

基于多模型融合与数据库进阶优化的高性能混合推荐系统

孙智立, 陈旭, 路泽宇, 谭征

(烟台大学 计算机与控制工程学院, 山东 烟台 264003)

摘要: 传统的单模型推荐系统考虑的衡量指标较为单一, 数据量较少或数据特征较多时准确率受到制约。繁冗的模型预训练与未经数据库优化的推荐过程, 导致系统无法及时更新用户和物品的最新数据, 推荐过程响应缓慢, 直接影响用户的使用体验。设计一种基于多模型融合与数据库进阶优化的高性能混合推荐系统。该系统在召回阶段对双塔模型进行改进, 使用内容基模型增强物品内容的向量化表达。在排序阶段使用NCF、DeepFM及改进了数据抽样方式和损失函数的RBPR进行排序, 兼顾满足准确性与解决冷启动问题的需求。在融合阶段使用瀑布融合与带权融合, 筛选出用户最感兴趣的物品, 创新性地应用Redis和向量数据库Faiss提高推荐准确率及推荐效率。实验结果表明, 相较于单模型且无数据库优化的推荐系统, 混合推荐系统在冷启动及常规场景下的推荐准确率及推荐速度均有明显提升。其为解决推荐系统中的冷启动和数据特征工程问题提供了新思路, 在媒体平台、电商等领域具有广阔的应用前景。

关键词: 混合推荐系统; 深度学习; 冷启动; 特征工程; 数据库优化

DOI: 10.11907/rjdk.251165

扫描二维码阅读全文: 

中图分类号: TP311.5

文献标识码: A

文章编号: 1672-7800(2025)012-0153-08

A High-Performance Hybrid Recommendation System Based on Multi Model Fusion and Advanced Database Optimization

SUN Zhili, CHEN Xu, LU Zeyu, TAN Zheng

(School of Computer and Control Engineering, Yantai University, Yantai 264003, China)

Abstract: The traditional single model recommendation system considers relatively single measurement indicators, and accuracy is constrained when there is a small amount of data or a large number of data features. The cumbersome model pre training and recommendation process without database optimization result in the system being unable to update the latest user and item data in a timely manner, leading to slow response times in the recommendation process and affecting the user experience. This article proposes a high-performance hybrid recommendation system based on multi model fusion and advanced database optimization. The system improves the dual tower model during the recall stage by using a content based model to enhance the vectorized expression of item content. In the sorting stage, NCF, DeepFM, and RBPR with improved data sampling and loss functions are used for sorting, balancing accuracy and solving cold start problems. In the fusion stage, waterfall fusion and weighted fusion are used to screen out the items that users are most interested in. Redis and vector database Faiss are innovatively applied to improve recommendation accuracy and efficiency. The experimental results show that compared to a single model recommendation system without database optimization, the hybrid recommendation system proposed in this paper has significantly improved recommendation accuracy and speed in both cold start and conventional scenarios. This study provides new ideas for solving the problems of cold start and data feature engineering in recommendation systems, and has broad application prospects in media platforms, e-commerce, and other fields.

Key Words: hybrid recommendation system; deep learning; cold start; feature engineering; database optimization

收稿日期: 2025-03-12

基金项目: 山东省自然科学基金项目(ZR2023MF063)

作者简介: 孙智立(2004-), 男, 烟台大学计算机与控制工程学院学生, 研究方向为机器学习与推荐系统; 陈旭(2004-), 男, 烟台大学计算机与控制工程学院学生, 研究方向为推荐系统; 路泽宇(2005-), 男, 烟台大学计算机与控制工程学院学生, 研究方向为推荐系统; 谭征(1968-), 男, 硕士, 烟台大学计算机与控制工程学院副教授, 研究方向为机器学习、数据挖掘和自然语言处理。本文通讯作者: 谭征。

0 引言

随着人工智能的快速发展,机器学习技术在推荐系统领域得到了广泛的应用。然而,传统单模型方法在处理稀疏数据或多特征数据时仍然面临挑战,包括预训练时间过长和无法充分利用多维度数据特征进行推荐等问题。如何在利用深度学习技术实现精准推荐的同时提高推荐速度,仍是一个亟待解决的问题。

本文旨在研究一种能够高效推荐且充分利用多维度数据特征的推荐系统。基于此,提出了一种基于多模型融合与数据库进阶优化的高性能混合推荐系统。该系统通过结合多种内容基、协同过滤和冷启动算法等多种模型,并搭载 MySQL、Redis 以及 Faiss 数据库,既显著提升了推荐的准确率,又大幅提高了推荐的响应速度,为复杂场景下推荐系统的输入训练及输出过程的工程构建提供了新的解决方案。实验结果表明,该混合推荐系统的推荐准确率和推荐速度均明显优于传统单模型推荐系统。此外,混合推荐系统在媒体平台和电商等领域具有广泛的应用前景。本文主要贡献总结如下:

(1)提出一种新的端到端的基于多模型融合与数据库进阶优化的高性能混合推荐系统,并对已有的模型进行改进。其中,双塔模型增加了基于 Word2Vec 的内容基模型,用于增强物品的向量化表达;RBPR 改进了损失函数并换用基于贝叶斯平均的负采样方法。

