智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

联邦推荐系统的协同过滤冷启动解决方法

王健宗， 肖京， 朱星华， 李泽远

引用本文:
王健宗, 肖京, 朱星华, 等. 联邦推荐系统的协同过滤冷启动解决方法[J]. 智能系统学报, 2021, 16(1): 178–185.
WANG Jianzong, XIAO Jing, ZHU Xinghua, et al. Cold starts in collaborative filtering for federated recommender systems[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(1): 178–185.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202009032

DOI: 10.11992/tis.202009032

联邦推荐系统的协同过滤冷启动解决方法

王健宗, 肖京, 朱星华, 李泽远
(平安科技(深圳)有限公司, 广东 深圳 518000)

摘 要: 基于联邦学习的推荐系统可以在保护用户隐私的情况下, 联合多方数据, 提升推荐系统的性能, 已经成为推荐领域的研究热点之一。联邦协同过滤是联邦推荐系统中最经典及最常用的算法之一。然而, 针对联邦协同过滤系统的冷启动问题的研究工作相对较少。针对这一问题, 本文提出了一种基于安全内积协议的解决方案。具体地, 在系统中添加新用户或新物品时, 联合多方评分矩阵, 利用安全内积的方法, 对多方数据进行相似矩阵的求解, 从而完成推荐输出。本文在 MovieLens 数据集上对所述方法进行了验证。结果表明: 本方法能够有效解决基于相似度的协同过滤中的冷启动问题, 并且推荐效果也会依据多方数据分布的比例变化。

关键词: 联邦学习; 隐私保护; 数据孤岛; 推荐系统; 协同过滤; 冷启动; 机器学习; 安全内积
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2021)01−0178−08

中文引用格式: 王健宗, 肖京, 朱星华, 等. 联邦推荐系统的协同过滤冷启动解决方法 [J]. 智能系统学报, 2021, 16(1): 178–185.
英文引用格式: WANG Jianzong, XIAO Jing, ZHU Xinghua, et al. Cold starts in collaborative filtering for federated recommender systems[J]. CAAI transactions on intelligent systems, 2021, 16(1): 178–185.

Cold starts in collaborative filtering for federated recommender systems

WANG Jianzong, XIAO Jing, ZHU Xinghua, LI Zeyuan
(Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518000, China)

Abstract: Recommender systems based on federated learning has become one of the research hotspots in the recommender field. However, few studies focused on the cold start problem of federated recommender systems. Under the framework of federated learning, we propose a novel collaborative filtering algorithm for its solution: through involving more rating matrices, we can get a similarity matrix with secure inner product method, and implement the recommendation for new users to the system. In this work, we verify the performance of our method on MovieLens. The results show that our proposal is effective in solving the cold start problem in similarity-based collaborative filtering, and the recommendation effects vary according to the data distribution among different parties.

Keywords: federated learning; privacy protection; data island; recommender system; collaborative filtering; cold start; machine learning; secure inner product

大数据时代政府关于隐私保护的法律法规盛行, 如《通用数据保护条例》(general data protection regulation)[1], 数据在政策的限制下分散在不同的载体/组织中, 因此保护隐私的机器学习在学术界和工业界都备受重视。在实现隐私保护机器学习的各种技术中, 联邦学习 (federated learning, FL)[2] 受到高度重视, 它最初由 Google 提出, 目标是基于分布在每个用户手机上的数据学习一个统一的模型, 用户的原始信息无需转移到中央服务器, 从而实现了隐私保护。经证明, FL 与在原始数据上学习到的传统模型相比, 预测能力几乎相同[3]。

随着互联网和电子商务的迅猛发展, 推荐系统成为企业提高市场竞争力的重要工具。其中, 协同过滤 (collaborative filtering, CF) 是最著名的一种推荐算法[4], 有 2 种实现方法: 一种是基于近邻[5], 另一种是基于矩阵分解[6]。其中基于近邻的协同过滤又分为基于用户的协同过滤和基于物品的协同过滤, 分别依据用户消费行为上的相似性或消费产品间的相似性来实现个性化的产品推

收稿日期: 2020−09−23.
基金项目: 国家重点研发计划“云计算和大数据”重点专项 (2018YFB1003503); 国家重点研发计划“高性能计算”重点专项 (2018YFB0204400); 国家重点研发计划“现代服务业共性关键技术研发及应用示范”专项 (2017YFB1401202).
通信作者: 王健宗. E-mail: jzwang@188.com.

