基于数据挖掘的高维数据协同过滤推荐算法*

朱木清[1],文　谧[2]
(1.广州华立学院,广州,511325;
2.广州应用科技学院,广州,511370)

摘　要:考虑到传统算法在推荐高维数据时存在覆盖率和准确率低、平均绝对误差和均方根误差大的问题,提出了基于数据挖掘的高维数据协同过滤推荐算法研究。根据高维数据属性特征的偏好值,预测评分高维数据属性特征,采用关联规则对分解后的属性特征重构,得到高维数据属性特征的数据挖掘分类树,提取出高维数据属性特征,利用空间向量法,判断高维数据信息间的权重信息值,根据权重信息设置相应的门限值,得到高维数据信息间相似度的分布情况,完成对高维数据间的相似度值的计算,利用数据挖掘技术,对高维数据信息预处理,结合高维数据协同过滤推荐算法设计,实现了高维数据的协同过滤推荐。实验结果表明,基于数据挖掘的推荐算法不仅可以通过提高覆盖率和准确率增强推荐效果,还可以通过降低平均绝对误差和均方根误差提高推荐性能。

关键词:数据挖掘;协同过滤;属性特征;高维数据;相似度计算;推荐算法
中图分类号:TP181　　**文献标识码**:A　　DOI 编码:10.14016/j.cnki.1001-9227.2022.01.091

High dimensional data collaborative filtering recommendation algorithm based on Data Mining

ZHU Muqing[1], WEN Mi[2]
(*1.Guangzhou Huali College, Guangzhou 511325, China*;
2.Guangzhou College of Applied Science and Technology, Guangzhou 511370, China)

Abstract: Considering the problems of low coverage and accuracy, large mean absolute error and root mean square error when traditional algorithms recommend high-dimensional data, a research on high-dimensional data collaborative filtering recommendation algorithm based on data mining is proposed. According to the preference value of high-level data attribute features, predict and score the high-dimensional data attribute features, reconstruct the decomposed attribute features by using association rules, obtain the data mining classification tree of high-dimensional data attribute features, extract the high-dimensional data attribute features, judge the weight information value between high-dimensional data information by using space vector method, and set the corresponding threshold value according to the weight information, Get the distribution of similarity between high-dimensional data information, complete the calculation of similarity value between high-dimensional data, use data mining technology to preprocess high-dimensional data information, combined with the design of high-dimensional data collaborative filtering recommendation algorithm, and realize the collaborative filtering recommendation of high-dimensional data. The experimental results show that the recommendation algorithm based on data mining can not only enhance the recommendation effect by improving the coverage and accuracy, but also improve the recommendation performance by reducing the average absolute error and root mean square error.

Key words: data mining; collaborative filtering; attribute characteristics; high dimensional data; similarity calculation; recommendation algorithm

0　引言

随着互联网时代的到来,线上产品的信息量不断地增多,导致网络信息数据也不断变多,面对海量的高维数据信息,如何对其负载信息有效筛选,成为人们关注的问题[1]。目前,面对激烈竞争的互联网市场,如何为用户提供更好的服务,已经成为供应商企业考虑的首要问题。现阶段,面对海量的高维数据信息,传统的协同过滤算法已经不能满足当今社会的需求,为了更好地服务于用户,利用推荐算法,为用户提供适用于个人偏好的个性化服务,已经成为目前电子商务领域最受关注的方式[2]。随着互联网的飞速发展和电子商务信息量的不断增多,传统的推荐算法,如基于内容的推荐算法等呈现出准确率低等问题,以保护用户隐私为出发点,对

收稿日期:2021-09-18
*基金项目:广东高校省级重点平台和重大科研项目(青年创新人才类)"基于学科知识库的问答系统研究及应用"(No.2016KQNCX212)
作者简介:朱木清(1982-),男,汉族,广东茂名,本科,讲师,研究方向为智能信息处理与知识工程研究。