(2)结合内存数据库 Redis 与向量数据库 Faiss,支持高并发高负载情况下的快速响应。同时,实现了端到端的增量训练过程,能够动态更新模型。

(3)使用基准数据集 ml-1m 对混合推荐系统进行评估。结果显示,与单模型或无数据库优化的系统相比,本文提出的混合推荐系统具有更加均衡稳定的性能。

1 相关工作

早期的推荐系统往往依赖于单一的推荐模型。协同过滤(Collaborative Filtering)是推荐系统中最经典的方法之一,其核心思想是通过学习用户-物品交互关系来为用户推荐相似物品,开创性地引入了相似度计算的概念^[1]。但在大数据场景下,存在较长的模型训练时间和推荐请求响应时间。内容基推荐则通过分析物品特征为用户推荐具有相似内容的物品,尽管内容基推荐在训练和推荐时占用的时间均较少,但该模型受限于物品标签质量和标签数量等因素,泛化性较差^[2]。总体而言,单模型推荐系统虽然在特定场景下表现良好,但面对复杂的推荐需求时,其性能表现往往不足。

随着深度学习的发展,研究者开始将深度学习技术与传统推荐模型融合起来,通过将传统的矩阵分解(MF)^[3]、

因式分解(FM)^[4]等引入神经网络,显著提高了推荐系统的准确率。

He 等^[5]提出的神经协同过滤(NCF)在 CF 的基础上,引入神经网络来处理显式反馈与隐式反馈。传统的 CF 通过构建用户-物品交互矩阵,并利用(MF)挖掘用户-物品交互中深层次的潜在特征。然而,传统的矩阵分解方法存在诸多局限,如无法捕捉到复杂的非线性交互关系,以及在处理大规模数据时面临较高的时间和空间成本等。为此,NCF 采用广义矩阵因式分解(GMF)与多层感知机(MLP)取代了传统的矩阵分解方法。

DeepFM 是由 Guo 等^[6]提出的一种结合因式分解机(FM)与深度神经网络(DNN)的混合模型,旨在同时捕捉低阶与高阶特征,广泛应用于点击通过率(CTR)的预测中。DeepFM 分为 FM 组件与 Deep 组件,组件间共享输入。通过组件融合,DeepFM 作为一个完整的端到端模型,无需进行任何额外的特征工程,节省了对原始数据进行特征工程所需的时间。在处理多特征数据时表现尤为出色。通过对特征进行嵌入和线性变换的组合,DeepFM 能够挖掘出更深层次的隐式信息。然而,在大数据场景下,单纯的神经网络推荐模型难以承受不断攀升的训练时间成本,无法及时更新数据。

Yin 等^[7]提出的 HEC-GCN 通过超图建模用户的多行为交互(如点击、收藏、分享),利用级联图卷积分层提取行为间的协同特征,有效解决了传统图模型难以捕捉复杂行为关联性的问题,在电商等推荐场景中显著提升了用户多行为数据的利用率。此外,Jia 等^[8]提出了一种偏好校正的多模态图卷积推荐网络,通过融合文本、图像等异构数据,并结合对抗学习校正用户偏好偏差,增强了推荐系统对多模态语义的理解能力,为处理跨模态数据稀疏性提供了新思路。

Feng 等^[9]提出的 RBPR,主要用于在数据稀疏的极端情况下处理用户冷启动问题。现有的大多数协同过滤方法严重依赖用户已有的大量数据。在冷启动场景下^[10-12],由于严重缺乏用户数据,协同过滤无法通过仅有的少量数据正确推理出用户的兴趣,准确为用户作出推荐。为了克服该问题,RPBR 融合了矩阵概率分解(PMF)与贝叶斯个性化排名(BPR)模型。通过在训练中使用面向成对排名的方法进行差异化区分,同时充分利用了显式反馈数据与隐式反馈数据。

Gao 等^[13]提出的 DR、Nguyen 等^[14]提出的基于 KNN 的 CF 模型以及 Chen 等^[15]提出的 ANN 搜索方法,为研究如何从大规模物品中获得召回物品提供了不同的方法。DR 使用光束搜索与 EM 优化方法,构建出物品簇以方便检索;基于 KNN 的 CF 模型使用 KNN 扩展提高 CF 的推荐性能;ANN 使用基于神经相似度度量,依照相似度快速检索召回物品。

多模型混合推荐系统成为近年来的研究热点,旨在通过结合多种不同推荐指标的模型来提高推荐性能。Chen

等^[16]提出了一种新的多模型混合推荐系统。通过采用 DeepFM 与图卷积网络 (DeepFM-GCN) 的多任务推荐算法, 图卷积网络能够深度挖掘知识图谱中的辅助实体信息, 从而有效补充推荐任务中的稀疏语义信息。通过交叉压缩单元结合神经网络的方式弥补了特征表示不足的影响, 提高了推荐的准确率。

尽管已有研究在多模型混合推荐系统中取得了重要进展, 但仍然存在一些问题。如在冷启动场景下, 因为数据稀疏导致的推荐效果不准确; 在大数据场景下, 无法高效且及时地处理大量数据, 完成特征的提取、模型的更新与推荐内容的快速检索。针对这些问题, 本文提出一种全新的基于多模型融合与数据库优化的混合推荐系统, 通过结合协同过滤模型、冷启动模型、深度学习算法与数据库性能优化, 有效解决了上述冷启动情景与大数据情景下的推荐问题。

2 混合推荐系统

推荐系统主要依靠属性信息 (如用户性别、用户年龄、物品标签等) 和交互信息 (如评分、点赞、转发、评论等) 进行推荐。

2.1 协同过滤模型

使用神经网络实现的 NCF 和 DeepFM 具有良好的预测性能。在排序阶段, 使用协同过滤模型, 能够很好地保证推荐的准确率与数据的利用率。在非冷启动的常规场景下, 协同过滤模型对推荐结果的质量起主导作用。

