基于 Spark 大数据的智能推荐系统设计与实现

杨运强
（辽宁生态工程职业学院　辽宁　沈阳　110101）

【摘要】针对大数据环境下推荐系统面临的效率与准确性挑战，本研究提出了基于 Spark 大数据技术的智能推荐系统设计方案。采用三层架构（数据层、计算层、应用层）组织系统，并结合协同过滤、基于内容的推荐及深度学习技术实现推荐算法。实验结果显示，所设计的系统在准确率、召回率等方面表现优异，尤其 DeepFM 算法在高阶非线性特征交互学习上具有明显优势，本研究为解决大数据环境下的推荐问题提供了可行的技术路径。

【关键词】Spark 大数据；智能推荐；协同过滤
【中图分类号】TP302　　【文献标识码】A　　【文章编号】1009-5624（2025）02-0150-03

0　引言

大数据时代，海量数据不断涌现，如何从中汲取有价值的信息成为各行业的重要课题。国务院发布的《促进大数据发展行动纲要》明确指出，运用大数据提升公共服务和产业发展水平[1]。推荐系统作为大数据应用的典型代表，在电商、内容平台等领域发挥重要价值。本文基于 Spark 大数据技术，设计并实现了一个智能推荐系统。

1　Spark 大数据技术的内涵与原理

Spark 作为一个新兴的大数据处理框架，以其高效、易用和通用性等特点在业界备受推崇。其核心架构基于内存计算，通过有向无环图来描述任务之间的依赖关系，从而构建阶段和任务集，并利用弹性分布式数据集来实现数据的存储与转换。弹性分布式数据集是一种只读、分区存储的数据结构，每个分区由多个块组成，默认块大小为 64 MB[2]。Spark 为用户提供了丰富的算子，可以支持复杂的数据处理逻辑。Spark 采用主从架构，驱动程序负责任务的调度，集群管理器负责资源的管理，而工作节点则负责具体任务的执行。通过“另一种资源协商器”和资源管理系统等资源调度框架，Spark 能够支持上万并发任务。由于弹性分布式数据集存储在内存中，避免了频繁的输入输出操作，使得 Spark 在迭代计算（如网页排名算法和聚类算法等）场景中的性能非常优越[3]。此外，Spark 流处理基于微批次架构，将流式数据按照 100 ms 等时间间隔进行切分，再处理每个批次的数据，因此非常适合准实时场景的应用。机器学习库是 Spark 的机器学习模块，提供了聚类、分类、回归等常用算法。而图计算引擎则是一个领先的图计算引擎，基于批处理模型，提供了丰富的图算法支持。

2　基于 Spark 大数据的智能推荐系统设计

2.1 系统总体架构与工作逻辑

本文设计的智能推荐系统采用三层架构：数据层、计算层和应用层。数据层基于分布式文件系统，分片存储用户行为日志、商品信息等结构化、非结构化数据。计算层以 Spark 为核心，利用其弹性分布式数据集、有向无环图等机制，构建数据预处理、特征工程、模型训练、推荐生成等计算流程。应用层包括推荐结果展示、用户反馈收集等，与前端系统对接。

2.2 系统关键技术实现

2.2.1 数据预处理与特征工程

本系统的数据预处理和特征工程主要基于 Spark 结构化查询语言和机器学习库实现。首先，利用 Spark 结构化查询语言对分布式文件系统中的原始数据进行提取、转换和加载处理，通过查询语句完成数据清洗、转换和集成。其次，对处理后的数据进行特征提取[4]。对于文本类数据如商品描述、用户评论等，采用词向量模型算法将其转换为词矢量。最后，词向量模型通过浅层神经网络学习词语的分布式表示，每个词语映射为一个固定维度（如 200 维）的实值矢量。其核心思想是基于上下文语境最大化似然概率。

$$\mathcal{L}(\theta)=\prod_{w\in C}\prod_{u\in \mathrm{Context}(w)}P(u\mid w;\theta) \tag{1}$$

式中：C 为语料库；$\mathrm{Context}(w)$ 为词语 w 的上下文窗口；$P(u\mid w;\theta)$ 为给定词语 w 生成上下文词语 u 的条件概率，通过 Softmax 函数计算。

$$P(u\mid w;\theta)=\frac{\exp(\boldsymbol{t}\cdot\boldsymbol{m})}{\sum_{v\in V}\exp(\boldsymbol{v}\cdot\boldsymbol{w})} \tag{2}$$