荐。一般来说, 客户的历史信息越详细, 推荐结果越准确。

由于没有足够多的客户数据, 许多中小型公司无法获得满意的推荐模型。为了解决这个问题, 通常采取的解决方案有: 1) 请求另一家拥有庞大客户数据库的公司帮助; 2) 与其他多家拥有相对较小客户数据库公司合作, 共同创建一个大的数据库[7]。公司间无法简单地共享或允许彼此完全访问其数据库, 因为这可能会造成客户隐私数据外泄。文献 [8] 表明, 70%~89.5% 的互联网用户认为个人隐私信息面临泄露风险。鉴于联邦学习处理数据孤岛和隐私保护问题的有效性和实用性, 与联邦学习相结合的协同过滤推荐算法成为目前推荐系统领域的一个研究热点[9]。

冷启动是协同过滤算法应用中经常会遇到的问题, 分为新用户冷启动、新项目冷启动、系统冷启动等。当系统中有新用户加入时, 由于该用户在系统中没有历史评分数据, 不能根据传统算法计算用户间的相似度, 也就无法为其进行推荐, 这就是协同过滤算法的新用户冷启动问题[10]。在现有的与联邦学习相结合的协同过滤推荐算法的研究中, 对用户冷启动问题的研究比较少, 因此对联邦学习协同过滤算法中用户冷启动问题的研究具有迫切的意义。

1　相关工作

本文的研究是 3 个研究主题的交叉点: 联邦学习、协同过滤推荐算法中的隐私保护问题和冷启动问题。

1.1　联邦学习

随着信息革命的发展, 海量的数据在不断产生, 如何合理有效地利用这些数据成为一个热点方向。由于隐私政策的保护, 很多数据不能被轻易地获取, 数据间相互隔离, 形成了一个个数据孤岛。如何建立数据孤岛间沟通的桥梁, 打破数据之间的界限, 成为一个热点方向。谷歌研究院提出了联邦学习的概念, 即通过只在各节点间传递模型参数, 而不分享模型间数据的方式训练一个共享的数据模型。联邦学习成为解决数据隐私保护的一个有利工具。联邦学习旨在满足数据隐私保护、数据安全和政府法规的前提下, 进行数据的使用和建模。根据数据划分的方式, 联邦学习可分为纵向联邦学习以及横向联邦学习[11]。迄今为止, 有许多研究致力于联邦学习算法, 以支持更多的机器学习模型, 包括深度神经网络(deep neural network, DNN)[12]、梯度提升树(gradient boosting decision tree, GBDT)[13]、逻辑回归、支持向量机[14]。

本文主要关注在纵向联邦的场景下实现推荐系统的冷启动问题。

1.2　推荐系统的隐私保护

推荐系统 (recommendation systems, RS) 收集和学习用户对一系列项目的偏好信息, 并预测用户对新物品或项目的兴趣程度, 产生推荐列表。用户的偏好信息可以是显性的 (基本上是通过收集用户的评分) 或隐性的 (基本上是通过监测用户的交互记录, 如访问过的网页、购买过的软件、阅读过的书籍和刷过的短视频等隐性推断关于某物品的兴趣程度)[15-17]。根据输入数据的类型, 推荐模型主要分为协同过滤式推荐系统[17]、基于内容的推荐系统[18] 和基于知识的推荐系统[19]。在实践中, 推荐系统已经被广泛地应用于各种应用中, 如电子商务[20-21]、娱乐[22-23]、新闻[24-25] 和社交平台[26-27]。

由于个人对物品的偏好往往涉及到个人的隐私信息, 长期以来, 推荐系统中如何保护隐私信息受到许多学者关注。许多研究使用差分隐私的方法保护用户评价记录的隐私性[1]。联邦学习通过数据不出本地、仅传输用户梯度的方式, 进一步保障用户的隐私不被窃取。联邦推荐系统可以与差分隐私、多方安全计算等技术结合, 灵活有效地在不泄露用户隐私的前提下实现推荐系统性能的提升。

协同过滤是推荐系统中最常用、应用范围最广的算法之一, 也是本文讨论的主要算法。针对协同过滤算法中的隐私保护问题, 有多种方法可以解决。如文献 [15] 针对集中式数据, 采用随机扰乱技术, 提出了一个保护隐私的协同过滤推荐方案; 文献 [28] 在差分隐私框架中提出了协同过滤算法; 文献 [29] 使用同态加密计算协同过滤过程的中间值, 中间值解密后通过奇异值分解和因子分析产生推荐建议; 文献 [30] 提出了一种基于同态密码的协同过滤算法; 文献 [31] 提出了一种新的兴趣点推荐隐私保护框,在联邦学习中采用安全聚合的策略来学习特征交互模型; 文献 [32] 提出了一种新的分布式矩阵分解框架用于兴趣点推荐, 该框架具有可扩展性, 能够保护用户隐私。

1.3　协同过滤及其冷启动问题

协同过滤是一种基于矩阵分解的推荐算法。在已知用户的历史评分矩阵 $\boldsymbol{R}$ 的前提下, 使用较低维的用户特征矩阵 $\boldsymbol{U}=\{\boldsymbol{u}_1\,\boldsymbol{u}_2\,\cdots\,\boldsymbol{u}_N\}$ 和物品特征矩阵 $\boldsymbol{V}=\{\boldsymbol{v}_1\,\boldsymbol{v}_2\,\cdots\,\boldsymbol{v}_M\}$ 的乘积 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}$ 拟合评分矩阵。在进行推荐时, 通过用户特征和物品特征向量的