2.2 RBPR

在原始的 RBPR 公式中, 损失 F 视作由 F_{PMF} 与 F_{BPR} 融合而成, 使用融合因子 α 对两部分损失进行贡献控制。 F_{PMF} 由已有的显式反馈的损失与使用奇异值分解 (SVD) 进行填补的反馈的损失计算得到, F_{BPR} 由对 2 个物品对进行排序预测的损失计算得到。尽管在使用机器学习的方法来实现 RBPR 时获得了良好的效果, 但当使用深度学习方法时却存在一些问题。在 PMF 的预先处理过程中, 需要使用 SVD 对用户-物品交互矩阵进行填充。当推荐系统中的用户与物品的数量逐渐增多时, SVD 填充矩阵所花费的时间与空间成本是令系统难以承受的。此外, 基于 SVD 进行填充物品的评分值的选取方法也显著影响着模型训练效果。基于此, 本文提出的一种新的处理未反馈数据的方式——负采样填充法, 并在此基础上改进原有的 RBPR。

2.2.1 负采样填充法

相较于类似“洪泛法”对所有未反馈数据进行补充的 SVD, 使用小批量精准负采样的方式能够显著加速模型的训练过程。此外, SVD 使用矩阵奇异值分解的方式对未反馈数据进行填充, 在面对物品的评分数量和评分均值差异过大时, 这种差异将被 SVD 进一步放大, 从而使冷启动问题更加难以处理。改进的 RBPR 采用参数修正后的贝叶

斯平均方法对未反馈物品数据进行补全, 如式 (1) 所示。

$$\hat{r}_i = \frac{b + \bar{R}_i k + \sum_{r \in R_i} r}{\left(b + \bar{R}_i k + \sum_{r \in R_i} r\right) + \left(b + (p_{revise} - \bar{R}_i) k + |\bar{R}_i| - \sum_{r \in R_i} r\right)} \quad (1)$$

其中, $\hat{r}_i$ 表示物品 i 的贝叶斯平均值, 用于表示物品的热度值。 $\bar{R}_i$ 表示物品 i 的评分总体的平均值, 平均值越高, 则表明该物品的总体评价越偏向正面, 反之则偏向负面。 b 表示偏置常量, 用于调整贝叶斯平均值的平衡性, 此处取 $b = 1.0$ 。 k 表示系数常量, 用于调节贝叶斯平均值的平滑程度, 此处取 $k = 6.0$ 。 p_{revise} 表示修正常量, 用作先验假设的常量值表示, 此处取 $p_{revise} = 5.0$ 。 $|\bar{R}_i|$ 表示物品 i 评分总体内的样本数量, 样本数量越多, 则可能表明物品 i 更加热门 (无论是正面的还是负面的)。 $\sum_{r \in R_i} r$ 表示该物品评分总体内所有评分的加和。

图 1 为根据公式 (1) 计算得到的物品的贝叶斯平均值。X 轴表示评分总体的平均值, 对应公式 (1) 中的 $\bar{R}_i$; Y 轴表示物品 i 评分总体内的样本数量, 对应公式 (1) 中的 $|\bar{R}_i|$; Z 轴表示物品 i 的贝叶斯平均值。由此计算得出贝叶斯平均值作为物品 i 的热度值。随着评分数量的增加, 即与所有用户与该物品交互的次数增加, 贝叶斯平均值逐渐上升, 表明更多的交互会提升物品的热度值。同理, 随着 $\bar{R}_i$ 的增加, 贝叶斯平均值逐渐上升, 表示物品评价越偏向正面, 热度越高。随着 $\bar{R}_i$ 与 $|\bar{R}_i|$ 的上升, 曲线将达到最大值, 代表着拥有大量正面评价的物品为当下最热门的物品; 反之则表示拥有极少交互且评价均偏向负面的物品为当下最冷门的物品。

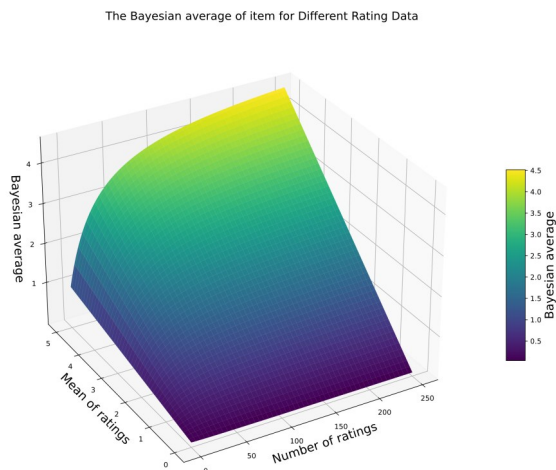


Fig. 1 Bayesian mean with parameter correction

图 1 经过参数修正的贝叶斯平均值

在参数修正的贝叶斯平均模型中, 通过使用额外的偏置常量 b 、系数常量 k 与修正常量 p_{revise} , 使最终的贝叶斯平均值的曲面整体较为平滑, 没有显著的值突变点。同时, 整体的数值被控制在一个较为紧凑的区间内, 避免了使用贝叶斯平均值作为热度值时可能出现的 2 个问题: 一是因

值域存在突变点导致的物品热度突变;二是因极差较大造成的热门物品常驻现象。此外,通过参数调整,那些即使评价非常正面但交互数量极少的物品,在初始时仍然获得较低的贝叶斯平均值。随着这类物品的交互数量增长,贝叶斯平均值在初期会以较高的增长率快速上升,随后逐渐趋于平稳。这一机制使得原本未被大众发现的小众优质物品能够迅速获得曝光,并最终成为稳定的热门优质物

$$F = -\frac{\alpha}{2} \frac{\left(\sum_{(u,i,j) \in T+T'} \left((r_{ui} - r_{\max} \sigma(\hat{r}_{ui}))^2 + (r_{uj} - r_{\max} \sigma(\hat{r}_{uj}))^2 \right) \right)}{2} + (1-\alpha) \sum_{(u,i,j) \in D_s} \ln(\sigma(\hat{r}_{uj})) - \lambda_{\theta} \|\Theta\| \quad (2)$$