式中：$\boldsymbol{t}$ 和 $\boldsymbol{m}$ 分别为词语 u 和 w 的矢量表示；V 为词表。通过负采样等优化手段，Word2Vec 可以在百万级词表上高效训练。对于类别型特征如性别、年龄段等，则采用独热编码将其转换为 0~1 矢量。此外，还引入词频-逆文档频率来刻画词语在文本中的重要程度。最终，将各类结构化、非结构化特征组合为高维稀疏特征矢量，用于后续的推荐模型训练。

基金项目：2023 年辽宁省教育厅基本科研项目（JYTMS20230997）；2023 年度辽宁省教育厅基本科研项目（JYTMS20230995）。
作者简介：杨运强（1979—），男，辽宁沈阳，硕士，研究方向：大数据技术。

2.2.2 基于协同过滤的推荐算法实现

协同过滤是本系统的核心推荐算法之一。它基于用户间的相似性，为目标用户推荐有相似兴趣用户喜欢的物品。本系统采用 Spark 机器学习库中的交替最小二乘法算法实现隐式反馈协同过滤。给定 m 个用户、n 个物品的隐式反馈数据，交替最小二乘法算法通过最小化损失函数学习用户矩阵 $\boldsymbol{U}$ 和物品矩阵 $\boldsymbol{V}$。

$$\min_{U,V}\sum_{(i,j)\in K}c_{ij}(p_{ij}-\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j)^2+\lambda(\|\boldsymbol{U}\|_F^2+\|\boldsymbol{V}\|_F^2) \quad (3)$$

式中：K 为已知的用户-物品交互集合；c_{ij} 为置信度；p_{ij} 为用户 i 对物品 j 的隐式反馈值（如点击、收藏等）；$\boldsymbol{u}_i$ 和 $\boldsymbol{v}_j$ 分别为用户 i 和物品 j 的 k 维隐矢量；λ 为正则化参数。通过交替固定 $\boldsymbol{U}$ 和 $\boldsymbol{V}$，交替最小二乘法算法迭代求解上述优化问题，得到用户和物品的隐语义表示。在预测阶段，对于目标用户 i，计算其与所有物品的评分矢量。

$$\hat{r}_i=\boldsymbol{U}_i\boldsymbol{V}^{\mathrm{T}} \quad (4)$$

选取 TOP-N 个得分最高的物品生成推荐列表。此外，为缓解协同过滤的冷启动和稀疏性问题，本系统还引入基于内容的推荐作为补充，利用物品元数据计算内容相似度，为新用户或长尾物品提供推荐。交替最小二乘法算法通过设置合适的隐矢量维度（如 10~200）、迭代次数（如 10~20）和正则化参数（如 0.01~0.1），在海量稀疏数据上表现出优异的扩展性和准确性[5]。

2.2.3 基于内容的推荐算法实现

为进一步提升推荐效果，本系统还实现了基于内容的推荐算法。该算法利用物品的内容信息如标题、描述、类别等，计算物品之间的相似度，然后为用户推荐与其历史喜欢物品内容相似的新物品。本系统主要采用词频-逆文档频率加权的余弦相似度来衡量物品间的内容相似性。首先，对物品的文本信息进行分词、去停用词等预处理，然后生成物品-词语矩阵 $\boldsymbol{X}$，其中 n 为物品数，p 为词语数。矩阵元素 $\boldsymbol{x}_{ij}$ 表示词语 j 在物品 i 中的词频-逆文档频率权重。

$$\boldsymbol{x}_{ij}=tf_{ij}\cdot\lg\left(\frac{n}{df_j}\right) \quad (5)$$

式中：tf_{ij} 为词语 j 在物品 i 中的词频；df_j 为包含词语 j 的物品数。词频-逆文档频率权重表示了一个词语在某个物品中的重要程度，同时考虑了该词语在整个物品集合中的稀有程度。在生成物品表示后，可以计算任意两个物品 i 和 j 之间的余弦相似度。

$$\mathrm{sim}(i,j)=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|} \quad (6)$$

式中：$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 分别为物品 i 和 j 的词频-逆文档频率矢量。对于给定用户，首先找到其历史喜欢的物品集合，然后对每个候选物品，计算其与历史物品集合的平均相似度，最后选取 TOP-N 个平均相似度最高的物品生成推荐列表。为平衡推荐的多样性和精确性，可以通过设置合适的 N 值（如 10~100）和相似度阈值（如 0.1~0.5）来调节推荐结果。此外，还可以利用局部敏感哈希等技术对物品进行快速近似最近邻搜索，提高算法的在线响应速度。

