东北农业科学 2020,45(6):140-144
Journal of Northeast Agricultural Sciences DOI: 10.16423/j.cnki.1003-8701.2020.06.032

基于大数据技术的农产品智能推荐方法研究

傅思维,陈桂芬*,赵 姗
(吉林农业大学信息技术学院,长春 130118)

摘 要:随着智慧农业与大数据智能的兴起,农产品电商平台智能推荐方法正成为高效满足个性化需求的重要手段。针对传统推荐方法存在的耗时长、效率低问题,本研究提出了基于大数据处理技术的农产品智能推荐方法。该方法首先将文档主题算法与矩阵分解算法混合,形成文档主题与矩阵分解混合算法;然后,将基于物品的协同过滤算法和文档主题与矩阵分解混合算法进行加权融合;最后,搭建Spark并行化计算平台,抓取京东商城和中国农产品网销售评分、评论等数据,进行特征提取、加权融合、智能推荐、误差测评。实验结果表明:文档主题与矩阵分解混合算法可有效提高推荐准确率;主题加权融合协同过滤算法可提高多样性;农产品智能推荐方法在推荐质量及执行效率方面具有明显提升。

关键词:混合算法;主题加权融合协同过滤算法;智能推荐;农产品;大数据处理技术
中图分类号:TP391.3 **文献标识码**:A **文章编号**:2096-5877(2020)06-0140-05

Research on Intelligent Recommendation Method of Agricultural Products Based on Big Data Technology

FU Siwei, CHEN Guifen*, ZHAO Shan
(*Information Technology of Jilin Agricultural University, Changchun* 130118, China)

Abstract: With the rising of intelligent agriculture and big data intelligence, the intelligent recommendation method of e-commerce platform for agricultural products has becoming an important measure to satisfy the personalized needs efficiently. Aiming at the problems of long time-consuming and low efficiency of traditional recommendation methods, this paper proposes an intelligent recommendation method of agricultural products based on big data processing technology. In this method, a kind of LDA-MF hybrid algorithm was formed by integrating the document theme algorithm and matrix factorization algorithm. Second, weighting the fusion of collaborative filtering algorithm based on the item and LDA-MF hybrid algorithm. Finally, a Spark parallel computing platform is built to capture the sales scoring and commentary data of jd.com mall and China Agricultural Products Network, then feature extraction, weighted fusion, intelligent recommendation and error evaluation will be carried out. Experimental results show that the LDA-MF hybrid algorithm can effectively improve the precision of recommendation, the theme weighted fusion collaborative filtering algorithm can improve diversity, and the intelligent recommendation method of agricultural products has been obviously improved in recommending quality and execution efficiency.

Key words: Hybrid algorithm; Theme weighted fusion collaborative filtering algorithm; Intelligent recommendation; Agricultural products; Big data processing technology

随着互联网、云计算、大数据技术的不断进步,面对农产品电商平台中随时更新的商品数据与海量用户信息,当前电商智能推荐研究的首要任务是如何快速准确地将用户感兴趣的商品推荐到用户客户端,使其产生兴趣并增加购买欲望。倘若能够在研究方法中结合用户评论文本等召回数据进行有效特征模型训练,将会达到加强深度学习的目的[1]。2017年,张敏等[2]提出将层叠降噪自动编码器与隐含因子模型相结合的混合推荐方法;黄璐等[3]提出了融合主题模型和协同过滤的多样化移动应用推荐;2018年,熊回香等[1]利用LDA 进行相似度计算提取资源内容关键词以构建标签混合推荐模型;王源龙等[4]结合Spark分布式计算平台提出一种混合推荐算法的增量迭代

收稿日期:2018-12-05
基金项目:国家星火计划项目(2015GA660004);吉林省重点科技研发项目(20180201073SF)
作者简介:傅思维(1992-),女,在读硕士,主要从事人工智能与智能计算研究。
通讯作者:陈桂芬,女,博士,教授,E-mail: guifchen@163.com

型;顾军华等[5]在Spark平台上创建一种基于图游走的计算间接相似度算法来实现文本方法的并行化,这些推荐方法在提高推荐准确率、多样性、新颖度方面做出很大贡献。本文提出了基于Spark分布式平台的文档主题与矩阵分解混合算法和协同过滤算法后融合的智能推荐方法。