其中,显式反馈损失的计算依赖用户已有的显式反馈数据与负采样数据,补充未反馈数据的方式由SVD变更为负采样。

贝叶斯平均公式可以通过统计物品*i*一段时间内的交互数据,计算出物品*i*在该段时间内的贝叶斯平均值,并将其作为热度值。通过在小批次(通常不超过10个)中选择用户u未交互且热度值较高的物品样本,可以认为这些物品虽然非常热门,但用户u始终未曾与之有过交互,因此将这些物品作为用户u的负样本空间T'。在改进的RPBR中,这些负样本物品将与该名用户u曾交互过的正样本物品进行配对排序训练。

图2表示使用了不同 α 的联合损失公式。其中,X轴表示正负样本间的预测评分差异,对应 F_{PMF} ;Y轴表示正负样本间的预测排名差异,对应 F_{BPR} ;Z轴表示 F_{RBPR} 。随着预测评分差异的扩大,联合损失值逐渐减小,模型在正样本的预测评分上更倾向于接近真实评分。当正负样本的预测排名差异增大时,联合损失值同样呈减小趋势,模型在排序上更倾向于将正样本排位在负样本之前。此外,由于 α 决定着 F_{PMF} 与 F_{BPR} 对联合损失的贡献程度,因此不同的 α 对联合损失的影响十分明显。在冷启动场景下,由于可用的用户显式反馈数据不足,若采用过高的贡献权重或贡献阈值,会形成过于激进的策略。在需要平衡PMF与

品。并且,随着 $\bar{r}_i$ 的降低,物品的评价也越来越偏向负面。在获得交互数量时,能够分配的增长率也越来越小。当物品的评价极度负面时,贝叶斯评价值的增加将趋近于0。

2.2.2 改进的RPBR

改进的RBPR损失计算公式如式(2)所示。原有 F_{BPR} 的计算方式保持不变,对PMF中的显式反馈损失计算与补充反馈损失的计算进行了修改。

BPR对整体RBPR性能贡献的情况下,并基于明确的优化目标,本研究参照Feng等^[7]在原始论文中通过实验得到的最佳的 α 值,改进的RBPR采用 $\alpha = 0.1$ 。

改进的RBPR使得PMF在填补数据时所需的时间大幅减少,同时,在训练模型时对用户显式反馈数据与负样本数据的出现更加敏感。在训练过程中,RBPR通过正负样本配对的方式,专注于学习正样本(即用户已交互的物品)与负样本(用户未反馈但具热度的物品)之间的联合特征表示,从而实现对两类样本的排序预测。基于此,改进的RBPR的输出为排序序列,在混合推荐系统中用于物品排序。

2.3 双塔模型(Dual Tower)

双塔模型(Dual Tower)因其在召回阶段处理大规模推荐任务的高效性而被业界广泛应用。然而,传统的基于协同过滤的双塔模型,不能充分利用用户与物品的大量特征建立特征关联,且面对大量数据时,模型的训练及使用均存在较高的时空复杂度,导致模型的表现不佳。因此,本文提出一种融合了基于Word2Vec模型实现的内容基模型的双塔模型。

2.3.1 基于Word2Vec实现的内容基模型

定义 S_i 表示物品*i*的标签集,用于准确概括物品*i*的内容。其中, $S_i = \{s_{i_1}, s_{i_2}, \dots, s_{i_n}\}$ 。 $s_{i_1}, s_{i_2}, \dots, s_{i_n}$ 为标签集合

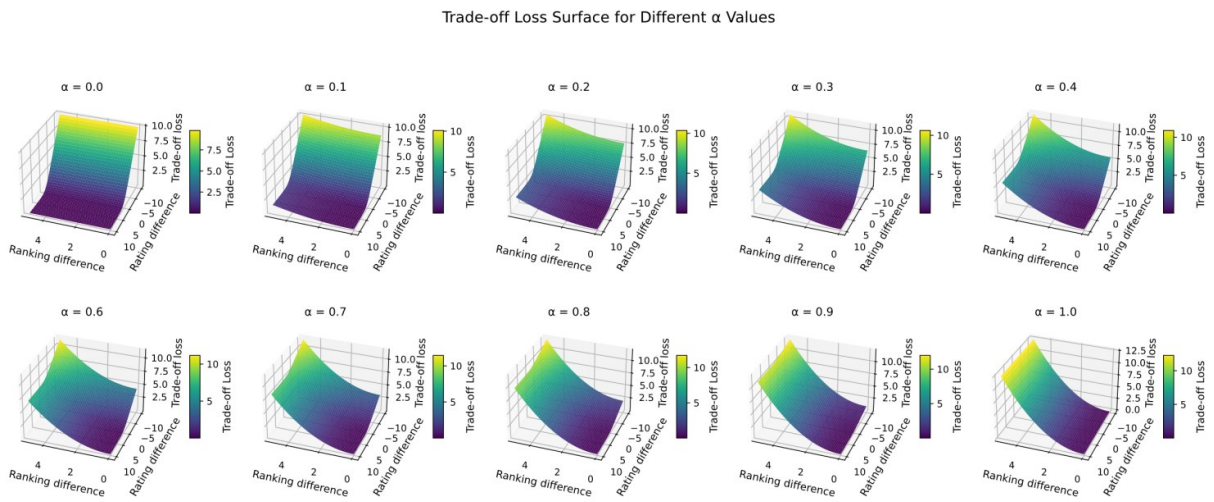


Fig. 2 Improved RBPR loss formula with visualization of different α values

图2 经过改进的RBPR损失公式使用不同 α 的可视化

中的元素,每个元素代表一个由自然语言表示的独立标签。 n 表示标签集的元素个数。由于物品涵盖多种类型,即便属于同一类型,其内容侧重点可能存在差异,因此不同物品标签集合的元素数量存在不同。同一物品的标签集合中的元素总是随着该物品的出现而同时显现,而相似物品之间的标签集合也往往具有高度的相同或相似性。基于此,在召回阶段以单个标签为起点进行扩展搜索时,应优先扩展到与该标签在某物品标签集合中曾经共现的标签,其次是那些与该标签在语义上具有相同或相似意义的标签。随后,根据所获取的所有标签,与物品池中的物品标签集进行匹配,从而获取初步的召回结果。