内积 $\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j$ 计算出用户对某一物品的预测评分，即用户可能对该物品感兴趣的程度。其中，用户特征和物品特征都是通过对历史评分的学习训练得到的，当系统中添加新的用户和新的物品时，它们的特征向量是未知的，即产生了初始推荐的问题，相关的算法称为协同过滤的冷启动。

针对协同过滤算法中冷启动问题，文献 [33] 提出了一种基于协同矩阵分解的用户冷启动推荐算法，来处理用户冷启动问题。文献 [34] 将计算相似性的不对称方法与矩阵分解和基于典型性的协同过滤 (tyco) 相结合，实现了一种改进的电影推荐算法。然而目前对冷启动问题的研究一般基于单方数据库，具有一定的局限性，有少数学者对多方参与的协同过滤冷启动问题进行了探索。文献 [35] 引入联邦多视图矩阵分解方法，将联邦学习框架扩展到具有多个数据源的矩阵分解。经证明，该方法对于多数据源冷启动推荐有较好的预测效果。

本文提出的方法联合多方数据，打通了数据孤岛，在进行隐私保护的同时，解决了协同过滤中的冷启动问题。

2　纵向联邦协同过滤中的冷启动定义

在本节中，对纵向联邦协同过滤中的冷启动问题给出正式的公式定义。

本文考虑采用基于纵向联邦学习的协同过滤方法。其中，联邦协同过滤可以由多方进行联合训练，为了方便起见，本文均以两方为例。假定联邦参与公司 A、B 都是半诚实的，这意味着他们会遵守协议，但也会尽可能地从执行中推断信息。因此在本文中，由于评分属于隐私信息，A、B 双方均不能直接交换评分。在此假定 A 需要解决新用户冷启动问题，与 B 联合进行基于物品的协同过滤训练，最终 A、B 均能获得两方物品的相似度矩阵，B 通过相似度矩阵与物品评分对 A 物品进行评分预测，将预测值排序并返回 A 推荐物品 id。最终，在不泄露评分信息的情况下，A 获得了针对新用户的物品推荐顺序。

如图 1，机构 A 为了解决本地新用户的冷启动问题，与机构 B 进行合作。假设 A 与 B 是纵向联合，A 与 B 共有用户 n 个，已经进行了对齐处理；共有 m 个物品，其中 A 有 m' 个物品，B 有 $m-m'm-m'$ 个物品。令评分为 $[v_{\min},v_{\max}][v_{\min},v_{\max}]$ 中的一个整数，对于物品 $i(1\leqslant i\leqslant m)$，评分向量为 $\boldsymbol{V}_i=(v_{1i},v_{2i},\cdots,v_{ni})$，其中 $v_{ui}(1\leqslant u\leqslant n)$ 代表用户 u 对物品 i 的评分。假设 n 个用户中，A 有 n' 个新用户，但对于 B 不为新用户。因此 A 拥有评分矩阵 $\boldsymbol{V}_{ui}(n'+1\leqslant u\leqslant n,1\leqslant i\leqslant m')$，相应地，B 机构拥有评分矩阵 $\boldsymbol{V}_{ui}(1\leqslant u\leqslant n,m'+1\leqslant i\leqslant m)$。该研究问题是设计一个联邦学习纵向协同过滤算法，让 A 能够在不泄露双方信息的前提下，通过 B 的信息完成对新用户 $u(1\leqslant u\leqslant n')$ 的推荐。

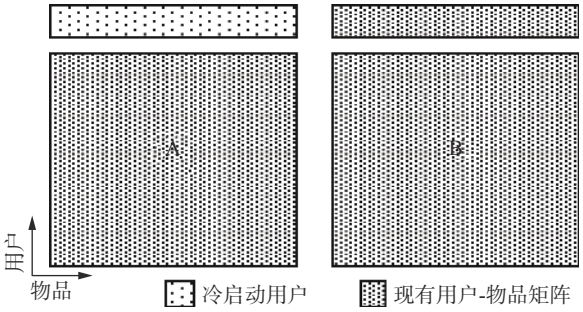

图 1　带有冷启动用户的纵向划分

Fig. 1　Vertical partition with cold start users.