1　文档主题与矩阵分解混合算法设计

1.1　文档主题与矩阵分解混合算法

文档主题算法可用于识别大型文档集合或语料库中的隐藏主题信息,主要用于文档主题建模,提高推荐的准确性,解决冷启动问题;矩阵分解算法可以有效地将高维用户-项目评分矩阵映射到低维近似矩阵中,解决数据稀疏性问题。根据已有学者的实践经验[3,6],提取物品主题及用户主题特点的文档主题算法与高效处理多维数据的矩阵分解算法融合,形成准确决策兴趣信息的混合算法。该混合算法框图如图1所示。

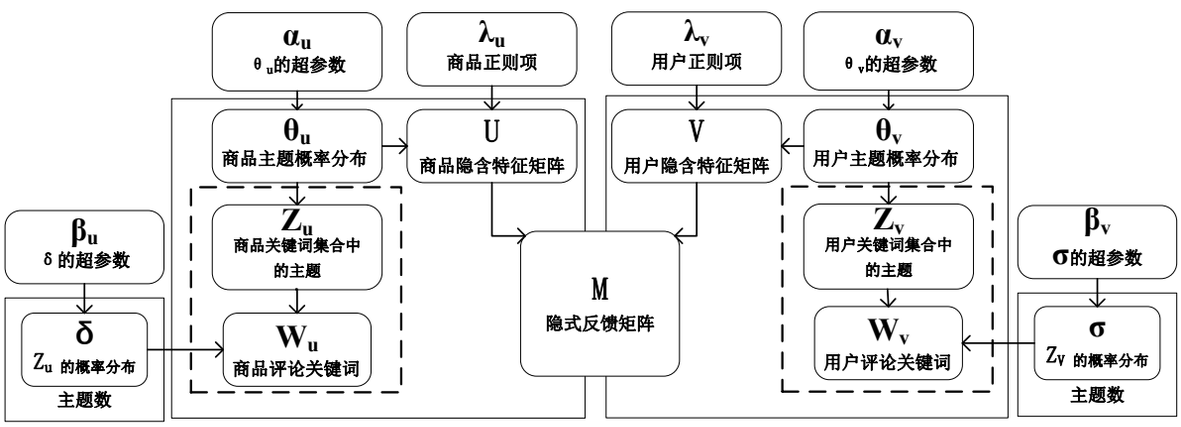

图1　文档主题与矩阵分解混合算法框图

在文档主题与矩阵分解混合算法框图中,通过商品u维度的评论关键词及评分概率分布,提取商品主题特征;商品主题特征由用户v维度的评论关键字及评分概率分布提取[7-8]。将商品主题概率分布θ_u与用户主题概率分布θ_v作为商品隐含特征矩阵U和用户隐含特征矩阵V的输入,使用ALS交替最小二乘法来求解,在商品和用户的特征分布中存在偏差值ε。计算用户的隐含特征V^*与商品的隐含特征U^*后,将融合结果的Top-80商品作为候选集,通过公式(1)计算用户x与商品y之间的兴趣关系r_{xy}^*。

$$r_{xy}^* = (v_x^*)^T u_y^* \quad \cdots\cdots(1)$$

1.2　文档主题与矩阵分解混合算法和文档主题算法性能对比

从京东商城大米购买评论数据集中随机选取80%评论超过30条的2013个用户作为训练集,并选取剩余的20%为测试集,同时每个用户在测试集中只有一条记录[9]。本对比实验选用准确率(precision)指标,首先定义X是数据集中用户集合,$hit(x_y)$表示根据训练集用户行为推荐给用户x_y的商品中,用户购买个数;$L(x_y)$表示测试集中推荐给用户x_y的商品列表长度。

$$precision = \frac{\sum_{x_y \in X} hit(x_y)}{\sum_{x_y \in X} L(x_y)} \quad \cdots\cdots(2)$$