Mikolov 等^[17]提出的 Word2Vec 模型通过神经网络模型来学习词汇的向量表示,并通过向量捕捉词汇间的语义关系。Word2Vec 模型分为连续词袋架构(CBOW)和跳字架构(Skip-gram)。由于 CBOW 的核心思想是根据上下文词汇预测目标词汇,而 Skip-gram 的核心思想则相反,旨在根据目标词汇预测上下文词汇。在内容基模型的召回机制中,目标标签会被扩展到目标标签集中。基于此,本文选择采用 Skip-gram 模型来实现内容基模型,且使用量化后的物品标签集来表征物品内容。物品 i 的内容表征如式(3)所示。

$$v_i = \frac{1}{|S_i|} \sum_{s \in S_i} v_s \quad (3)$$

其中, v_i 表示物品的内容表征向量, v_s 表示标签 s 的 Skip-gram 向量化表示, $v_i, v_s \in \mathbb{R}^d$, d 为表征向量的维度。通过聚合所有的标签向量并求取平均向量,找到在 d 维空间中物品 i 标签集的“质心”,用作物品 i 的内容表征向量。通过该方法得到的物品表征向量,能够捕捉到物品 i 的主要内容特征,并且不会因为标签数量产生激烈的数值变化,使物品内容的表示更加稳定与可靠,从而解决了大数据场景下物品内容无法被有效表示导致的模型性能损失的问题,适用于物品标签丰富的场景。

2.3.2 改进的双塔模型

改进的双塔模型融合了原始的基于 CF 的双塔模型与基于 Word2Vec 实现的内容基模型。图 3 展示了改进的双塔模型的处理流程图。

模型结合显式数据与隐式数据,使用负采样方法从隐式数据中获得用户的负样本物品。对于用户数据,直接使用基于协同过滤的用户塔进行处理;对于物品数据,使用基于 Word2Vec 的内容基模型获得内容表征向量,再将物品内容表征向量及其他特征使用基于协同过滤的物品塔进行处理,最终在双塔模型内部聚合用户塔与物品塔各自得到的嵌入向量。通过损失函数进行训练,如式(4)所示。

$$F = \sum_{(u,i,j) \in T+T'} \log \left(1 + \exp \left(\varepsilon \left(f(v_u, v_i) - f(v_u, v_j) \right) \right) \right) \quad (4)$$

其中,三元组 (u, i, j) 表示用户 u 对应的正样本物品 i 与负样本物品 j 。所有三元组均来自用户正样本空间 T 与

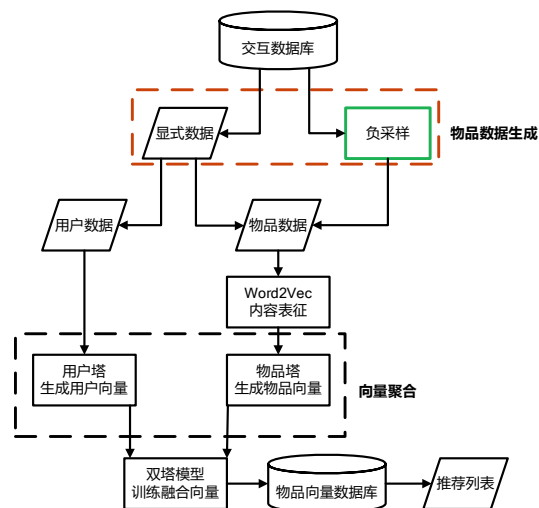


Fig. 3 Improved dual-tower model processing flowchart

图 3 经过改进的双塔模型处理流程图

用户负样本空间 T' 的联合空间。 $f(v_u, v_i)$ 表示向量 u 与向量 i 的余弦相似度,通过用户向量与物品向量直接的余弦相似度,预测用户与物品的评分。 ε 表示额外的损失系数,用于控制损失函数对正负样本相似度差值的敏感性。基于此损失函数,双塔模型将更倾向于区分正样本与负样本,使得模型预测用户对正样本更可能给予正面评价,而对负样本更可能给予负面评价。

综上,双塔模型的模块化设计使得系统能够在任意处理流程节点轻松扩展功能。通过在双塔模型的召回阶段使用数据库存储物品的嵌入向量,可以在大数据场景下显著加速召回过程,减少密集计算量,从而获得更加精确的近似近邻召回结果。

2.4 数据库优化

随着用户与物品数量的增长,传统的无数据库优化的推荐系统将面临巨大的数据吞吐压力。现有研究(如基于树的深度神经网络)多通过改进模型架构提升计算效率^[18]。在实际运行时,若额外考虑模型的计算用时与 Web 页面的响应用时,缓慢的推荐请求处理速度将导致用户获得极差的使用体验。为解决传统推荐系统无法有效处理大规模数据吞吐的情况,本文使用额外的数据库进行数据库吞吐优化,提高推荐系统整体的数据处理能力与响应速度。

