

基于知识图谱的行为路径协同过滤推荐算法



陈源毅^{1,3} 冯文龙^{2,3} 黄梦醒^{2,3} 冯思玲^{2,3}

1 海南大学计算机与网络空间安全学院 海口 570228

2 海南大学信息与通信工程学院 海口 570228

3 海南大学南海海洋资源利用国家重点实验室 海口 570228

(thisismike@foxmail.com)

摘要 针对个性化推荐,常用的推荐算法有内容推荐、物品协同过滤(Item CF)和用户协同过滤(User CF),但是这些算法以及它们的改进算法大多偏向于关注用户的显性反馈(标签、评分等)或评分数据,缺少对多维度用户行为和行为顺序的利用,导致推荐准确率不够高及冷启动等问题。为了提高推荐精度,文中提出了一种基于知识图谱的行为路径协同过滤推荐算法(BR-CF)。首先根据用户行为数据,考虑行为顺序创建行为图谱(behavior graph)和行为路径(behavior route),然后采用向量化技术(Keras Tokenizer)将文本类型的路径向量化,最后计算多维度行为路径向量之间的相似度,对各维度分别进行路径协同过滤推荐。在此基础上,文中提出了两种 BR-CF 与 Item CF 相结合的改进算法。实验结果表明,在阿里天池数据集 UserBehavior 上, BR-CF 算法能够有效地在多个维度中进行推荐,实现数据的充分利用和推荐的多样性,并且此改进算法很好地提升了 Item CF 的推荐性能。

关键词: 推荐算法;行为顺序;行为图谱;行为路径;路径协同;多维度推荐

中图法分类号 TP391

Collaborative Filtering Recommendation Algorithm of Behavior Route Based on Knowledge Graph

CHEN Yuan-yi^{1,3}, FENG Wen-long^{2,3}, HUANG Meng-xing^{2,3} and FENG Si-ling^{2,3}

1 College of Computer Science and Cyberspace Security, Hainan University, Haikou 570228, China

2 College of Information Science & Technology, Hainan University, Haikou 570228, China

3 State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University, Haikou 570228, China

Abstract For personalized recommendation, common recommendation algorithms include content recommendation, Item CF and User CF. However, most of these algorithms and their improved algorithms tend to focus on users' explicit feedback (tags, ratings, etc.) or rating data, and lack the use of multi-dimensional user behavior and behavior order, resulting in low recommendation accuracy and cold start problems. In order to improve the recommendation accuracy, a collaborative filtering recommendation algorithm based on knowledge graph (BR-CF) is proposed. Firstly, according to the user behavior data, behavior graph and behavior route are created considering the behavior order, and then the vectorization technology (Keras Tokenizer) is used. Finally, the similarity between multi-dimensional behavior route vectors is calculated, and the route collaborative filtering recommendation is carried out for each dimension. On this basis, two improved algorithms combining BR-CF and Item CF are proposed. The experimental results show that the BR-CF algorithm can recommend effectively in multiple dimensions on the user behavior dataset of Ali Tianchi, realize the full utilization of data and the diversity of recommendation, and the improved algorithm can improve the recommendation performance of Item CF.

Keywords Recommendation algorithm, Behavior order, Behavior graph, Behavior route, Route coordination, Multi-dimensional recommendation

1 引言

互联网的飞速发展导致数据呈指数级增加,进而出现了

搜索引擎技术,使得用户可以在海量信息中搜索感兴趣的信息。随着搜索引擎的发展,用户的个性化需求越来越明显,于是出现了个性化推荐系统,其通过预测用户偏好进行推荐,帮

收稿日期:2020-10-01 返修日期:2021-01-19 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家重点研发计划项目(2018YFB1404400)

This work was supported by the National Key R & D Project(2018YFB1404400).