将评论文档测试集应用文档主题与矩阵分解混合算法和文档主题算法进行准确率(precision)对比,结果如图2所示。

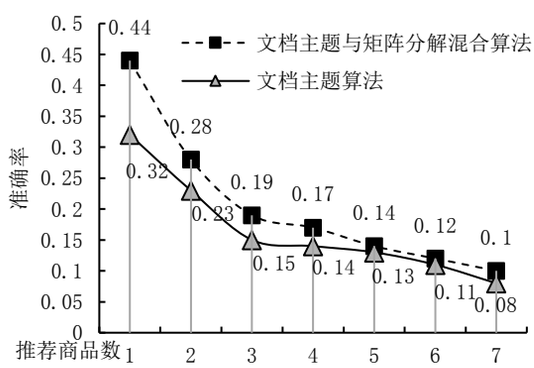

图2　混合算法和文档主题算法准确率对比图

推荐商品数越多,准确率逐渐降低,当推荐商品数达到7时,文档主题与矩阵分解混合算法准确率较文档主题算法提高25%,与参考文献[10]测评趋势一致,因此本文选用文档主题与矩阵分解混合算法。该混合算法优点在于使准确率有较好的提高,缺点是推荐序列商品差异性小,多样性有待提高[10]。

2 主题加权融合协同过滤算法

2.1 加权融合

为了充分利用文档主题与矩阵分解混合算法准确率高、改善数据稀疏性的优势，弥补其推荐序列商品差异性小的不足，将该混合算法与基于物品的协同过滤算法进行加权融合。基于物品的协同过滤算法是一种对用户行为的评分预测分析，通过余弦相似度方法计算物品之间属性相似度，降低热门物品权重来实现长尾商品推荐，具有多样性强的优点。李艳等[11]通过改进LDA-MURE主题模型，形成准确率更高的个性化音乐推荐方案；廉涛等[12]提出结合潜在因素模型和邻域方法的混合协同过滤算法，能够有效缓解数据稀疏性问题；蒋明明[13]深入挖掘评分评论信息，提出引入话题模型的协同过滤算法，多样性强且更加直观；王末等[14]采用动态加权方式有效将多种混合过滤算法融合，较传统算法的准确度及召回率有显著提高；姜信景等[15]提出组合推荐算法，将推荐结果由混合算法和协同过滤算法两种方式生成，该组合推荐算法性能优于同类算法。受到上述研究成果的启发，以加权融合的方式结合两种算法的优势进行深层次研究，本文提出一种改进主题加权融合协同过滤算法，可更好地基于用户评分及评论信息进行智能推荐。

首先设置自变量符号，针对商品a在基于物品的协同过滤算法推荐候选集得分是P_a^{item}，在文档主题与矩阵分解混合算法推荐候选集得分是P_a^{LDAMF}，则商品a后融合的最终总得分为$P_a^{LDAMFitem}$。基于物品的协同过滤算法的权重是β，那么文档主题与矩阵分解混合算法的权重是(1-β)，在对两种模型得分进行指数加权[16]后，可获得最终推荐列表的得分，即为商品被推荐给用户的概率。

$$P_a^{LDAMFitem} = \beta \cdot P_a^{item} + (1-\beta) \cdot P_a^{LDAMF} \quad \cdots\cdots\cdots\cdots (3)$$

公式(3)表示将每个用户在两种算法中Top-80结果生成推荐候选集合。通过指数加权来融合算法，并最终将Top-10的推荐序列反馈给用户。

2.2 主题加权融合与基于物品的协同过滤算法的性能对比

多样性Diversity描述了候选集物品之间差异，因此多样性和相似性具有对应关系，假设sim(a,b)∈[0,1]表示商品a和b之间的相似度，用户v的推荐列表R(v)的多样性定义如公式(4)所示，推荐系统整体多样性由所有用户推荐列表的平均值而得，见公式(5)。

$$Diversity = 1 - \frac{\sum_{a,b \in R(v), a \neq b} s(a,b)}{0.5|R(v)|(|R(v)|-1)}$$

$$Diversity = 1 - \frac{\sum_{x,y \in R(v), x \neq y} s(x,y)}{0.5|R(v)|(|R(v)|-1)} \quad \cdots\cdots (4)$$

$$Diversity = \frac{1}{|V|}\sum_{v \in V} Diversity(R(v)) \quad \cdots\cdots\cdots\cdots (5)$$