2.4.1 Faiss

向量数据库 Faiss 是由 Facebook 开发的一套针对聚类和相似性搜索的近似近邻搜索库,其主要为稠密向量提供高相似度搜索与聚类,支持十亿级别向量的搜索^[19]。Faiss 支持多种索引结构,包括 Flat 索引、倒排文件索引 (IVF)、分层导航小世界索引 (HNSW) 与局部敏感哈希索引 (LSH) 等。在召回阶段,双塔模型需要高效地在大量物品向量中进行近似近邻的查找。由于大数据场景下数据规模庞大,向量数据库中将会存储大量稠密向量,且物品

间的关系可以抽象转为图结构。本文将使用 Faiss 中的 HNSW 索引构建向量数据库,实现大规模向量的相似度搜索,具体如图 4 所示^[20]。

图 4 展示了 Faiss 在召回阶段的使用流程图。召回阶段预训练后将物品向量存储至 Faiss 中,在系统运行时使用向量数据库快速检索出用户向量的近似最近邻,使用双塔模型得到召回结果。由于用户的兴趣变换较快,因此在召回中实时计算用户向量,不再将用户向量存储至 Faiss 中。

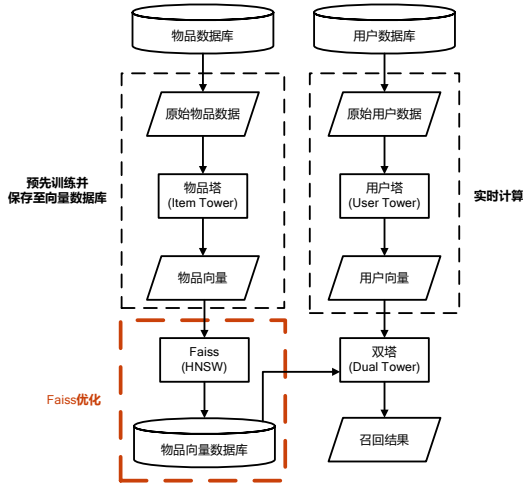


Fig. 4 Faiss processing flowchart

图 4 Faiss 使用流程图

基于 HNSW 索引实现的向量数据库,在搜索过程中拥有良好的性能。通过相似度对大量节点进行筛选,将相似的节点聚集在一起,仅保留必要的通路用以保证搜索的可达性。对于差异度较大的节点,则分散在不同层的不同簇内。在召回阶段,使用基于 HNSW 实现的向量数据库,能够确保以较低的成本存储大量物品向量的同时,仍然保持良好的搜索性能用于大量物品的精准召回。

2.4.2 Redis

Redis 是一个开源的基于内存的高性能键值对数据库,能够提供极高的读写速度,非常适合作为缓存使用。Redis 缓存数据算法伪代码如下:

算法 1 Redis 缓存数据算法

输入:本次查询任务的需求参数, data_key、lock_key 将根据需求参数生成

输出:本次查询任务的查询结果

```

1. BEGIN
2.   IF Redis.exists(data_key) THEN
3.     data = Redis.get(data_key)
4.   ELSE
5.     lock = Redis.setnx(lock_key, expire_time)
6.     IF lock THEN
7.       data = MySQL.query(data_key)
8.       Redis.set(data_key, data)
9.       Redis.delete(lock_key)

```

```

10.    ELSE
11.      WHILE Redis.exists(lock_key) DO
12.        WAIT(1s)
13.      END WHILE
14.      IF Redis.exists(data_key) THEN
15.        data = Redis.get(data_key)
16.      ELSE
17.        lock = Redis.setnx(lock_key, expire_time)
18.        IF lock THEN
19.          data = MySQL.query(data_key)
20.          Redis.set(data_key, data)
21.          Redis.delete(lock_key)
22.        ELSE
23.          THROW "Failed to acquire lock after retry"
24.        END IF
25.      END IF
26.    END IF
27.  END IF
28.  RETURN data
29. END

```

算法 1 描述了 Redis 在模型训练、物品召回、召回物品排序以及推荐热门物品时的使用逻辑。该算法专门针对高并发场景中的缓存击穿问题,提出了有效的解决方案。通过使用分布式锁,确保只有一个请求能够访问数据库,并将新数据写入缓存。同时,采用懒加载(Lazy Load)机制,使得仅在缓存中没有相应数据时,请求才会尝试从数据库中获取最新数据。通过过期时间、分布式锁和懒加载机制,有效降低了缓存击穿的风险,保证过期时间后由访问请求进行新数据的更新与写入,保证了数据一致性。

2.5 混合推荐系统

设 r_m 为模型 m 的推荐运算结果, $r_m = i \rightarrow s_i | i \in \mathcal{I}$ 。其中, s_i 表示物品 i 在模型 m 中获得的评分, $\mathcal{I}$ 表示物品池集合。在融合多模型推荐结果时,首先采用瀑布式融合策略:对于通过双塔模型召回的物品,优先使用 NCF 模型进行打分排序。设 r 为系统中的评分,规定 $r \in [0, r_{max}]$ 。在此设置不低于 $0.8 \times r_{max}$ 的预测评分作为 NCF 的输出限制。DeepFM 将 NCF 的输出作为自身的输入,通过使用较 NCF 更复杂的特征挖掘,进一步过滤候选物品,完成了初步的瀑布融合。此处设置不低于 $0.8 \times r_{max}$ 的预测评分作为 DeepFM 的输出限制。RBPR 以召回物品作为输入数据,其设计初衷主要用于解决冷启动问题。由于在数据不足的情况下, RBPR 的预测结果准确率较低,因此使用不低于 $0.6 \times r_{max}$ 的预测评分作为 RBPR 的输出限制。在排序阶段与初步的瀑布融合阶段完成后,对 DeepFM 的输出与 RBPR 的输出使用线性加权融合的方式进行融合,融合公式如式(5)所示。

$$s_i = \sum_{m \in M} w_m \times e^{(\alpha_m + s_{m,i})} \quad (5)$$

其中, s_i 表示物品 i 的融合得分, $M = \{m_{DeepFM}, m_{RBPR}\}$, s_{m_i} 表示使用模型 m 时对物品 i 的预测分数。 w_m, α_m 表示模型 m 的常量参数, 在系统运行时根据侧重需求动态调整。通过公式(5), 融合阶段可以将多模型的输出结果赋予权重, 最终将排序获得权重最高的一批物品作为融合推荐结果输出。