高维数据协同过滤推荐算法进行研究[3]。

高海燕等人[4]提出了一种基于低维特征提取的协同过滤推荐算法,首先构建低维信任模型,利用模型技术对用户的低维数据进行特征提取,并根据提取的多维特征推算出各个特征之间的关联关系,利用信任模型表示其关联关系,最后根据用户的信任度对产品的权重进行评分,提高协同过滤推荐算法对数据稀疏问题的数据挖掘准确率,并通过多组数据进行验证,结果表明,基于低维特征提取的协同过滤推荐算法可以解决传统协同过滤算法不能解决的数据稀疏问题,但是推荐效果较差,只能针对用户搜索过的相关商品进行推荐,无分析功能。任永功等人[5]提出了一种基于决策规则的协同过滤推荐算法,用来解决传统的协同过滤推荐算法不能准确计算出用户喜好的问题。首先,采集用户日常购物的商品属性,并构建喜好偏爱矩阵,根据决策规则对矩阵中的每一条规则进行约束,完成对用户日常购物的属性特征的确定,并根据决策对象对属性特征进行约简,完成对用户喜好的预测。实验结果表明,该方法可以提高推荐的准确率,但是其稳定性不足,容易为客户推荐不喜欢的风格及产品。

基于以上研究,将数据挖掘应用到了高维数据协同过滤推荐算法设计中,从而提高高维数据推荐的效果和性能。

1　高维数据协同过滤推荐算法设计

1.1　基于数据挖掘提取高维数据属性特征

基于数据挖掘对高维数据属性特征进行提取,首先对高维数据进行采集,并通过总线进行传输,构建高维数据属性架构图,通过数据挖掘算法[6],对高维数据的属性特征进行偏好修正,并利用自适应算法实现对高维数据属性特征的提取,利用下式表示高维数据属性特征的偏好值:

$$C_{uu}^{*}=\sqrt{\frac{d_{in}(v_1)}{d_{out}(u_1)+d_{in}(v_1)}} \tag{1}$$

其中,$d_{out}(u_1)$是高维数据属性特征值,u_1为采集的高维数据属性特征的关联信息集合,$d_{in}(v_1)$为高维数据特征挖掘过程中的偏好值,v_1表示高维数据属性特征的节点集合。

高维数据属性特征集合u_1的信息属性特征偏好会受噪音N_u的影响[7],如果基于数据挖掘规则u的高维数据属性特征表示为v,将挖掘的高维数据属性特征的偏好特征C_{uu}^{*}汇入进去,通过下式对这种关系进行表示:

$$\bar{R}_{ik}=\sum_{u_1\in v}C_{uu}^{*}R_{jk} \tag{2}$$

其中,$\bar{R}_{ik}$表示挖掘结果u_i对高维数据v_j的偏好预测,R_{jk}表示挖掘结果u_i对高维数据属性特征v_k的偏好评分,通过下式完成对高维数据属性特征的预测评分:

$$\bar{R}=T\bar{R}_{ik} \tag{3}$$

上式中,T表示高维数据属性特征的权重信息。

把上式内得到的预测评分代入数据挖掘算法,并对高维数据属性特征进行分解,利用关联规则对分解后的属性特征进行重构,从而得到高维数据属性特征向量[8]。依靠得到的高维数据特征向量,得到高维数据属性特征的数据挖掘分类树:

$$p(R\mid\sigma_R^2)=\prod_{u_j=1}^{r}[N(R_{ij}\mid\mu,\sigma_R^2)\,g(x)I_{ij}^{R}(U_i^{T}V_\lambda)] \tag{4}$$

其中,$N(R_{ij}\mid\mu,\sigma_R^2)$表示利用数据挖掘技术获取到的高维数据属性特征$r$中的一个属性特征,高维数据属性特征的标准方差为$\sigma_R^2$,$I_{ij}^{R}$表示特征提取函数,假如高维数据信息$U_i^{T}$的特征属性为$V_\lambda$,得到其偏好特征。