通信作者:冯文龙(fwfwl@163.com)

助用户快速决策和寻找信息^[1]。推荐系统的关键在于推荐算法,主要通过显性反馈和隐性反馈^[2]来获取用户偏好。常用的算法有内容推荐、物品协同(Item CF)、用户协同(User CF)及其改进算法,但是这些算法大多偏向于关注用户的显性反馈(文本、标签等),容易出现冷启动问题^[3-4]和马太效应^[5]。相比显性反馈,隐性反馈数据容易收集,例如点击和收藏等行为数据可以很好地解决冷启动问题和马太效应,提高了个性化 Web 文档结果排名的准确率^[6]。

在使用隐性反馈数据做推荐时,主要有 3 个研究方向:用户的近期行为和历史行为会对推荐结果产生影响^[6-9];行为数据的时间顺序会对推荐结果产生影响,即用户兴趣存在时变性^[10-12];在用户行为序列中,相邻行为之间存在着相似性和关联性^[13-15]。但是已有方法存在以下不足:仅使用用户评分数据而缺乏对多维行为数据的利用;无法应用在缺乏用户评分数据的场景中;只考虑了单方面的因素,未把这 3 个研究方向结合在一起。为了提高推荐精度,本文结合以上 3 个研究方向及其不足之处,提出了行为路径的概念和相应的改进算法。

本文综合考虑长短期行为、行为顺序和行为关联等因素,提出了行为路径的概念,定义行为路径为用户针对同一目标实体在长周期内产生的所有行为的有序集合,用于表示用户行为的演化发展过程。同一用户针对不同目标实体形成不同的行为路径,其中每条行为路径包含着用户的偏好和行为习惯信息,而行为习惯是用户长期形成的比较稳定的自动化行为方式^[9]。因此,挖掘出用户的所有行为路径,就可以大致得到用户的偏好和行为习惯信息。更进一步地,将行为路径按照行为类型分类,结合协同过滤的思想,通过计算不同类型的行为路径之间的相似度,找到相似度最高的行为路径,就可以预测用户最可能喜欢的、最可能发生当前类型行为的目标实体,从而实现推荐。

另一方面,在大数据环境下,数据呈现出多源异构的特征,难以融合、转化、创建关联关系及数据模型^[16],用户行为数据的处理和如何有效地挖掘行为路径存在着较大困难,而知识图谱技术可以很好地解决这一问题。知识图谱由谷歌公司于 2012 年提出,也称作知识地图,由节点和节点之间的关系组成。真实世界中的实体或概念都可以创建为节点,并用关系表示它们之间的关联^[17]。知识图谱能够从异构数据源中得到用户和项目的特征信息并进行融合^[18],丰富数据的语义信息,从而更好地推理用户数据。

结合行为路径和知识图谱,本文提出了一种基于知识图谱的行为路径协同过滤推荐算法(BR-CF),该算法根据用户长周期内的行为数据创建行为图谱和行为路径,通过计算多维度行为路径向量之间的相似度来进行路径协同过滤推荐,从而实现数据的充分利用和推荐的多样性。在此基础上,本文提出了两种 Item CF 与 BR-CF 相结合的改进算法。实验结果表明:1)在阿里天池数据集 UserBehavior 上,BR-CF 算法的推荐效果优于 Item CF;2)Item CF 与 BR-CF 相结合的两种改进算法在 3 个维度上的平均准确率均明显高于 Item CF;3)与最新的 Spark Hierarchical CF 和 Pheromone-based Algorithm 相比,本文提出的两种改进算法的推荐性能均更优。可

见,BR-CF 算法能够进行有效推荐,并很好地改进了 Item CF 算法。

2 相关工作

为了提高推荐系统的实时性和扩展性,Che 等^[19]提出了一种基于 Spark 的分层协同过滤推荐算法(Spark Hierarchical CF),利用用户偏好模型和聚类算法把用户按不同的偏好特征划分为不同的用户簇,之后针对不同的用户簇作协同过滤推荐,与基于 MapReduce 的 Item-based 协同过滤算法相比提高了推荐准确率,节约了运算时间,但是聚类的中心点难以确定。

目前关于用户行为路径的研究,主要包括基于强化学习的方法、基于统计分析的方法、基于活动轨迹聚类的方法和基于蚁群算法的方法 4 种。

为了改进协同过滤算法忽略项目属性的问题,Han 等^[20]提出了一种基于强化学习的动态个人推荐方法。首先从操作行为中挖掘用户的属性标签,然后根据操作路径(operate path)和召回路径(recall path)调整属性标签的奖惩模型,实现标签权重的动态调节,最后使用强化学习实现标签推荐,该方法的缺点是仅考虑了近期行为,没有充分利用行为路径隐含的用户偏好信息。