表1 主题加权融合协同过滤算法与基于物品的协同过滤算法性能对比

测试指标	主题加权融合协同过滤算法	基于物品的协同过滤算法
准确率	0.27	0.22
多样性	0.28	0.27

由表1可知，主题加权融合协同过滤算法在准确率方面较基于物品的协同过滤算法提高22.7%，同时又保证了良好的多样性推荐效果。

3 基于大数据技术的农产品智能推荐方法实验结果与分析

3.1 数据来源与实验架构

基于大数据技术的农产品智能推荐方法研究实验构建了Spark大数据集群计算平台，以其内存计算的优势减少了迭代计算时的IO开销，交互式编写程序更为便捷。Spark是一个全面的软件栈，包括Spark SQL、Spark Streaming、MLlib、GraphX、Spark Core、独立调度器、YARN、Mesos模块。

实验数据来源于京东商城(www.jd.com)和中国农产品网(www.zgncpw.com)，依托解析器、JQUERY选择器抓取8 324 153条农产品销售记录[17]，包括时间、商品ID、商品名称、用户ID、用户名称、商品详情、评分、评论等。本实验选取京东商城米面杂粮销售记录子集下50 088名用户对634种商品信息的评分评论数据，平均评论词条达103个单词。随机选取80%数据集为训练集，剩余20%数据为测试验证集。根据农产品销售特性[18-19]，该文件部分数据如表2所示。农产品智能推荐方法实验架构共分为三层：智能推荐层、数据处理层、数据存储层，如图3所示。

智能推荐层：评分评论源数据经过预处理后，分别通过文档主题与矩阵分解混合算法和基于物品的协同过滤算法分析处理，生成商品推荐候选集，经加权融合权重，最终形成个性化排序后，从而形成推荐列表。

数据处理层：采用容错性好、易与机器学习和图计算相结合的Spark Streaming流计算框架；训

表2　京东米面杂粮销售平台大米部分实验抓取数据

时间	商品ID	商品名称	用户ID	用户名	商品详情	评分	评论
2017-09-08 17:24	https://item.jd.com/958912.html	十月稻田 长粒香大米 东北大米	02	j***唐	重量:5 kg 产地:黑龙江省哈尔滨市 包装:普通 价格:39.0元	4	一直买这,两年多了吃这个大米,好吃！以前那个快递员重的东西就送到单元楼下了,现在这个直接停到小区外！
2017-11-09 23:15	https://item.jd.com/885987.html	金龙鱼 东北大米 盘锦大米	01	叔叔叔叔去哪了	重量:5 kg 产地:辽宁 当季新米 包装:普通 价格:29.9元	5	第一次买的吃起来不错,后来买的吃着就不行了。
2018-06-18 18:16	https://item.jd.com/3479262.html	华润五丰 东北大米 优选东北珍珠米5 kg	04	j***i	重量:5 kg 产地:黑龙江省	4	买了两种米不知道哪个好吃,吃过以后再评价,两个袋子的米放在一个箱子里又没送到家。
2018-08-01 19:39	https://item.jd.com/418657.html	福临门 东北大米 水晶米 中粮出品大米	03	亲***i	重量:5 kg 产地:吉林 包装:普通 价格:29.9元	5	吃着还行,已经是n次购买,物流很快。