通过对高维数据属性特征的挖掘,完成对高维数据属性特征的提取。

1.2　计算高维数据间的相似度

在对高维数据间的相似度进行求解过程中,主要利用空间向量法,判断高维数据信息间的权重信息值[9],根据权重信息设置相应的门限值,防止高维数据信息的权重门限值过高,导致相似度较低。高维数据信息的权重计算公式如下:

$$Weight\ (t)_d=[\alpha(t)_d+\beta(t)_d]\times TFIDF\ (t) \tag{5}$$

式中,$Weight\ (t)_d$表示高维数据d中的某一个数据信息t的权重,数据信息t在数据集d中的初始权重和结束权重分别为$\alpha(t)_d$和$\beta(t)_d$,对于数据信息t在数据集d中的TFIDF值,可以通过式(6)计算,即:

$$TFIDF\ (t)_d=\frac{TF(t)_d\times\log10\left(\frac{N_k}{n_i}+0.01\right)}{\sqrt{\sum_{xd}\left[TF(t)_d\times\log10\left(\frac{N_k}{n_i}+0.01\right)\right]^2}} \tag{6}$$

上式中,N_k表示高维数据信息集中的特征量,$TF(t)_d$表示高维数据t在数据集d中的存在概率,n_i表示数据集d中所有的高维数据数量。

通过自适应分析法[10],对高维数据间的相似度进行定义为:

$$\mathrm{Sim}\ (s_i,s_j)=\frac{|C_i\cap C_j|}{|C_i\cup C_j|} \tag{7}$$

上式中,C表示高维数据间的特征信息集合,通过对两个高维数据特征向量s_i与s_j的求解,可以获取关于高维数据特征向量的对比值,从而确定高维数据间的相似度。通过公式(7)可以确定高维数据间的相似度在0~1之间取值,其中0表示完全不存在相似的两个高维数据信息,1表示完全相似的高维数据信息。

通过对高维数据的相似度进行计算,得到数据集中不同高维数据相似度的分布情况[11],采用以下矩阵来表示:

$$S=\begin{bmatrix}s_{11}&s_{12}&\cdots&s_{1n}\\s_{21}&s_{22}&\cdots&s_{2n}\\\vdots&\vdots&\ddots&\vdots\\s_{n1}&s_{n2}&\cdots&s_{nm}\end{bmatrix} \tag{8}$$

式中,s_{nm}表示数据集中不同高维数据s_n与s_m二者间的相似度。

通过对数据集中不同高维数据相似度的分析,设置好高维数据相似度权重门限值,通过对权重值的计算,获得数据集中高维数据间的相似度,得到高维数据信息间相似度的分布情况,完成对高维数据间的相似度值的计算。

1.3 设计高维数据协同过滤推荐算法

利用数据挖掘技术,对高维数据信息进行预处理,实现对信息的分类[12],构建协同过滤评分模型,用 $m\times n$ 协同过滤评分矩阵表示,如下:

$$R(m,n)=\begin{Bmatrix} r_1 & r_2 & \cdots & r_n \\ r_{11} & r_{12} & \cdots & r_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{Bmatrix} \tag{9}$$

上式中,m 代表对高维数据进行协同过滤过程中评分矩阵的行数,n 为协同过滤过程中评分矩阵的列数,高维数据 i 对目标数据 j 的评分结果表示为 R_{ij}。对高维数据的评价等级用 0~5 的常数表示,将其表示对高维数据协同过滤的评分,该常数值越大说明对推荐结果的评分越高,0 则表示没有对推荐结果进行评价[13]。