Li 等^[21]提出用户的体验感知能提升推荐效果,根据行为数据分析用户行为路径和创建用户体验感知模型,其行为路径被定义为吸引注意、表现兴趣、点击意愿、视频观看等连续性指标,能够较好地展现出用户的兴趣发展趋势,但是该方法基于统计分析,存在冷启动、数据稀疏的问题。

为了更好地预测用户行为,Syaekhoni 等^[22]将顾客在商店内的活动轨迹定义为购物路径,并使用 K-means 聚类算法分析用户的隐藏行为,但是聚类算法的重点在于初始中心点的选取,而该方法随机确定初始中心点,易导致结果不准确或偶然性误差。

考虑到用户访问和蚂蚁觅食的相似性,Wang 等^[23]引入蚁群算法来实现推荐,根据用户浏览行为和偏爱浏览路径算出转移概率,从而动态地进行推荐。但是该方法存在冷启动问题和马太效应,并且根据转移概率进行推荐存在偶然性误差。Xiong^[24]模仿蚁群算法中信息素的概念提出了商品信息素,并提出了 Pheromone-based Algorithm 算法,通过商品信息素将用户的浏览行为与购买行为相结合,继而针对用户的浏览轨迹对用户进行推荐,但是商品信息素的特性会导致马太效应。与本文方法类似的是,该算法定义浏览轨迹为商品之间的轨迹,而本文方法定义行为路径为行为之间的轨迹。

3 基于知识图谱的行为路径协同过滤推荐算法 BR-CF

本文提出了一种基于知识图谱的行为路径协同过滤推荐算法(BR-CF)。首先将用户行为数据中的 n 条行为初始化为 n 个节点 $nodes = (a_1, a_2, \dots, a_n)$,每个节点 a 存储该行为的属性,并在 Neo4j 中将这些节点构建成为行为知识图谱,简称行为图谱。其次,定义行为路径为用户针对同一目标实体在长周期内产生的所有行为的有序集合 $Route_{uid, iid} = (b_1, b_2, \dots, b_n)$,并在行为图谱中根据行为顺序进行构建,uid 和 iid 分别表示

用户 ID 和物品 ID。然后,导出行为路径进行向量化处理。最后,计算多维度行为路径之间的相似度,从而得到推荐结果。该算法的流程如图 1 所示。本节主要介绍该算法的具体步骤。

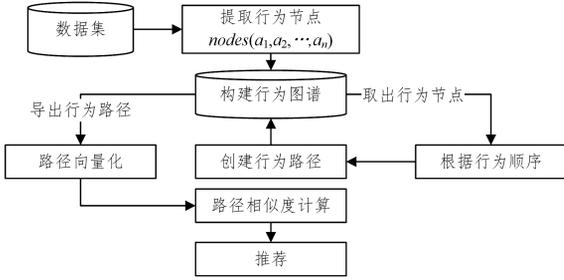


图 1 BR-CF 算法的流程

Fig. 1 Flowchart of BR-CF algorithm

3.1 构建行为图谱

以每条行为数据为一个节点,使用 Cypher 语言将每个行为的 m 个属性 (f_1, f_2, \dots, f_m) 封装为节点,存入 Neo4j 数据库,构建算法如式(1)所示:

$$\begin{aligned}
 &g = \text{Graph}(\text{host}, \text{http_port}, \text{user}, \text{password}) \\
 &\text{node} = \text{Node}('f_1' = f_1, 'f_2' = f_2, \dots, 'f_m' = f_m) \quad (1) \\
 &g.\text{create}(\text{node})
 \end{aligned}$$

其中, $\text{host}, \text{http_port}, \text{user}, \text{password}$ 分别代表主机地址、端口地址、数据库用户名和密码,用于连接 Neo4j 数据库。Node 将每条行为数据封装成一个节点, $\text{create}(\text{node})$ 命令将每个节点存入 Neo4j 数据库,初步构建的行为图谱如图 2 所示。

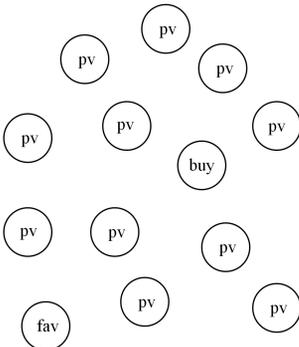


图 2 初步构建的行为图谱

Fig. 2 Initially constructed behavior graph

3.2 创建行为路径

首先,确定行为优先级。根据行为图谱中包含的 n 种行为类型,定义行为优先级 $(b_1, b_2, b_3, \dots, b_n)$, b 代表单个行为, b_1 的优先级最低, b_n 的优先级最高。