2.2.4 基于深度学习的推荐算法实现

为进一步提升推荐效果，本系统在传统推荐算法的基础上引入了深度学习技术。在推荐模型训练阶段，采用深度因子分解机算法学习用户和物品的高阶交互特征。深度因子分解机由因子分解机和深度神经网络两部分组成。因子分解机部分用于提取用户和物品的二阶交互。

$$\hat{y}_{\mathrm{FM}}=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle x_ix_j \quad (7)$$

式中：$w_0\in\mathbb{R}$ 为全局偏置项；$w_i\in\mathbb{R}$ 为第 i 维特征的权重；$\langle\cdot,\cdot\rangle$ 表示矢量点积；$\boldsymbol{v}_i\in\mathbb{R}^k$ 为第 i 维特征的隐矢量（k 取 10~100）。深度神经网络部分则用于学习高阶非线性特征交互。

$$\hat{y}_{\mathrm{DNN}}=\mathrm{ReLU}(\boldsymbol{W}_n\cdot\mathrm{ReLU}(\boldsymbol{W}_{n-1}\cdot\cdots\mathrm{ReLU}(\boldsymbol{W}_1x+\boldsymbol{b}_1)\cdots)+\boldsymbol{b}_n) \quad (8)$$

式中：$\boldsymbol{W}_n$ 和 $\boldsymbol{b}_n$ 分别为第 i 层的权重矩阵和偏置矢量；$\mathrm{ReLU}(x)=\max(0,x)$ 为修正线性单元激活函数。最终，深度因子分解机将因子分解机和深度神经网络的输出结合生成预测评分。

$$\hat{y}=\sigma(\hat{y}_{\mathrm{FM}}+\hat{y}_{\mathrm{DNN}}) \quad (9)$$

式中，$\sigma(x)=1/(1+\mathrm{e}^{-x})$ 为 Sigmoid 函数，将预测评分映射到 0~1 范围。模型通过最小化交叉熵损失函数进行端到端训练。

$$\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}(y_i\lg\hat{y}_i+(1-y_i)\lg(1-\hat{y}_i))+\lambda\|\Theta\|_2^2 \quad (10)$$

式中：$y_i\in\{0,1\}$ 为第 i 个样本的真实标签（用户是否对物品感兴趣）；θ 为模型参数集合，λ 为 L2 正则化系数（取 10^{-4}~10^{-2}）。通过自适应矩估计等自适应梯度下降算法优化模型，可在百万级特征维度的超大规模数据上高效训练[6]。在推荐生成阶段，利用训练好的深度因子分解机模型对用户-物品对进行评分预测，为每个用户选取 TOP-N 个评分最高的物品生成推荐列表。

3 系统典型应用案例评估

3.1 实验方案

为验证本文设计的基于 Spark 大数据的智能推荐系统的有效性，选取电商领域的典型应用案例进行评估。实验在 4 台配置为 32 核 CPU、128 GB 内存、2 TB 固态硬盘的 CentOS 7.2 服务器上进行，Spark 版本为 2.4.0。实验数据集包括亚马逊电商平台 2015 年 5 月到 2016 年 4 月的用户行为日志和商品元数据，共 1 亿用户、2 亿商品和 10 亿交互记录。数据预处理阶段，利用 Spark SQL 对原始 JSON 格式数据进行清洗、转换和集成，得到统一的用户-商品评分矩阵。特征工程阶段，采用 200 维词向量模型（Word2Vec）矢量表示商品文本，独热编码表示类别型特征。推荐算法选取交替最小二乘法（alternating least

squares,ALS)(隐矢量维度 100、迭代次数 10、正则化系数 0.01)、词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)加权余弦相似度(Top-50 相似商品)、深度因子分解机(deep factorization machines,DeepFM)、深度神经网络(deep neural network,DNN)结构(512, 256, 128]、随机失活率 0.5、L2 正则化系数 0.001),对比评估其 Top-10 推荐结果的准确率、召回率、覆盖率和多样性等指标。最后,对推荐系统的实时响应能力进行压力测试。

3.2 结果分析

从表 1 可以看出,DeepFM 在准确率和召回率上显著优于 ALS 和 TF-IDF 推荐,这主要得益于其利用 DNN 学习高阶非线性特征交互的能力。在覆盖率方面,TF-IDF 推荐以 0.67 的高覆盖率胜出,说明内容推荐在发掘长尾物品方面有独特优势。而 ALS 推荐的多样性指标最高,达到 0.82,这是由于其利用随机负采样增加训练样本的多样性。综合来看,DeepFM 在各方面实现了较好的平衡,但仍需进一步优化以提升覆盖率和多样性。