3　联邦推荐冷启动

3.1　协同过滤冷启动算法

根据文献 [36] 提出的框架，协同过滤主要分为 3 个步骤：

1) 物品相似度计算：根据评分信息，计算物品 i 与其他物品相似度；

2) 邻近样本选择：这一步主要是为了提高推荐算法的效率和精确度；

3) 预测推荐：对用户 u 的评分预测，并将排序最高的前 N 个推荐给用户。

1) 计算物品相似度：为了计算 A 方物品与 B 方物品的相似度，采用皮尔森相关系数进行计算，令 i 为 A 方物品，j 为 B 方物品，则皮尔森系数为

$$\mathrm{sim}_{ij}=\frac{\sum_{u=n'+1}^{n}(v_{ui}-\overline{v_i})(v_{uj}-\overline{v_j})}{\sqrt{\sum_{u=n'+1}^{n}(v_{ui}-\overline{v_i})^2}\sqrt{\sum_{u=n'+1}^{n}(v_{u,j}-\overline{v_j})}}$$

式中：$v_{u,i}$ 表示用户 u 给物品 i 的评级；$\bar{v}_i=\dfrac{\sum_{u=n'+1}^{n}v_{ui}}{n-n'}$；$\mathrm{sim}_{ij}$ 代表物品 i 与 j 相似度，范围在 [−1, 1]。为了使 A、B 两方能够联合计算，将其分为 c_{ui}、c_{uj} 2 个部分：

$$\mathrm{sim}_{ij}=\sum_{u=n'+1}^{n}c_{ui}c_{uj}$$

其中：

$$c_{ui}=\frac{v_{ui}-\bar{v}_i}{\sqrt{\sum_{u=n'+1}^{n}(v_{ui}-\bar{v}_i)^2}} \tag{1}$$

$$c_{uj}=\frac{v_{uj}-\bar{v}_j}{\sqrt{\sum_{u=n'+1}^{n}(v_{uj}-\bar{v}_j)^2}} \tag{2}$$

显然, 计算 $c_{ui}(c_{uj})$ 只需要 $v_{ui}(v_{uj})$ 的信息, 因此 A、B 两方均可在本地计算 c_{ui}、c_{uj}。令 $C_i=(c_{n'+1i},c_{n'+2i},\cdots,c_{ni}),C_j=(c_{n'+1j},c_{n'+2j},\cdots,c_{nj})$, 计算 sim_{ij} 即可转为计算 C_i 和 C_j 的内积。

2) 选择近邻: 由于本文针对的是新用户的冷启动推荐问题, 将所有物品都作为邻近样本, 允许使用不相似的物品来进行计算, 增加推荐覆盖率。

3) 预测推荐: 为了对机构 A 的新用户进行推荐, 采用 B 机构里的评分信息进行预测:

$$\mathrm{pred}_{ui}=\frac{\sum_{j=m'+1}^{m}v_{uj}\mathrm{sim}_{ij}}{\sum_{j=m'+1}^{m}|\mathrm{sim}_{ij}|} \tag{3}$$

式中: pred_{ui} 代表用户 u 对物品 i 的预测评分, 对于 A 机构的新用户, 有 $1\leqslant u\leqslant n',1\leqslant i\leqslant m'$。其中预测值计算可在 B 机构进行, 且 B 机构对预测值结果排序, 将排在前 N 个物品发送给 A 机构, 完成对新用户的推荐。

3.2　安全内积

文献 [37] 提出了一种基于第 3 方的一种安全内积计算协议。

为了计算 C_i 与 C_j 的内积 $<C_i,C_j>$, 加入一个第 3 方, 在不泄露各方数据下进行数据的汇总, 如算法 1 所示。

算法 1　安全内积算法 $<C_i,C_j>$

多方　A、B 以及第 3 方 T;

输入　A: C_i; B: C_j;

输出　A: r_A; B: r_B, 且有 $r_A+r_B=<C_i,C_j>$。

1) T 方随机产生 2 个随机向量 $\boldsymbol{x},\boldsymbol{y}$, 以及一个随机数 r, 且令 $z=<\boldsymbol{x},\boldsymbol{y}>-r$, 将 $(\boldsymbol{x},r)$ 传给 A, $(\boldsymbol{y},z)$ 传给 B;

2) A 将 $C_i+\boldsymbol{x}$ 传给 B;

3) B 将 $C_j-\boldsymbol{y}$ 传给 A;

4) A 计算可公开信息 $r_A=<C_i,C_j-\boldsymbol{y}>-r$;

5) B 计算可公开信息 $r_B=<C_i+\boldsymbol{x},C_j>-z$。

由于 r_A 及 r_B 里面各包含 2 个随机项, 对 C_i 及 C_j 的隐私信息均进行了保护, 因此不会泄露各方的隐私数据, 最终 A, B 双方均可利用公开的 r_A、r_B 进行内积 $<C_i,C_j>$ 的计算, 达到了隐私保护的作用, 流程如图 2 所示。

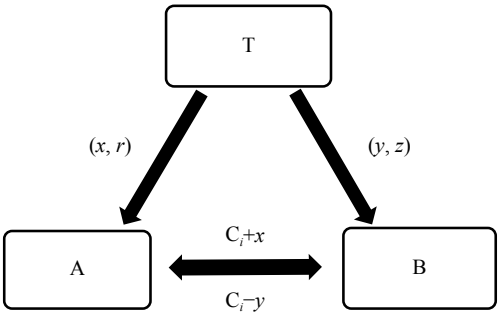

图 2　安全内积流程

Fig. 2　Overview of secure inner product.