然后,定义行为路径。本文定义行为路径为用户针对同一目标实体在长周期内产生的所有行为的有序集合 $\text{Route}_{uid, iid} = (b_1, b_2, \dots, b_n)$, uid 和 iid 分别表示用户 ID 和物品 ID。

最后,创建行为路径。根据行为顺序,遍历行为图谱,对同一用户关于同一目标实体做出的连续行为创建“连续动作”关系,对于不构成路径的单节点则跳过,完成遍历后得到所有用户关于不同目标实体的 $\text{Route}_{uid, iid}$,具体步骤如图 3 所示。

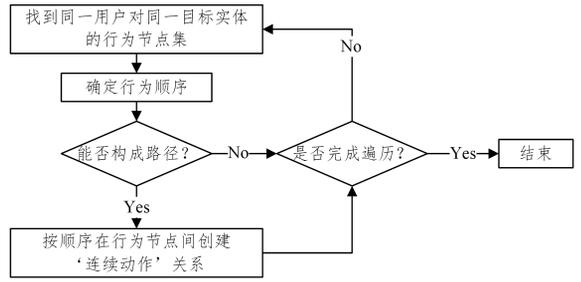


图 3 创建行为路径的步骤

Fig. 3 Steps to create a behavior route

创建完成的行为路径如图 4 所示。

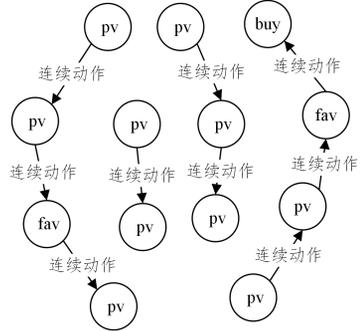


图 4 行为路径图

Fig. 4 Behavior route diagram

3.3 路径向量化

由于上一步中创建的行为路径的数据类型是文本类型,而进行相似度计算需要数字类型的数据,因此对路径进行向量化及对齐处理,具体步骤如图 5 所示。

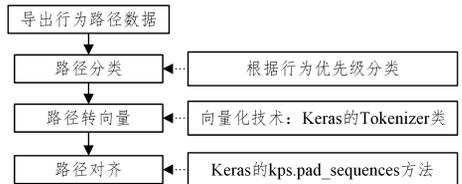


图 5 路径向量化流程图

Fig. 5 Flowchart of route vectorization

首先,将行为路径按照行为优先级分类,如表 1 所列。

表 1 路径分类表

Table 1 Route classification

分类	分类依据
b_1 类	路径中最高行为优先级为 b_1
b_2 类	路径中最高行为优先级为 b_2
...	...
b_n 类	路径中最高行为优先级为 b_n

其次,将每一类中的每一条路径向量化,从文本格式转为数字格式。本文实验利用 Keras 的 Tokenizer 类进行向量转换,向量映射表如表 2 所列。例如行为路径“ $b_1 - b_2 - b_4 - b_1 - b_1$ ”,转换后的向量为 $[1, 2, 4, 1, 1]$ 。

表 2 向量映射表

Table 2 Vector mapping

Behavior	mapping
b_1	1
b_2	2
...	...
b_n	n

最后,需要向量对齐。由于路径长度各不相同,各个向量也长度不一,这意味着向量的空间维度不统一,无法计算相似度。因此,需要计算所有行为路径中最大的路径长度,然后使用 Keras 的 `kps.pad_sequences` 方法令每条路径的空间维度与最大长度相同,补齐原理为补 0。数据变化过程如表 3 所列。

表 3 数据变化过程示例

Table 3 Example of data change process

路径状态	数值
初始行为路径集	{“ $b_1-b_2-b_4-b_1-b_1$ ”“ $b_1-b_1-b_1$ ”}
路径分类后	b_1 类: [“ $b_1-b_1-b_1$ ”],
	b_2 类: [],
	b_3 类: [],
	b_4 类: [“ $b_1-b_2-b_4-b_1-b_1$ ”]
路径转向量后	b_1 类: [[1,1,1]],
	b_2 类: [],
	b_3 类: [],
	b_4 类: [[1,2,4,1,1]]
路径对齐后	b_1 类: [[1,1,1,0,0]],
	b_2 类: [],
	b_3 类: [],
	b_4 类: [[1,2,4,1,1]]