根据高维数据的协同过滤评分模型,对用户的喜好评分预测,对采集的每个高维数据样本进行协同过滤得到 $T=\{(q_1,u_1),\cdots(q_n,u_n)\}$,其中,$q_i=(q_{1i},\cdots,q_{ni})$,第 i 个协同过滤训练样本的第 j 种高维数据属性表示为 q_{ji}。根据用户的评价体系,利用式(10)和式(11)计算出用户对高维数据 $P(c_k)$ 进行协同过滤的条件概率:

$$P(c_k)=\sum_{i=1}^{N}\frac{I(c_k)}{N} \tag{10}$$

$$P(q_i^j \mid c_k)=\frac{\sum_{i=1}^{N} I(q_i^j,c_k)}{\sum_{i=1}^{N} I(c_k)} \tag{11}$$

其中,N 表示用户点击喜好的频率,$I(c_k)$ 表示用户搜索结果与评价之间的概率,$I(q_i^j,c_k)$ 表示用户搜索的条件概率,$\sum_{i=1}^{N} I(c_k)$ 表示高维数据协同过滤训练样本与用户偏好的相似概率。对模型进行训练后,即可对高维数据 q^j 进行协同过滤推断:

$$U=P(c_k)\prod_{j=1}^{n}P(q_i^j \mid c_k) \tag{12}$$

上式中,根据高维数据的特征不同会得到不同的用户喜好值,该值表示用户喜好度与推荐结果之间的关系[14]。

在高维数据协同过滤推荐过程中,假设 $U=\{(u_1,a_1),\cdots,(u_n,a_n)\}$ 表示数据节点信息,n 个推荐结果和 m 个用户搜索高维数据的二分图为 G,a_i 为高维数据协同过滤结果的预测值,对用户喜好协同过滤 a_i 求和,即:

$$w_{qi}=\sum_{j=1}^{n}\frac{a_j}{k(u_j)} \tag{13}$$

其中,$k(u_j)$ 表示高维数据协同过滤的数量,a_j 表示对协同过滤结果的预测值。根据用户的偏好对推荐列表进行预测,即:

$$w_{qi}=\sum_{j=1}^{n}\frac{a_j}{k(u_j)}\frac{\text{click}\ (u_j,q_i)}{N} \tag{14}$$

其中,$click\ (u_j,q_i)$ 表示高维数据 q_i 被推荐的概率。

在最终的高维数据推荐列表中,存在:

$$v_{qi}=\left[\sum_{j=1}^{n}\frac{\text{click}\ (u_i,q_i)}{Num(q_i)}+\alpha\frac{|q\cap q_i|}{|q\cup q_i|}\right] \tag{15}$$

其中,α 表示时间影响因素,根据不同用户的喜好系数设置不同的时间因子[15]。通过对用户喜好 q_i 协同过滤,将相似度最高的高维数据推荐给用户:

$$N=\sum_{j=1}^{n}t(q,q_i) \tag{16}$$

其中,$t(q,q_i)$ 表示 q 与 q_i 的关联性。

根据以上过程,设计了高维数据协同过滤推荐算法,实现了高维数据的协同过滤推荐。

2 实验分析

2.1 搭建实验数据集

协同过滤推荐实验数据来自某一高维数据库,包含 200 名用户和 500 组高维数据,数据集中包括高维数据类型、维数以及需求量。将实验数据分为训练集和测试集,先将用户推荐的高维数据设置成 10 条,利用协同过滤推荐算法来测试高维数据,最后将推荐实验结果展示出来。

2.2 设置实验指标

高维数据协同过滤推荐实验分两个阶段进行,先利用覆盖率和准确率指标衡量高维数据的推荐效果,计算公式为:

$$Q=\frac{N_d}{N_r} \tag{17}$$

$$p=\frac{N}{M} \tag{18}$$

其中,Q 表示推荐覆盖率,N_d 表示用于推荐的高维数据量,N_r 表示高维数据总量,p 表示推荐准确率,N 表示推荐中用户感兴趣的高维数据量,M 表示向用户推荐的高维数据总量。