3.3　联邦协同过滤冷启动算法

为了使 A、B 双方能够在数据不泄露的情况下进行新用户的推荐, 将协同过滤与安全内积相结合, 构建联邦学习协同过滤冷启动解决算法。框架内容如图 3 所示, 由受信任的第 3 方 T 以及 A、B 两方构成。如前文所述, 假设 A、B 两方都是半诚实的, 同时第 3 方 T 也不与 A、B 两方勾结。

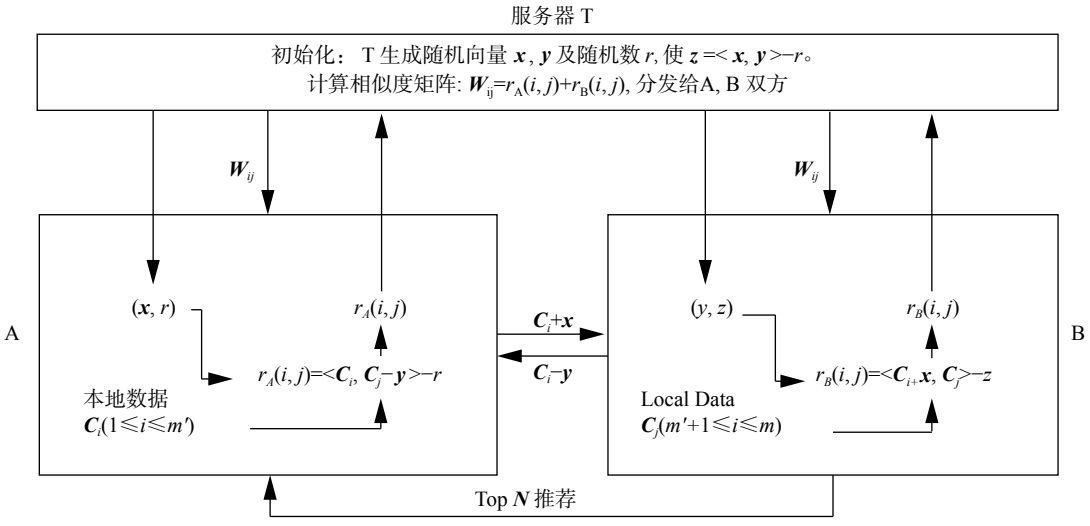

图 3　联邦协同过滤冷启动框架

Fig. 3　Overview of federated learning system for cold start in collaborative filtering

联邦协同过滤冷启动算法的主要步骤包括：

1) 基于安全内积算法，A、B 端根据式 (1)、(2) 在本地分别计算 $C_i(1 \leqslant i \leqslant m')$，$C_j(m'+1 \leqslant i \leqslant m)$；

2) 第 3 方 T 随机产生 2 个随机向量 $\boldsymbol{x}$、$\boldsymbol{y}$ 以及一个随机数 r，且令 $z=<\boldsymbol{x},\boldsymbol{y}>-r$，将 $(\boldsymbol{x},r)$ 传给 A，$(\boldsymbol{y},z)$ 传给 B；

3) A 将 $C_i+\boldsymbol{x}(1 \leqslant i \leqslant m')$ 传给 B；B 将 $C_j-\boldsymbol{y}$ $(m'+1 \leqslant j \leqslant m)$ 传给 A；

4) A 分别计算 $r_A(i,j)=<C_i,C_j-\boldsymbol{y}>-r(1 \leqslant i \leqslant m',$ $m'+1 \leqslant j \leqslant m)$ 传输给 T；同理 B 相应计算 $r_B(i,j)$ 传输给 T；

5) T 计算 $\mathrm{sim}_{ij}=r_A(i,j)+r_B(i,j)$，并构建相应的相似矩阵 $\boldsymbol{W}_{ij}$(由 $\mathrm{sim}_{i,j}$ 元素构成) 分发给 A、B 双方；

6) B 根据 $\boldsymbol{W}_{ij}$ 相似矩阵以及本地用户 $u \in [1,n']$ 对物品 $j \in [m'+1,m]$ 的评分信息，由式 (3) 对 $\mathrm{pred}_{ui}, u \in [1,n'], i \in [1,m']$ 进行预测，并对预测结果进行排序，最终将预测值最高的前 N 个物品 ID 发送给 A，完成用户冷启动推荐过程。

由图 3 中可看出，每一方并不能直接收到对方的原始评分，且由于增加了随机项，不能从中间结果进行对原数据的反推，因此达到了隐私保护的作用。

4　实验与结果

4.1　度量标准

本文采用 Top-N 推荐，采用 2 种不同的评估方法进行模型评价。

1) 阈值击中评价

推荐评估指标采用精确率 Precision、查全率 Recall 以及 $F_1-\mathrm{score}$。

$$\mathrm{Precision}=\frac{\sum_{u=1}^{n'}|R_u \cap T_u|}{\sum_{u=1}^{n'}|R_u|}$$

$$\mathrm{Recall}=\frac{\sum_{u=1}^{n'}|R_u \cap T_u|}{\sum_{u=1}^{n'}|T_u|}$$

$$F_1-\mathrm{score}=2\frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}$$