3.4 路径相似度计算

这里用 A 类中的某路径分别与 B 类中的各路径的距离总和来度量路径相似度,如式(3)所示,总距离越小则相似度越大。

由表 3 中的 b_2 类、 b_3 类数组为空可知,路径分类中可能存在空类,直接计算相似度会出现对象为空的情况,因此需要分辨路径分类中哪些类是空类,最后决定哪种行为是可进行推荐的。

根据排列组合原理, n 种行为类型存在 $2^n - 1$ 种分类情况,例如 4 个分类共存在 15 种情况,如表 4 所列。其中,组合(b_1)表示除了 b_1 类其他类都为空,组合(b_2, b_3, b_4)表示 b_2, b_3 和 b_4 类不为空, b_1 类为空,以此类推。

表 4 路径类别组合表

Table 4 Route category combination

Possible combination of route categories	
1. (b_1)	9. (b_2, b_4)
2. (b_2)	10. (b_3, b_4)
3. (b_3)	11. (b_1, b_2, b_3)
4. (b_4)	12. (b_1, b_2, b_4)
5. (b_1, b_2)	13. (b_1, b_3, b_4)
6. (b_1, b_3)	14. (b_2, b_3, b_4)
7. (b_1, b_4)	15. (b_1, b_2, b_3, b_4)
8. (b_2, b_3)	

确定路径组合后,计算路径相似度。本文采用欧氏距离来计算路径的相似度, n 维空间的欧氏距离的计算式如式(2)所示:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

其次,分别计算 $class_1$ 类中每条路径与 $class_2$ 中所有路径的距离总和,如式(3)所示,其中 $class_1$ 和 $class_2$ 代表路径分类中的两个类。其他路径分类情况依次类推。

$$distances = \sum_{i=1}^m d(x, y_i), x \in class_1, y \in class_2 \quad (3)$$

最终,针对 b_n 类路径可得到对应总距离列表,如式(4)所示:

$$List_distances_b_n = [distances_1, distances_2, \dots], n \in [1, n] \quad (4)$$

如组合为(b_2, b_3, b_4),则计算可得到总距离列表,如式(5)所示:

$$List_distances_b_2 = [distances_1, distances_2, \dots, distances_m]$$

$$List_distances_b_3 = [distances_1, distances_2, \dots, distances_n] \quad (5)$$

其中, m 和 n 分别为 b_2 和 b_3 类中的路径数量, $List_distances_b_2$ 表示 b_2 类中每条路径与 b_3 类中所有路径的总距离列表, $List_distances_b_3$ 表示 b_3 类中每条路径与 b_4 类中所有路径的总距离列表。

3.5 多维度推荐

本文定义从多维度($d_1, d_2, d_3, \dots, d_{n-1}$)实现推荐, d 代表可能发生 b 行为的推荐集, d_1 对应可能发生 b_2 的推荐集, d_{n-1} 对应可能发生 b_n 的推荐集,分别用 $possible_b_n$ 表示。 d 的值为 $\min(List_distances)$ 对应的目标实体的集合,即总距离最小的路径对应的目标实体的集合。得到的推荐结果如式(6)所示:

$$\{d_1, d_2, \dots, d_{n-1}\} = \{“possible_b_2”: [\dots], “possible_b_3”: [\dots], \dots, “possible_b_n”: [\dots]\} \quad (6)$$

根据式(5)中的示例结果可知,存在两个推荐维度 d_2 和 d_3 ,返回推荐结果如式(7)所示,分别表示本文算法所推荐的用户可能发生行为 b_3 和 b_4 的目标实体集合,值为目标实体的 ID 或其他唯一标识。

$$\{d_3, d_4\} = \{“possible_b_3”: [2733371, 332733], “possible_b_4”: [124451]\} \quad (7)$$

4 基于 Item CF 与 BR-CF 结合的改进算法

物品协同过滤推荐算法简称 Item CF,用于给用户推荐与他们之前喜欢的目标实体相似的目标实体。

为了进一步分析 BR-CF 的推荐性能,本节分别将 Item CF 和 BR-CF 进行级联和混合,提出了两种基于 Item CF 与 BR-CF 结合的改进算法。