在实验第二阶段,利用平均绝对误差和均方根误差衡量高维数据的推荐性能,计算公式为:

$$MAE=\frac{1}{n}\sum_{i,j}|r_{ij}-x_{ij}| \tag{19}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i,j}(r_{ij}-x_{ij})^2} \tag{20}$$

上式中,n 表示实验测试次数,r_{ij} 表示推荐结果,x_{ij} 表示符合用户需求的推荐结果。

2.3 结果分析

为了更加突出基于数据挖掘的推荐算法的优势,引入基于图嵌入模型的推荐算法和基于检索内容提取的推荐算法作对比,测试了三种推荐算法的效果和性能,结果如下。

三种算法的高维数据协同过滤推荐覆盖率测试结果如图 1 所示。

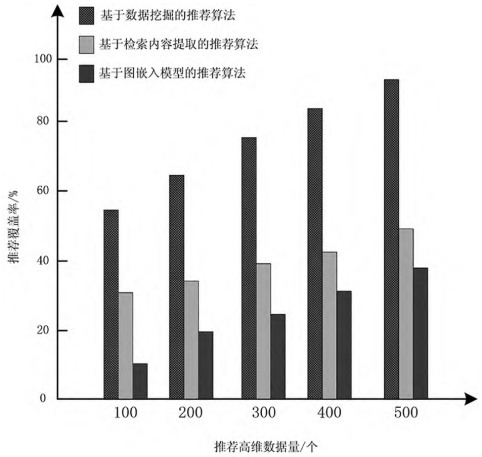

图 1　高维数据协同过滤推荐覆盖率测试结果

从图 1 的结果可以看出,基于图嵌入模型的推荐算法和基于检索内容提取的推荐算法得到的高维数据协同过滤推荐覆盖率测试结果比较接近,协同过滤推荐覆盖率都低于 50%,最高的覆盖率也只有 47%和 38%,而采用基于数据挖掘的推荐算法时,最低的覆盖率都达到了 55%,随着推荐的高维数据量增加,协同过滤推荐覆盖率越来越大,最高达到了 93%,因此,说明文中方法通过提高推荐覆盖率,提高高维数据的推荐效果。

三种算法的高维数据协同过滤推荐准确率测试结果如图 2 所示。

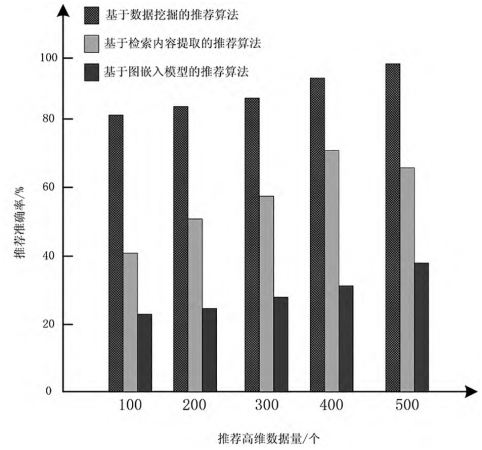

图 2　高维数据协同过滤推荐准确率测试结果

从图 2 的结果可以看出,采用基于图嵌入模型的推荐算法时,推荐准确率在 20%~40%之间,采用基于检索内容提取的推荐算法时,推荐准确率在 40%~70%之间,基本可以满足用户的兴趣需求,采用基于数据挖掘的推荐算法时,推荐准确率都高于 80%,甚至最高准确率达到了 99%,说明文中方法在推荐高维数据时的效果完全满足用户的兴趣需求。

三种算法在高维数据协同过滤推荐平均绝对误差方面的测试结果如图 3 所示。

从图 3 的结果可以看出,在高维数据推荐的平均绝对误差测试中,与基于图嵌入模型的推荐算法和基于检索内容提取的推荐算法相比,基于数据挖掘的推荐算法具有更好的推荐性能,平均绝对误差值在 0.25 以内,说明文中算法更适用于高维数据的协同过滤推荐。