图 5 展示了本文提出的混合推荐系统的整体架构。系统的数据主要来源于运行日志和在线统计, 其中, 核心

数据为用户与物品的交互记录。数据源中的原始数据经过处理后存入数据库层的 MySQL 中。处理推荐任务时, 数据处理层从数据库层中获取数据并进行清洗与包装, 随后将格式化数据送交推荐算法层。在推荐算法层, 物品池中的物品依次通过召回、排序及融合模块, 最终获得推荐结果。推荐结果通过基于 WSGI(Web 服务网关接口)的中间件传输至 Web 展示层, 并由 Nginx 服务器统一管理展示层与中间件层的访问配置。

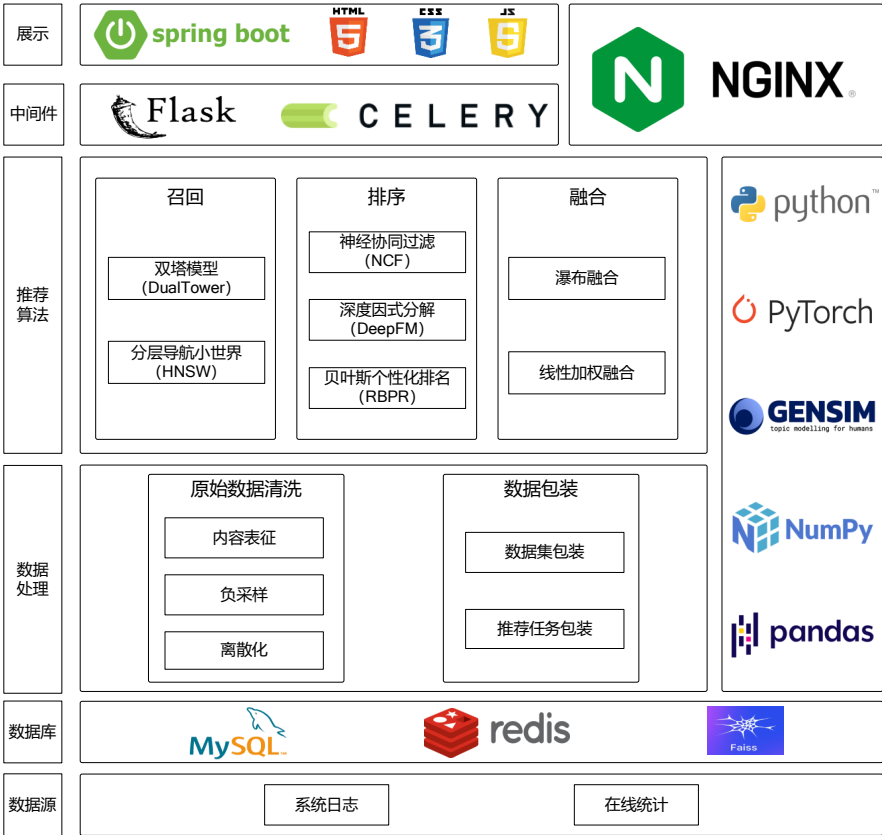


Fig. 5 Overall architecture of hybrid recommendation system

图 5 混合推荐系统整体架构

3 实验

为评估混合推荐系统性能, 本文设计了以下实验方案: 所有实验均在一台配备了 AMD Ryzen 7 5800H CPU 和 16GB RAM 的机器上进行。实验采用 ml-1m 数据集。在实验开始前, 对数据进行了清洗, 并将其随机分为训练集 (80%) 与验证集 (20%)。各模型的训练参数具体设置如表 1 所示。

其中, epoch 表示训练迭代次数, lr 表示学习率, L2 表示 L2 正则权重, patience 表示最大能够接受的测试集上损失不继续下降的迭代次数。实验结果表明, 所提出的混合推荐系统取得了优于现有方法的性能。

表 2 展示了融合模型与其他基线模型的性能指标对比, 评估指标采用 NDCG(归一化折损累计增益)、MRR(平

Table 1 Model training parameters

表 1 模型训练参数表

模型名	训练参数			
	epoch	lr	L2	patience
DualTower	100	1e-5	1e-2	5
NCF	100	1e-4	1e-2	10
DeepFM	100	1e-4	1e-2	10
RBPR	100	1e-3	0.0	5

均倒数排名)、MAP(平均精度)和 RT(响应时间)。对比基线模型为 KGCN-DAE^[21]、RippleNet-Node2Vec^[22]、PIFSA-GNN^[23]等。

由表 2 可知, 融合模型在 NDCG@15 和 MAP@15 指标上表现最佳, 表明其具有良好的精度。尽管在 MRR@15 指标上, 融合模型的表现逊于 DeepFM 和 PIFSA-GNN, 但融合模型具有更低的响应时间, 更适合高并发状态下作出准确的推荐。实验结果表明, 融合模型不仅具备均衡稳定的