4.1 Item CF 与 BR-CF 的级联推荐算法

Item CF 与 BR-CF 的级联推荐算法简称为 Combine Item CF and BR-CF。该算法在物品协同的推荐结果中进一步计算出存在行为路径且与验证集中所有目标实体的行为路径总距离最小的目标实体集合,流程图如图 6 所示。

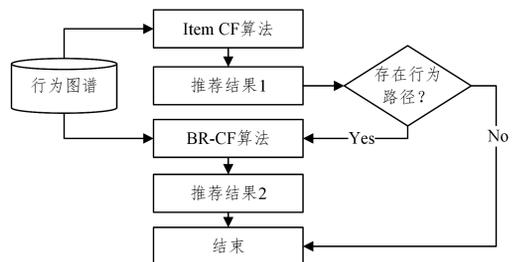


图 6 Combine Item CF and BR-CF 算法的流程图

Fig. 6 Flowchart of Combine Item CF and BR-CF algorithm

4.2 Item CF 与 BR-CF 的混合推荐算法

Item CF 与 BR-CF 的混合推荐算法简称为 Mix Item CF and BR-CF。该算法将二者的推荐结果合并,得到最终的推荐结果,流程图如图 7 所示。

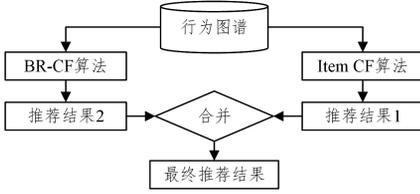


图 7 Mix Item CF and BR-CF 算法的流程图

Fig. 7 Flowchart of Mix Item CF and BR-CF algorithm

5 实验

为了验证 BR-CF 算法和 Item CF 结合 BR-CF 的改进算法的有效性,本节在公开数据集上进行测试和对比实验,并对实验结果进行评价。

5.1 实验环境

本文实验环境的配置如下:Windows 7 系统,AMD A8-5550M,CPU: 3.1 GHz,4 GB 内存,500 GB 硬盘,PyCharm Community Edition 2019.3.3 x64,Neo4j-community_windows-x64_3_3_1,Python 3.6.9,Py2neo 4.3.0,Keras2.2.5,Numpy 1.18.1。

5.2 数据集介绍

本文实验在阿里云天池的公开数据集 UserBehavior^[25]上进行验证,其中包含淘宝用户的真实行为数据。类似的公开数据集还有电子商务网站访问者的 Retailrocket,其中包含用户的行为数据。经过实验,由于其中同一用户的数据很少,无法形成有效的行为路径,进而无法应用 BR-CF 算法,因此本文略去该数据集的介绍及实验部分。

UserBehavior 数据集包含了 2017 年 11 月 25 日至 2017 年 12 月 3 日之间约 100 万淘宝随机用户的行为(点击、购买、加购、收藏)。每条行为数据由用户 ID、商品 ID、商品类目 ID、行为类型和时间戳组成。统计信息如表 5 所列。

表 5 UserBehavior 数据集的统计信息

Table 5 Statistical information of UserBehavior dataset

数据集	用户数	商品数量	商品类目数量	行为总数
UserBehavior	987 994	4 162 024	9 439	100 150 807

5.3 对比实验

基于 Spark 的分层协同过滤推荐算法^[19]简称 Spark Hierarchical CF,基于信息素的推荐算法^[24]简称 Pheromone-based Algorithm。在相同数据集和评价指标下,分别进行 Item CF、BR-CF、Item CF 改进算法、Spark Hierarchical CF 和 Pheromone-based Algorithm 的对比实验。由于用户协同算法要求用户之间的交集必须达到 3 及以上,经实验发现 UserBehavior 数据集中满足该条件的用户为 0,说明用户协同对数据有一定要求,导致推荐结果为空,因此本文略去用户协同算法的对比实验,但是显然 BR-CF 算法实现的结果可以

改善用户协同算法因数据要求导致的推荐结果。对比实验如表 6 所列。

表 6 对比实验

Table 6 Comparative experiments

序号	算法
方法 1	Item CF
方法 2	BR-CF
方法 3	Combine Item CF and BR-CF
方法 4	Mix Item CF and BR-CF
方法 5	Spark Hierarchical CF
方法 6	Pheromone-based Algorithm

