关联规则的推荐算法^[15] (Association Rules Based Algorithm, ARA), 通过改进的 Apriori 算法,来挖掘出数据之间所隐藏的潜在关系,为商户推荐与其交易记录关联程度较高的商品; 3) 基于热门商品的推荐算法 (Popular Products Based Algorithm, PPA), 将全域范围下商户交易数量最多的商品作为热门商品进行推荐.

3.2.3 算法稳定性分析

考虑不同的推荐数量对算法推荐质量的影响,进行算法稳定性的研究. 本文将推荐数 N 依次设定为 5, 10, 15, 20, 25 进行实验, 发现随着 N 的增加, 各算法的准确率呈现出先升高后下降趋势, 召回率则一直保持上升, 且在后期趋于平缓, 而 F 值变化曲线表明, 当 $N=15$ 时, 各算法的准确率和召回率之间均开始出现失衡. 因此, 本实验将选定 $N=15$ 作为平衡准确率和召回率的临界值. 具体结果如图 2-图 4 所示.

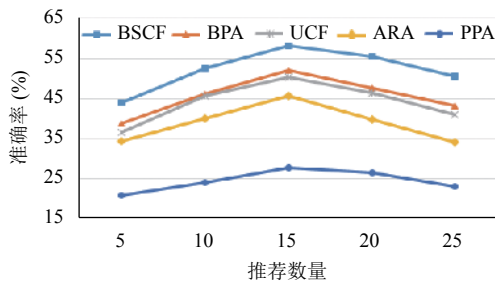


图2 准确率随推荐数量变化

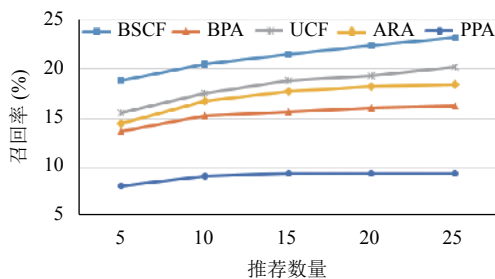


图3 召回率随推荐数量变化

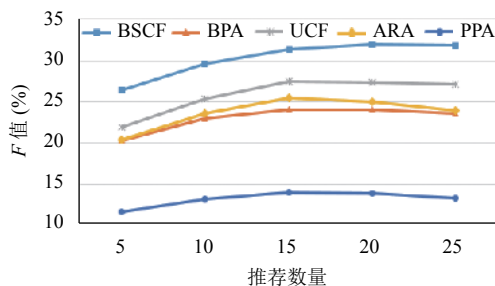


图4 F值随推荐数量变化

由实验可知, 本文所提出的面向商圈流行度的推荐算法具有较好的推荐准确性, 但由于将流行商品进行推荐, 存在长尾效应, 即其召回率和 F 值均较低. 而面向商圈相似度的协同过滤算法充分利用商户位置数据, 结合商圈, 在一定程度上增加了可用的推荐信息, 使推荐准确率和召回率明显优于其他几种算法, 具有

良好的稳定性. 并且该算法还缓解了冷启动问题, 针对可用信息较少, 兴趣特征难以提取的商户也可提供较为准确的推荐, 对比其他推荐算法, 面向商圈相似度的协同过滤算法具有明显的推荐优势.

3.2.4 算法应用与分析

企业 3 月份针对某款新品在浙江省内进行为期 6 个月的投放试验, 并在基于商圈位置区域模型的基础上采用基于商圈相似度的协同过滤算法来改进传统的新品投放策略, 以增强该规格新品与目标群体、零售商户所在商圈的消费群体匹配程度, 从而提高新品在消费者中的关注度与接受度. 具体投放策略流程如下:

- 1) 按商品的规格定位搜寻出与投放新品最为相似的商品;
- 2) 根据相似商品, 找出 5 个具有该新品最高流行度的商圈, 记为 A 组商圈;
- 3) 通过商圈相似度, 搜寻出 10 个与 A 组商圈最为相似的商圈, 记为 B 组商圈;
- 4) 为验证 A, B 两组商圈针对所投放的新品具有一定的投放意义, 通过传统投放策略, 即通过销量、结构或者档位进行排序, 择优选取 50 名零售户, 进行对比实验, 记为 C 组;
- 5) 分别从 A, B 两组商圈中随机选择 50 个零售户作为投放户, 与 C 组一起进行新品的等量投放试验, 统计各组商圈在每个月份针对所投放的新品的销量, 绘制销量趋势图, 如图 5 所示.