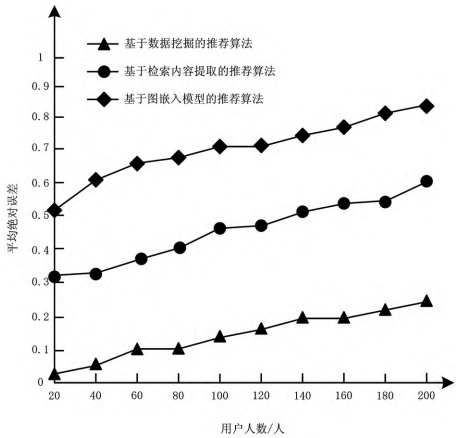

图 3　平均绝对误差测试结果

三种算法在高维数据协同过滤推荐均方根误差方面的测试结果如图 4 所示。

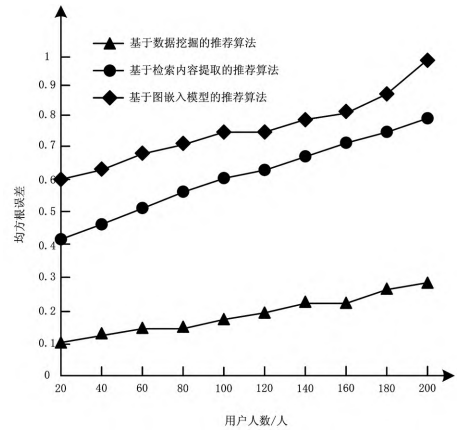

图 4　均方根误差测试结果

从图 4 的结果可以看出,文中算法在推荐高维数据时的均方根误差在 0.1~0.3 之间,采用基于图嵌入模型的推荐算法和基于检索内容提取的推荐算法时,推荐高维数据的最小均方根误差分别为 0.6 和 0.4,随着用户人数的增加,高维数据推荐的均方根误差逐渐变大,当用户人数达到 200 人时,数据推荐的均方根误差达到了 0.79 和 0.95,说明文中方法在推荐的均方根误差中具有更好的性能。

3　结束语

提出了基于数据挖掘的高维数据协同过滤推荐算法研究,经实验测试发现,该推荐算法在高维数据中的应用具有更好的效果和性能。但是本研究仍然存在很多不足,在今后的研究中,希望可以引入蚁群算法搜索出一条最优的推荐路径,从而提高高维数据的推荐效率。

参考文献

[1]　吴宾, 娄铮铮, 叶阳东. 一种面向多源异构数据的协同过滤推荐算法[J]. 计算机研究与发展, 2019, 56(005):1034-1047.

[2]　王涵,夏鸿斌.LDA 模型和列表排序混合的协同过滤推荐算法[J].计算机科学,2019,46(09):216-222.

(下转第 99 页)

时监控系统,极大地缩短了系统响应时间,提升了监控有效率,为自动售货机的发展与应用提供有力的系统支撑,也为实时监控研究提供一定的参考。

参考文献

[1] 张绘敏. 基于 PLC 的蔬菜自动售货机控制系统设计[J]. 广西农业机械化, 2019, 219(5):39-40.

[2] 金薇. 基于 PLC 的自动售货机控制系统设计[J]. 中国新技术新产品, 2020, 420(14):20-21.

[3] 王晓丽. 一种基于 PLC 的自动售货机系统的设计[J]. 集成电路应用, 2019, 36(2):74-75.

[4] 关斯斯, 于帆. 基于眼动追踪的自动售货机人机界面设计研究[J]. 包装工程, 2019, 40(8):242-248.

[5] 孙娜, 潘振华, 于金秀. 基于灰色预测模型的自动售货机商品销售量研究[J]. 商场现代化, 2020, 911(2):12-13.

[6] 高一歌. 基于服务设计的智能自动售货系统人机交互研究[J]. 设计, 2019, 32(7):40-43.