实验从 UserBehavior 数据集中随机抽取 20 059 条连续的用户行为作为实验数据,并将其 80% 划分为训练数据,将其 20% 划分为验证数据,以进行推荐和验证。为了避免实验结果的偶然性误差,实验结果采用 K-折交叉验证方法进行验证,K 值按照数据比例取值为 5。

5.4 实验评价

推荐算法常用的评价指标有准确率和召回率,由于 BR-CF 是从行为数据中挖掘出最可能是用户喜爱的目标实体,而没有对某个目标实体是否是用户所喜爱的进行判断,无法计算召回率,因此本文仅使用准确率来评判推荐结果。

准确率是相对于预测结果而言的,它表示预测为正的样本中有多少是对的,如式(8)所示:

$$precision = \frac{TP}{TP + FP} \quad (8)$$

其中,TP 表示正确预测数量,FP 表示错误预测数量。

由 3.5 节可知,BR-CF 算法从多个维度进行推荐,分别是 $(d_1, d_2, d_3, \dots, d_{n-1})$,因此存在 $n-1$ 个推荐准确率,分别对应一种推荐,如式(9)所示:

$$precision_{d_1} = \frac{d_1 \text{ 中预测正确的数量}}{d_1 \text{ 中目标实体的总数}}$$

$$precision_{d_2} = \frac{d_2 \text{ 中预测正确的数量}}{d_2 \text{ 中目标实体的总数}}$$

$$\dots$$

$$precision_{d_{n-1}} = \frac{d_{n-1} \text{ 中预测正确的数量}}{d_{n-1} \text{ 中目标实体的总数}}$$

5.5 实验结果

5.5.1 方法 1—方法 4 的准确率散点图对比

实验中根据 20059 条行为数据(来自 181 个用户)成功创建了 2622 条行为路径,数据统计如表 7 所列。

表 7 实验数据统计

Table 7 Statistics of experimental data

类别	数量
行为信息	20 059
行为路径	2 622
用户总数	181
K 折交叉验证的 K 值	5

根据 5.2 节中数据集 UserBehavior 的定义,将数据中的 4 种行为初始化,得到行为优先级(pv, fav, cart, buy),然后从 fav, cart 和 buy 3 个维度进行推荐,推荐结果如式(10)所示:

$$\{d_1, d_2, d_3\} = \{\text{possible_fav}, \text{possible_cart}, \text{possible_buy}\} \quad (10)$$

由式(9)可知,3个维度需要计算3个准确率($precision_{d1}, precision_{d2}, precision_{d3}$)。

通过实验发现,当数据中不存在优先级更低的路径类 b_n 或者用户数据不足以形成行为路径时,Item CF 和 BR-CF 都无法在维度 d_n 进行推荐。因此针对方法1—方法4进行可推荐人数的统计,如表8所列。由表8可知,可进行 d_1 推荐的人数较少,而式(10)中 BR-CF 将多个维度的推荐结果合并,从多个维度进行推荐,可以避免推荐为空的情况,同

时提高了推荐的多样性。

表8 在不同维度中可进行推荐的人数

Table 8 Number of people who can be recommended in different dimensions

算法	可进行 d_1 推荐的人数	可进行 d_2 推荐的人数	可进行 d_3 推荐的人数
方法1	28	84	27
方法2	23	68	28
方法3	28	88	27
方法4	40	120	53