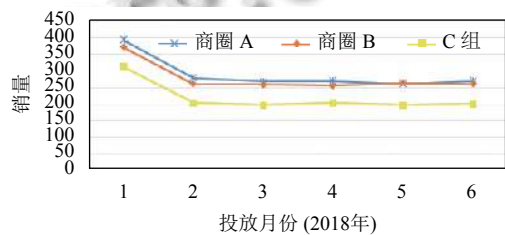


图5 新品投放效益对比图

可以发现, 经过 6 个月的新品投放试验, A, B 两组商圈每个月的新品销量明显高于 C 组, 因此, 该新品投放方法具有较好的投放效果. 并且, 该基于商圈相似度的新品投放策略已在今年 8 月在全国范围内正式部署, 具有较好的投放效益.

4 结论

本文通过建立商圈位置区域模型, 将地理位置信息作为重要特征数据纳入推荐算法, 有效弥补了传统

推荐算法中地理位置或区域信息因素缺失的缺陷。相较于其他基于模型的推荐算法,引入商圈作为市场划分的标准,并进行分区推荐,增强了市场拟合度,使得推荐更具有市场针对性和可靠性。对于所提出的两种基于商圈位置区域模型的推荐算法:面向商圈流行度的推荐算法和面向商圈相似度的协同过滤算法,本文在实验过程中采用 *Spark* 进行算法计算效率加速^[12,16,17]。实验结果表明,基于商圈位置区域模型的推荐算法相较于传统推荐算法,具有较好的推荐准确率,可以一定程度地缓解冷启动问题。而面向商圈相似度的协同过滤算法通过推荐相似商圈中的流行商品,很好地提高了推荐结果的新颖性和惊喜度,避免了长尾效应的产生,在保证推荐准确率的前提下,具有良好的稳定性。

下一阶段,将继续对基于商圈位置区域模型的推荐算法进行研究,并从商圈范围划分、数据指标的设定、计算效率的提升等几方面改进算法及实验,提高商品的推荐准确率,满足用户个性化商品推荐的需求。

参考文献

- 1 邱爽,葛万成,汪亮友,等.个性化推荐中基于用户协同过滤算法的优化.信息技术,2016,(3):62-63,72. [doi: 10.3969/j.issn.1674-2117.2016.03.025]
- 2 胡新明,罗建军,夏火松.基于商品领域知识的交互式推荐系统.现代图书情报技术,2014,(10):56-62. [doi: 10.11925/infotech.1003-3513.2014.10.09]
- 3 江周峰,杨俊,鄂海红.结合社会化标签的基于内容的推荐算法.软件,2015,36(1):1-5. [doi: 10.3969/j.issn.1003-6970.2015.01.001]
- 4 张志威.个性化推荐算法研究综述.信息与电脑(理论版),2018,(17):27-29.
- 5 孟晓龙.基于机器学习的推荐技术研究[硕士学位论文].长春:吉林大学,2016.
- 6 王聪.基于深度学习的推荐系统均模型研究[硕士学位论文].太原:太原理工大学,2018.
- 7 冯永,张备,强保华,等.MN-HDRM:长短兴趣多神经网络混合动态推荐模型.计算机学报,2019,42(1):16-28.
- 8 郝斌,董硕,胡引翠,等.多维特征融合的城市商圈划分方法.地理与地理信息科学,2017,33(5):56-62. [doi: 10.3969/j.issn.1672-0504.2017.05.009]
- 9 秦佳睿,徐蔚鸿,马红华,等.自适应局部半径的DBSCAN聚类算法.小型微型计算机系统,2018,39(10):2186-2190. [doi: 10.3969/j.issn.1000-1220.2018.10.010]
- 10 董晓君,程春玲.基于核密度估计的K-CFSFDP聚类算法.计算机科学,2018,45(11):244-248. [doi: 10.11896/j.issn.1002-137X.2018.11.038]
- 11 翁璇.基于商圈理论的商业建筑设计策略[硕士学位论文].广州:华南理工大学,2012.
- 12 吴庭.基于位置的推荐计算:Spark实现[硕士学位论文].杭州:浙江大学,2016.
- 13 Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 1997, 27(3): 313-331. [doi: 10.1023/A:1007369909943]
- 14 吴颜,沈洁,顾天竺,等.协同过滤推荐系统中数据稀疏问题的解决.计算机应用研究,2007,24(6):94-97. [doi: 10.3969/j.issn.1001-3695.2007.06.028]
- 15 王立华,肖慧,徐硕,等.基于关联规则的渔业信息推荐系统设计与实现.农业工程学报,2013,29(7):124-130.
- 16 陈恩杰.基于Spark的实时推荐系统的研究与设计[硕士学位论文].西安:西安科技大学,2018.
- 17 岑凯伦,于红岩,杨腾霄.大数据下基于Spark的电商实时推荐系统的设计与实现.现代计算机,2016,(24):61-69. [doi: 10.3969/j.issn.1007-1423.2016.24.015]