表 1　不同推荐算法的离线评估结果　　%

算法	准确率	召回率	覆盖率	多样性
ALS	0.213	0.285	0.590	0.820
TF-IDF	0.184	0.246	0.670	0.740
DeepFM	0.238	0.326	0.620	0.780

表 2 展示了推荐系统在不同并发用户数下的平均响应时间。当并发数从 100 增加到 10 000 时,响应时间从 25 ms 增加到 268 ms。系统在 1 000 并发用户时响应时间为 73 ms,满足了 100 ms 的实时交互需求。当并发数进一步增大时,响应时间开始恶化,这主要是由于单个执行器的内存和 CPU 资源逐渐耗尽。后续可以通过增加执行器数量、内存和核数来提升系统吞吐量。

表 2　不同并发用户数下的系统平均响应时间

并发用户数/个	平均响应时间/ms
100	25
500	46
1 000	73
5 000	187
10 000	268

4　结语

本研究通过引入协同过滤、基于内容的推荐以及深度学习等多种推荐算法,有效提升了推荐的准确率和召回率。实验评估表明,特别是 DeepFM 算法在处理高阶非线性特征交互时展现显著的优势。尽管如此,系统在推荐的覆盖率和多样性方面仍存在一定的局限性,且随着并发用户数的增加,响应时间有所延长。未来的工作可以着眼于优化算法以提高推荐的覆盖面和多样性,同时探索更高效的负载均衡策略和技术手段来增强系统的实时响应能力和整体吞吐量,以更好地服务于实际应用场景。

【参考文献】

[1] 周杨玥,李世锋,李林. 基于 Spark 的智能菜品推荐系统设计与实现[J]. 软件工程,2024,27(2):69-73.

[2] 云有为. 基于 LFM-MBGD 的特色农产品智能推荐系统研究[D]. 合肥:安徽农业大学,2023.

[3] 李娜,蒋晓敏. 基于大数据挖掘技术的 IPTV 智能推荐系统的设计与实现[J]. 广播电视网络,2022,29(6):73-76.

[4] 张敏,程鹏翔. Spark 平台下基于聚类挖掘的影视资源智能推荐[J]. 信息技术,2021,45(9): 30-33,38.

[5] 刘丽娜. 基于改进 Apriori 算法的毕业生就业智能推荐系统[J]. 辽东学院学报(自然科学版),2021,28(2):130-135.

[6] 沈黄金. 基于 Spark 的农产品智能推荐系统研究[D]. 大庆:黑龙江八一农垦大学,2021.

(上接第 143 页)

【参考文献】

[1] 张楷,孙超,刘家豪,等. 基于计算机视觉的输电塔位移监测 ROI 关键点法[J]. 振动测试诊断,2024,44(5):849-856,1033.

[2] 洪书颖,张东霖. 语义信息处理方式分类的车道线检测技术研究综述[J]. 计算机工程与应用,2024,44(5):1-8.

[3] 李琼,考月英,张莹,等. 面向无人机航拍图像的目标检测研究综述[J]. 图学学报,2023,42(5):1-6.

[4] 宋涛. 图像中值滤波算法在智能采摘机器人中的应用[J]. 农机化研究,2024,12(5):1-6.

[5] 陈红江,丁宏权,曹雄恒,等. 基于机器视觉的无标记图像低频加速度计动态校准方法[J]. 中国测试,2024,15(7):1-7.

[6] 田胜景,韩一男,赵宪通,等. 3D 稀疏卷积结构下融合空间点与体素关系建模的 LiDAR 点云跟踪方法[J]. 电子学报,2023,23(16):1-14.

[7] 杨乐,何大阔,王正松. 基于机器视觉的二维码检测教学实验系统设计[J]. 控制工程,2024,31(9):1722-1728.

[8] 王成豪,王煜,高庭辉,等. 基于机器视觉的预制节段梁锈迹检测[J]. 科学技术与工程,2024,24(26):11432-11440.

[9] 刘慧,董振阳,田帅华. 融合点云和体素信息的目标检测网络[J]. 计算机工程与设计,2024,45(9):2771-2778.

[10] 熊春宝,孙长保,牛彦波. 基于视觉与加速度测量的结构动态位移融合估计[J]. 天津大学学报(自然科学与工程技术版),2024,57(9):891-901.

[11] 闫玉刚,李成族,陈玉洁,等. 实时机器视觉技术在纺织智能生产中的应用[J]. 东华大学学报(自然科学版),2024,12(6):1-12.

[12] 王超, 齐天玉, 杨青祥. 基于机器视觉和最优组合应变影响线的车辆荷载识别方法研究[J]. 桥梁建设, 2024, 54(4): 61-68.