Table 2 Model evaluation

表2 模型评价

模型名	评价指标			
	NDCG@15	MRR@15	MAP@15	RT@15/ms
融合模型	0.660 582	0.584 412	0.496 745	3.934 616
CF	0.602 368	0.510 598	0.469 229	1.701 821
NCF	0.412 135	0.238 154	0.110 557	13.859 759
DeepFM	0.609 368	0.776 319	0.087 081	19.082 460
RBPR	0.513 383	0.269 741	0.078 019	15.495 709
KGCN-DAE	0.401 675	0.346 418	0.357 749	10.836 642
RippleNet-Node2Vec	0.619 416	0.446 516	0.214 675	12.478 654
PIFSA-GNN	0.659 134	0.647 162	0.456 473	8.456 968

指标性能,还具有好的组件扩展性与推荐实时性。

本文对混合推荐系统的性能进行了全面评估。实验结果表明,本文数据处理方法与混合推荐算法在推荐任务中表现优异,兼具高效性和鲁棒性。此外,在高负载场景下,系统相比现有模型实现了更低的延迟和更高的吞吐量。

4 结语

本文通过实验证明了基于多模型融合与数据库优化的高性能混合推荐系统的有效性。实验结果表明,本文系统在标准数据集 ml-1m 上取得了优于现有方法的性能,特别是在处理多模型推荐结果融合与高负载情况下快速响应展现了显著的优势。此外,提出了一种隐式数据利用方法,进一步提高了隐式数据的利用率,并显著提升了推荐结果的准确率。尽管本文取得了一定的研究成果,但仍存在一些局限性,如在处理大量训练数据时,模型训练速度较慢,且无法充分利用小样本进行有效学习等。未来的工作将重点集中于提升模型的计算准确率以及探索基于集群的优化训练方案。综上,本文为解决推荐系统中的冷启动和数据特征工程问题,提供了新的视角与方法,并为相关领域的技术发展提供了重要参考。

参考文献:

- [1] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International Conference on World Wide Web, 2001: 285-295.
- [2] PAZZANI M, BILLSUS D. Learning and extracting structural knowledge from the Web[C]// Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, 1997: 21-24.
- [3] HE X, ZHANG H, KAN M, et al. Fast matrix factorization for online recommendation with implicit feedback[C]// Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2016: 549-558.
- [4] RENDEL S. Factorization machines[C]// IEEE International Conference on Data Mining, 2010: 995-1000.
- [5] HE X, LIAO L, ZHANG H, et al. Neural collaborative filtering[C]// Proceedings of the 26th International Conference on World Wide Web, 2017: 173-182.
- [6] GUO H, TANG R, YE Y, et al. DeepFM: a factorization-machine based neural network for CTR prediction[DB/OL]. <https://arxiv.org/abs/1703.04247>.
- [7] YIN Y, ZHU X, WANG W, et al. HEC-GCN: hypergraph enhanced cascading graph convolution network for multi-behavior recommendation[DB/OL]. <https://arxiv.org/abs/2401.08281>.
- [8] JIA X, DONG Y, ZHU F, et al. Preference-corrected multimodal graph convolutional recommendation network[J]. Applied Intelligence, 2023, 53: 3947-3962.
- [9] FENG J, XIA Z, FENG X, et al. RBPR: a hybrid model for the new user cold start problem in recommender systems[J]. Knowledge-Based Systems, 2021, 214: 106732.
- [10] LI H, WANG Y, XIAO Z, et al. DisCo: graph-based disentangled contrastive learning for cold-start cross-domain recommendation[DB/OL]. <https://arxiv.org/abs/2412.15005>.
- [11] LEE H, IM J, et al. MeLU: meta-learned user preference estimator for cold-start recommendation[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 1073-1082.
- [12] SILVA N, CARVALHO D, et al. A deep study about how to conquer first-time users in recommendations domains[J]. Information Systems, 2019, 80: 1-12.
- [13] GAO W, FAN X, WANG C, et al. Deep retrieval: learning a retrievable structure for large-scale recommendations[DB/OL]. <https://arxiv.org/abs/2007.07203>.
- [14] NGUYEN L V, VO Q T, et al. Adaptive KNN-based extended collaborative filtering recommendation services[J]. Big Data and Cognitive Computing, 2023, 7: 106.
- [15] CHEN R, LIU B, ZHU H, et al. Approximate nearest neighbor search under neural similarity metric for large-scale recommendation[C]// Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022: 3013-3022.
- [16] CHEN L, BI X, FAN G, et al. A multitask recommendation algorithm based on DeepFM and graph convolutional network[J]. Concurrency and Computation: Practice and Experience, 2023, 35(2): e7498.
- [17] MIKOLOV T, CHEN K, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.
- [18] ZHU H, ZHANG P, LI G, et al. Learning tree-based deep model for recommender systems[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1079-1088.
- [19] DOUZE M, GUZHVA A, DENG C, et al. The faiss library[DB/OL]. <https://arxiv.org/abs/2401.08281>.
- [20] MALKOV Y, YASHUNIN D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 42: 824-836.
- [21] KAUR G, LIU F, CHEN Y P, et al. The integration of knowledge graph convolution network with denoising autoencoder[J]. Engineering Applications of Artificial Intelligence, 2024, 135: 108792.
- [22] NI W, DU Y, MA X, et al. Research on hybrid recommendation model for personalized recommendation scenarios[J]. Applied Sciences, 2023, 13(13): 7903.
- [23] LI M, MA W, CHU Z, et al. User preference interaction fusion and swap attention graph neural network for recommender system[J]. Neural Networks, 2025, 184: 107116.

(责任编辑:陈维捷)