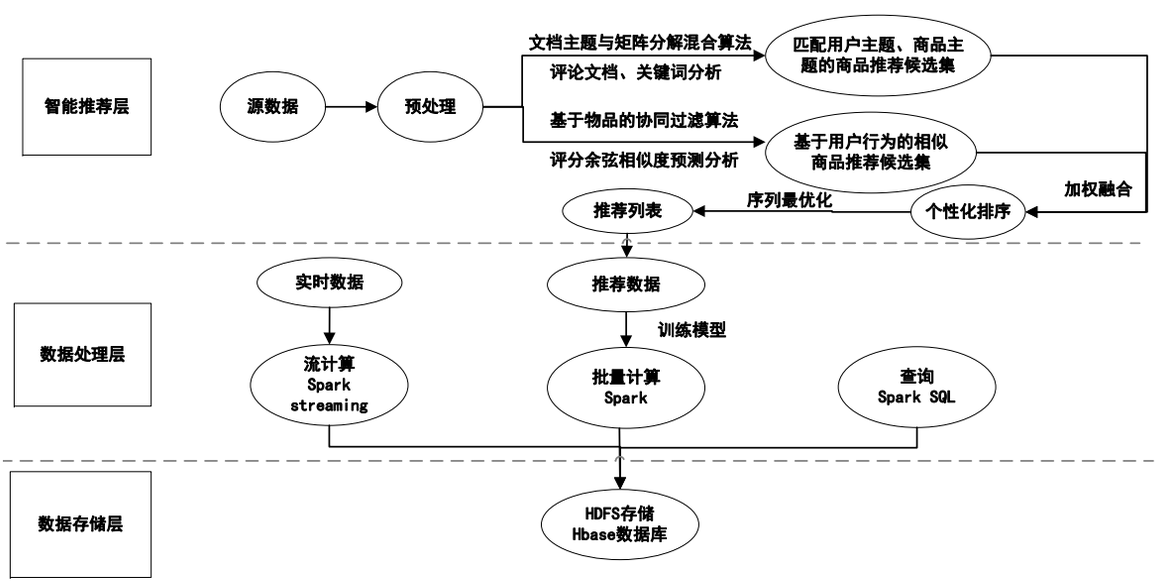

图3　农产品智能推荐方法研究实验架构流程图

练模型后进行Spark批处理;同时查询Spark SQL数据库获得最终结果[20]。

数据存储层:将处理数据传送至分布式文件系统HDFS和HBase数据库中存储调用。

3.2　结果与分析

对于推荐候选集,计算其融合2种算法权重后的平均推荐概率,提取匹配特征信息,由公式(3)推导出加权指数β与单一推荐候选集间的关系。

$$\beta = \frac{P_a^{LDAMFitem} - P_a^{LDAMF}}{P_a^{item} - P_a^{LDAMF}} \quad \cdots\cdots(6)$$

若$P_a^{LDAMFitem}$为用户实际购买评分,加权训练后得到基于物品的协同过滤算法融合权重为0.455,则文档主题与矩阵分解混合算法融合权重为0.545。在测试集中得到AUC值为0.66,模型分类较好,说明该实验融合权重适用。选择综合值高的前10个进行序列穿插最优化,生成推荐结果。

3.3　误差测评

采用均方根误差(RMSE)来评价模型预测准确度,实际值与实验预测值相差越小,说明该方法准确度越高。其公式为:

$$RMSE = \sqrt{\frac{\sum_{u,i \in T}(r_{ui} - \hat{r}_{ui})^2}{|T|}} \quad \cdots\cdots\cdots\cdots\cdots\cdots(7)$$

其中：r_{ui}表示农产品的实际评分，$\hat{r}_{ui}$是智能推荐方法预测的评分，|T|为测试验证集总量。

如下代码便可求出：

```
val RMSE=math.sqrt(MSE)
println(Root Mean Squared Error = "+RMSE)
```

其输出的均方根误差为：

Root Mean Squared Error=0.787

如图4所示，在同等情况下，智能推荐方法比

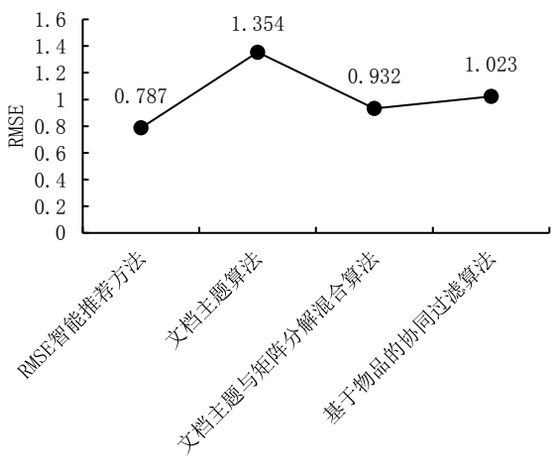

图4 智能推荐方法与传统算法均方根误差对比图

单一基于物品的协同过滤算法和单一文档主题算法、单一基于文档主题与矩阵分解混合算法的均方根误差更小，预测更加准确，说明算法在预测准确度和多样性上有所提高。

4 结 论

本研究在文档主题算法、文档主题与矩阵分解混合算法、基于物品的协同过滤算法、主题加权融合协同过滤算法基础上，进行基于大数据的农产品智能推荐方法实验，并得出以下结论：