[7] 王志鹏, 段浩, 卢郑兴. 基于 NB-IoT 的工业设备报警实时远程监控系统设计[J]. 电子设计工程, 2020, 28(5):67-71+76.

[8] 李琦. 基于 PLC 通信的料位实时监测系统的设计与应用[J]. 机械管理开发, 2019, 34(6):220-221.

[9] 史瑞刚, 周亮, 秦琴琴. 可移动支付的导航购物车系统设计与实现[J]. 自动化技术与应用, 2019, 38(12):82-86.

[10] 刘晴, 董平军. 时间序列长度对基于 Prophet 的自动售货机销量预测影响研究[J]. 管理科学与工程, 2020, 9(4):11-11.

[11] 祝婕. 新零售时代的线下连锁商户移动支付渠道战略转型分析[J]. 通讯世界, 2019, 26(3):65-66.

[12] 刘青苑. 双廓复合移动凸轮式售货机出罐机构设计[J]. 常州工学院学报, 2019, 32(4):17-20.

[13] 邢生. 京东支付"刷脸付"自助售货机 亮相世界互联网大会[J]. 中国周刊, 2019, 232(11):61-62.

[14] 崔棚飞. 基于单片机的自动售货机[J]. 中国新通信, 2019, 21(23):86-86.

[15] 宗临. 自动售货机的前世今生[J]. 现代班组, 2019, 154(10):19-19.

[16] 徐凤芹, 杨娟娟, 张文健. 自助售货终端的人机分析和改进设计[J]. 机械工程与自动化, 2019, 216(5):205-206+211.

(上接第 94 页)

[3] 陆航,师智斌,刘忠宝.融合用户兴趣和评分差异的协同过滤推荐算法[J].计算机工程与应用,2020,56(07):24-29.

[4] 高海燕,毛林,窦凯奇,等.基于图嵌入模型的协同过滤推荐算法[J].数据采集与处理,2020,35(03):483-493.

[5] 任永功,张云鹏,张志鹏.基于粗糙集规则提取的协同过滤推荐算法[J].通信学报,2020,41(01):76-83.

[6] 宋月亭, 吴晟. 基于相似度优化和流形学习的协同过滤算法改进研究[J]. 计算机工程与科学, 2020, 42(2):351-357.

[7] 王根生,潘方正.融合语义相似度的协同过滤推荐算法[J].中国科学技术大学学报,2019,49(10):835-841.

[8] 孔麟,黄俊,马浩,等.融合多层相似度与信任机制的协同过滤算法[J].计算机工程与设计,2020,41(12):3405-3411.

[9] 崔春生,王辉,李群.基于用户标签和信任关系的协同过滤推荐算法研究[J].系统科学与数学,2019,39(03):437-448.

[10] 张志鹏, 张尧, 任永功. 基于时间相关度和覆盖权重的协同过滤推荐算法[J]. 模式识别与人工智能, 2019, 032(004):289-297.

[11] 李悦,谢珺,侯文丽,等.融合用户偏好优化聚类的协同过滤推荐算法[J].郑州大学学报(理学版),2020,52(02):29-35.

[12] 陈碧毅, 黄玲, 王昌栋,等. 融合显式反馈与隐式反馈的协同过滤推荐算法[J]. 软件学报, 2020, 000(003):794-805.

[13] 李维乾,张艺,郑振峰,等.基于多属性的动态采样协同过滤推荐算法[J].计算机应用研究,2020,37(09):2640-2644+2683.

[14] 李昆仑,王萌萌,于志波,等.融合信任度值与半监督密度峰值聚类的改进协同过滤推荐算法[J].小型微型计算机系统,2020,41(08):1613-1619.

[15] 王卫红, 曾英杰. 基于聚类和用户偏好的协同过滤推荐算法[J]. 计算机工程与应用, 2020, 056(003):68-73.