式中：R_u 是 B 方基于相似矩阵以及用户 u 的评分信息进行对 A 方的推荐列表；T_u 是新用户 u 对于 A 中物品的非空且评分大于阈值 c 的用户行为列表。

2) 留一法 (leave-one-out) 评价

对于测试集里的每个用户，都保留其最新一条评分信息进行验证。由于在评价过程中对每个用户都进行排序比较耗时，所以本位采用了常用的策略[35-36]，即对于每个用户随机抽取 30 个没有评分的物品，并且对其评分预测排序。在此设定推荐个数为 10，将该 30 个物品与最新评分物品的预测评分进行排序，若最新评分物品排在前 10 则视为击中。评估指标采用命中率 (hit ratio, HR)[37] 以及归一化折现累计增益 (normalized discounted cumulative gain, NDCG) 来判断排序列表的性能。HR 直观地衡量测试物品是否在前 10，而 NDCG 根据击中的位置进行评估。

4.2　数据集

本文实验采用 GroupLens 提供的网络开源数据集 MovieLens 1M 为例，整个数据集包括了 6 040 个电影用户对 3 706 部电影的 1 000 209 条评分，评分范围为 1~5。由于本文针对基于物品协同过滤的冷启动问题，因此只取数据集中的用户−电影评分集作为实验数据。

该评分数据集均对用户及电影进行了 ID 处理，因此在进行计算过程中用户 ID 及电影 ID 不视为隐私数据。同时有较多数的电影只被极少数的用户进行评分，该部分电影对新用户的推荐起较少的作用，因此将用户评分率低于 10% 的电影进行删除，处理后的用户−电影数据维度为 6 040 498。

为了衡量冷启动推荐的效果，将用户分割为 2 部分，新老用户比例为 2∶8，在新用户与老用户中的数据进行了纵向分割，随机分成 A、B 2 个部分，并且通过改变 A、B 的分配比例 $P \in (0.1,0.9)$进行多组实验。评估时，将 A 中已有的新用户评分行为与相应的推荐作比较。

4.3　验证

第 1 个实验是通过变化 A、B 间分配比例 P 的取值来观察当联合多方数据训练模型时，取何值效果会比较理想。第 2 个实验中，将该算法和一个基于物品平均评分的无偏差推荐基线算法进行评估指标的比较。

1) 不同 B 机构比例的实验结果

令 P 代表 B 数据量在总数据量中的占比，当 P 越大时，代表 B 所拥有的信息越多。由于在实际场景中，两方机构所含有的物品特征不全是一致的，A 可能会与规模较大的机构合作，也可能会与规模较小的机构合作。

从表 1 中可以看出当 B 的占比越大，无论阈值 c 取 3 或 4(当用户评分>=阈值时，才算击中)，

对于 A 机构的冷启动推荐效果都更好, 这与直观相符合, 当外部提供的信息越多, 对自身的帮助会越大。对于 HR 及 NDCG, 也可以看到随着 B 的占比上升, 总体趋势也为上升。阈值 c 为 4 时召回率均大于阈值为 3, 这说明本文的方法对于高评分物品推荐的效果较好。

表 1　取不同 P 的实验结果

Table 1　Results change population distribution

P	c=3			c=4			命中率	归一化折现累计增益
	精确率	查全率	F_1-score	精确率	查全率	F_1-score		
0.1	0.952 9	0.090 6	0.165 5	0.802 9	0.103 2	0.183 0	0.331 3	0.158 0
0.2	0.962 8	0.113 6	0.203 2	0.829 5	0.132 1	0.227 9	0.365 7	0.174 2
0.3	0.967 4	0.138 7	0.242 6	0.846 3	0.163 7	0.274 4	0.388 2	0.183 8
0.4	0.972 0	0.169 1	0.288 0	0.851 0	0.200 4	0.324 4	0.430 0	0.210 6
0.5	0.972 8	0.203 2	0.336 1	0.853 4	0.239 6	0.374 2	0.423 7	0.208 4
0.6	0.972 1	0.253 2	0.401 8	0.846 4	0.295 9	0.438 5	0.426 4	0.207 4
0.7	0.970 1	0.320 2	0.481 5	0.837 4	0.371 9	0.515 1	0.459 9	0.220 8
0.8	0.965 7	0.456 3	0.619 7	0.823 2	0.517 8	0.635 7	0.450 5	0.218 1
0.9	0.942 9	0.771 6	0.848 7	0.758 5	0.830 0	0.821 4	0.493 4	0.253 2