（1）文档主题与矩阵分解混合算法准确率较文档主题算法提高25%，文档主题与矩阵分解混合算法优点在于使准确率有较好的提升，不足是推荐序列商品差异性小。

（2）主题加权融合协同过滤算法能够充分利用文档主题与矩阵分解混合算法准确率高、改善数据稀疏性的优势，同时结合了基于物品的协同过滤算法多样性强的优点，保证了推荐商品的高准确率和多样性。

（3）改进后的主题加权融合算法通过Spark大数据处理平台验证后表明：基于大数据的农产品智能推荐方法能够有效提取商品特征并进行序列最优化推荐，保证在多种算法验证测评情况下误差最小，推荐质量及执行效率得到提升。

该方法也可应用于现有涉农网站中的农产品智能推荐，从而加大和促进农产品网上销售。

参考文献：

[1] 熊回香，窦 燕.基于LDA主题模型的标签混合推荐研究[J].图书情报工作，2018，62(3)：104-113.
[2] 张 敏，丁弼原，马为之，等.基于深度学习加强的混合推荐方法[J].清华大学学报(自然科学版)，2017，57(10)：1014-1021.
[3] 黄 璐，林川杰，何 军，等.融合主题模型和协同过滤的多样化移动应用推荐[J].软件学报，2017，28(3)：708-720.
[4] 王源龙，孙卫真，向 勇.基于Spark的混合协同过滤算法改进与实现[J].计算机应用研究，2019，36(3)：855-860.
[5] 顾军华，谢志坚，武君艳，等.基于图游走的并行协同过滤推荐算法[J].智能系统学报，2019，14(4)：743-751.
[6] 董亚楠.主题模型与矩阵分解模型在信息流推荐中的应用[D].北京：北京工业大学，2017.
[7] 吴飞飞，姬东鸿，吕超镇.基于LDA和CTR的用户模型分析[J].计算机工程与应用，2016，52(6)：50-54.
[8] 艾 黎.基于商品属性与用户聚类的个性化服装推荐研究[J].现代情报，2015，35(9)：165-170.
[9] 孙 红，韩 震.融合物品热门因子的协同过滤改进算法[J].小型微型计算机系统，2018，39(4)：638-643.
[10] 项 亮.推荐系统实践[M].北京：人民邮电出版社，2012：64-70.
[11] 李 艳，李葆华，王金环.一种新的基于LDA-MURE模型的音乐个性化推荐算法[J].吉林大学学报(理学版)，2017，55(2)：371-375.
[12] 廉 涛，马 军，王帅强，等.LDA-CF：一种混合协同过滤方法[J].中文信息学报，2014，28(2)：129-135，150.
[13] 蒋明明.引入话题模型的协同过滤推荐算法研究[D].北京：北京理工大学，2016.
[14] 王 末，郑晓欢，王卷乐，等.基于混合过滤的地学数据个性化推荐方法设计与实现[J].地理研究，2018，37(4)：814-824.
[15] 姜信景，齐小刚，刘立芳.个性化信息推荐方法研究[J].智能系统学报，2018，13(2)：189-195.
[16] 宋 峰，陈桂芬，王国伟.基于GIS与空间数据库技术的土壤肥力评价研究[J].吉林农业科学，2014，39(6)：43-46.
[17] 魏倩男，贺正楚，陈一鸣.基于网络爬虫的京东电商平台数据分析[J].经济数学，2018，35(1)：77-85.
[18] 陈娉婷，官 波，沈祥成，等.大数据时代开放式农业信息知识库构建研究[J].东北农业科学，2018，43(5)：60-64.
[19] 马成忠，邓继峰，魏亚伟，等.基于灰色理论的辽宁省农业产业结构分析与预测[J].东北农业科学，2016，41(4)：106-112.
[20] 许明杰，蔚承建，沈 航.Spark并行化基于物品协同过滤算法[J].计算机工程与设计，2017，38(7)：1817-1822.

（责任编辑：王丝语）