2) 基准算法与本文算法的结果对比

为验证方法可行性, 采用一个基准算法来作对比。采用的基准算法为只取 A 方的数据对各个物品评分取平均, 并将其平均分排序最高的前 N 个对新用户进行无偏差的推荐。本文以 A、B 机构均占 50%, 即 P = 0.5 为例, 实验结果如表 2、3。

表 2　阈值击中评价

Table 2　Hit rate with different threshold

c	方法	精确率	查全率	F_1-score
3	本文算法	0.972 8	0.203 2	0.336 1
	基准算法	0.964 5	0.187 0	0.313 2
4	本文算法	0.853 4	0.239 6	0.374 2
	基准算法	0.854 6	0.222 8	0.353 4

从表 2 结果可以发现, 当阈值取 3 以及取 4 时, 本文算法均比基准算法有一定的提高。当阈值取 3, 可以看到无论是精确度还是查全率都有一定的提升, 其中 F_1 值较基准算法提升了约 7%。当阈值取 4 时, F_1 值较基准算法提升了约 6%。

表 3　留一法评价

Table 3　Leave-one-out evaluation

方法	命中率	归一化折现累计增益
本文算法	0.423 7	0.208 4
基准算法	0.376 5	0.190 0

从表 3 中可以看到留一法的评估结果, 相较基准算法, 本文算法的击中率 (HR) 有较大的提升, 约 12.5%, 但 NDCG 的提升效果较小, 约 9.6%。说明该算法对于推荐的击中效果有较大的提升, 而对于击中的位置提升的效果较小。

5　结束语

本文首先介绍了基于物品的协同过滤算法以及安全内积算法, 并针对新用户的冷启动问题, 提出了一种基于联邦学习的协同过滤冷启动解决算法。该算法针对某一方的新用户冷启动问题, 通过与其他方数据进行联合, 在不泄露信息的情况下进行相似矩阵的计算, 最终解决冷启动问题。实验结果表明, 本文提出的联邦学习冷启动方法在准确度均有一定的提升, 同时实验证明当联合规模较大的数据进行联合训练时, 对于本地的推荐效果会有较大的提升。该方法不仅提供了一种解决协同过滤冷启动的方法, 也在运用联邦学习解决冷启动的方向带来了一定的启发。

参考文献:

[1] ALBRECHT J P. How the GDPR will change the world[J]. European data protection law review, 2016, 2(3): 287–289.

[2] SHI E, CHAN T H H, RIEFFEL E, et al. Distributed private data analysis: lower bounds and practical constructions[J]. ACM transactions on algorithms, 2017, 13(4): 50.

[3] KIM S, KIM J, KOO D, et al. Efficient privacy-preserving matrix factorization via fully homomorphic encryption: extended Abstract[C]//Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security.

Xi'an, China, 2016: 617–628.

[4] 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. Hong Kong, China, 2001: 285–295.

[5] HERLOCKER J L, KONSTAN J A, BORCHERS A, et al. An algorithmic framework for performing collaborative filtering[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. California, Berkeley, USA, 1999: 230–237.

[6] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30–37.

[7] RENNIE J D M, SREBRO N. Fast maximum margin matrix factorization for collaborative prediction[C]//Proceedings of the 22nd International Conference on Machine Learning. New York, USA, 2005: 713–719.

[8] KOBSA A. Privacy-enhanced web personalization[M]. BRUSILOVSKY P, KOBSA A, NEJDL W. The Adaptive Web: Methods and Strategies of Web Personalization. Berlin: Springer, 2007: 628–670.

[9] JECKMANS A, TANG Qiang, HARTEL P. Privacy-preserving collaborative filtering based on horizontally partitioned dataset[C]//Proceedings of 2012 International Conference on Collaboration Technologies and Systems. Denver, USA, 2012.

[10] QIAO Yu, LI Lingjuan. Research on resolving strategies of the cold start problem of recommendation system [J]. Computer technology and development, 2018, 28(2): 83–87.

[11] YANG Qiang, LIU Yang, CHEN Tianjian, et al. Federated machine learning: concept and applications[J]. ACM transactions on intelligent systems and technology, 2019, 10(2): 12.

[12] 乔雨, 李玲娟. 推荐系统冷启动问题解决策略研究 [J]. 计算机技术与发展, 2018, 28(2): 83–87.
QIAO Yu, LI Lingjuan. Research on solution of solving cold start problem in recommender systems[J]. Computer technology and development, 2018, 28(2): 83–87.

[13] ZHAO Lingchen, NI Lihao, HU Shengshan, et al. InPrivate digging: enabling tree-based distributed data mining with differential privacy[C]//Proceedings of 2018 IEEE Conference on Computer Communications. Honolulu, USA, 2018: 2087–2095.

[14] SMITH V, CHIANG C K, SANJABI M, et al. Federated multi-task learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA, 2017: 4427–4437.

[15] HSU C C, YEH M Y, LIN Shude. A general framework for implicit and explicit social recommendation[J]. IEEE transactions on knowledge and data engineering, 2018, 30(12): 2228–2241.

[16] CHEN Shulong, PENG Yuxing. Matrix factorization for recommendation with explicit and implicit feedback[J]. Knowledge-based systems, 2018, 158: 109–117.

[17] JAWAHEER G, SZOMSZOR M, KOSTKOVA P. Comparison of implicit and explicit feedback from an online music recommendation service[C]//Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems. New York, USA, 2010: 47–51.

[18] PAZZANI M J, BILLSUS D. Content-based recommendation systems[M]. BRUSILOVSKY P, KOBSA A, NEJDL W. The Adaptive Web: Methods and Strategies of Web Personalization. Berlin, Heidelberg, Germany: Springer, 2007: 325–341.

[19] BURKE R. Knowledge-based recommender systems[J]. Encyclopedia of library and information systems, 2000, 69(S32): 175–186.

[20] WANG Jizhe, HUANG Pipei, ZHAO Huan, et al. Billion-scale commodity embedding for E-commerce recommendation in Alibaba[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, United Kingdom, 2018: 839–848.

[21] HWANGBO H, KIM Y S, CHA K J. Recommendation system development for fashion retail E-commerce[J]. Electronic commerce research and applications, 2018, 28: 94–101.

[22] HERLOCKER J L, KONSTAN J A, RIEDL J. Explaining collaborative filtering recommendations[C]//Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. Pennsylvania, Philadelphia, USA, 2000: 241–250.

[23] SUBRAMANIYASWAMY V, LOGESH R, CHANDRASHEKHAR M, et al. A personalised movie recommendation system based on collaborative filtering[J]. International journal of high performance computing and networking, 2017, 10(1/2): 54–63.

[24] ZHENG E, KONDO G Y, ZILORA S, et al. Tag-aware dynamic music recommendation[J]. Expert systems with applications, 2018, 106: 244–251.

[25] ZHENG Guanjie, ZHANG Fuzheng, ZHENG Zihan, et al. DRN: a deep reinforcement learning framework for news recommendation[C]//Proceedings of the 2018 World Wide Web Conference. Lyon, France, 2018: 167–176.

[26] COLOMO-PALACIOS R, GARCÍA-PEÑALVO F J, STANTCHEV V, et al. Towards a social and context-aware mobile recommendation system for tourism[J]. Pervasive and mobile computing, 2017, 38: 505–515.

[27] GENTRY C. A fully homomorphic encryption

scheme[D]. Palo Alto: Stanford University, 2009.
[28] CANNY J. Collaborative filtering with privacy via factor analysis[C]//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002: 238–245.
[29] CHEN Chaochao, ZHOU Jun, WU Bingzhe, et al. Practical privacy preserving POI recommendation[J]. ACM transactions on intelligent systems and technology, 2020, 11(5): 52.
[30] CHEN Chaochao, LIU Ziqi, ZHAO Peilin, et al. Privacy preserving point-of-interest recommendation using decentralized matrix factorization[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA, 2018: 257–264.
[31] ERKIN Z, BEYE M, VEUGEN T, et al. Efficiently computing private recommendations[C]//Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic, 2011: 5864–5867.
[32] KATARYA R, VERMA O P. Effective collaborative movie recommender system using asymmetric user similarity and matrix factorization[C]//Proceedings of 2016 International Conference on Computing, Communication and Automation. Noida, India, 2016: 71–75.
[33] HERLOCKER J L, KONSTAN J A, BORCHERS A, et al. An algorithmic framework for performing collaborative filtering[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. California, Berkeley, USA, 1999: 230–237.
[34] GASCON A, SCHOPPMANN P, BALLE B, et al. Secure linear regression on vertically partitioned datasets[J]. IACR cryptology ePrint archive, 2016, 2016: 892.
[35] HARPER F M, KONSTAN J A. The movielens datasets: history and context[J]. ACM transactions on interactive intelligent systems, 2015, 5(4): 19.
[36] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Nevada, Las Vegas, USA, 2008: 426–434.
[37] NIKOLAENKO V, WEINSBERG U, IOANNIDIS S, et al. Privacy-preserving ridge regression on hundreds of millions of records[C]//Proceedings of 2013 IEEE Symposium on Security and Privacy. Berkeley, USA, 2013: 334–348.

作者简介:

王健宗, 高级工程师, 博士, 主要研究方向为联邦学习算法、金融智能平台。主持国家重点研发计划基金项目 3 项、校企联合课题 2 项, 授权发明专利 100 余项。发表学术论文 50 余篇, 出版著作 3 部。

肖京, 教授级高级工程师, 博士, 主要研究方向为人工智能与大数据分析挖掘。国际授权专利 101 项, 授权国内发明专利 109 项。2019 年吴文俊人工智能科学技术奖“杰出贡献奖”获得者, 发表学术论文 130 余篇。

朱星华, 博士研究生, 主要研究方向为联邦学习、机器视觉算法。