方法1—方法4关于3个维度(d_1, d_2, d_3)的推荐准确率散点图分别如图8—图10所示。

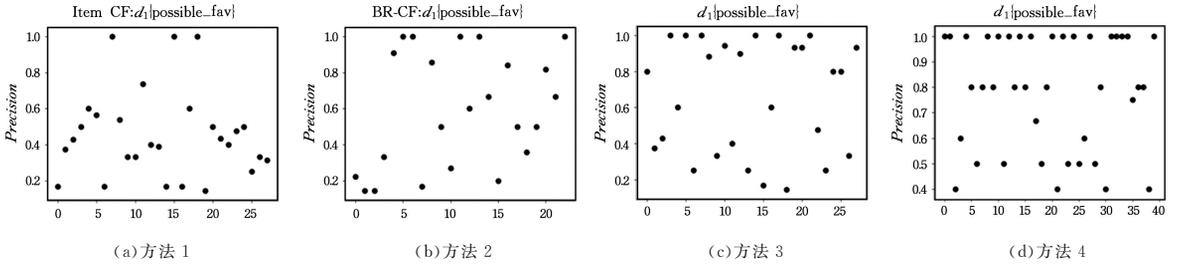


图8 d_1 的推荐准确率散点图对比

Fig. 8 Comparison of precision scatter plot of d_1

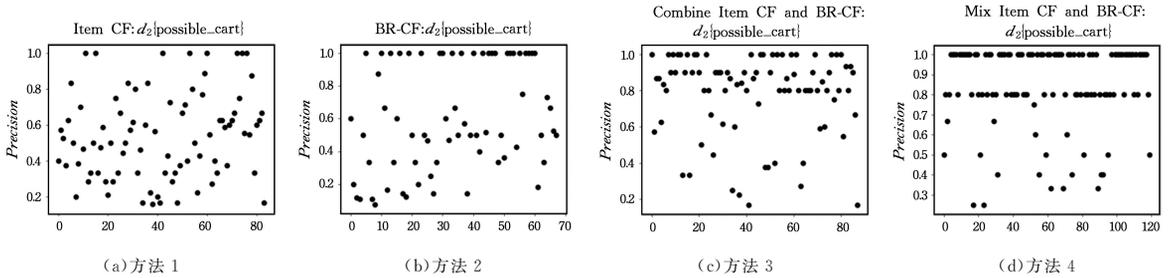


图9 d_2 的推荐准确率散点图对比

Fig. 9 Comparison of precision scatter plot of d_2

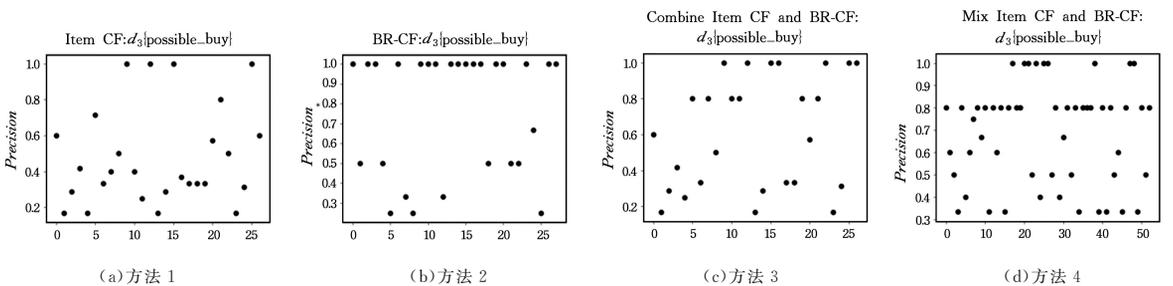


图10 d_3 的推荐准确率散点图对比

Fig. 10 Comparison of precision scatter plot of d_3

由图8中 d_1 的推荐结果可以看出,表现最佳的是方法4,准确率主要集中在0.4及以上,说明了方法4的有效性。由图9中 d_2 的推荐结果可以看出,方法3和方法4优于方法1和方法2,准确率多数较高,说明了方法3和方法4的有效性。由图10中 d_3 的推荐结果可以看出,方法2、方法3和方法4优于方法1,准确率多数更高,其中方法4的表现最好,准确率高且多数聚集在0.8附近,说明了方法4的有效性。

5.5.2 方法1—方法6的平均准确率对比

由于方法5和方法6的实验数据集和评价指标^[19,24]与

本文一致,因此本文将引用其评价结果进行对比实验。在可进行推荐的用户中,分别计算方法1—方法4在3个维度(d_1, d_2, d_3)上的推荐平均准确率,并与方法5和方法6的评价结果作对比,结果如图11所示。

由图11可以得出以下结论:

1)方法2在3个维度上的平均准确率分别为59.5%,60.8%,77.1%,分别比Item CF高出13.7%,6.7%,28.9%,说明BR-CF的推荐性能优于Item CF,同时,BR-CF明显优于方法6。

2)方法3在3个维度上的平均准确率分别为66.2%,78.3%,61.2%,分别比Item CF提高了20.4%,24.2%,13.0%;方法4在3个维度上的平均准确率分别为78.5%,84.7%,68.6%,分别比Item CF提高了32.7%,30.6%,20.4%,说明BR-CF与Item CF相结合的改进算法能够显著提升Item CF的推荐性能。

3)方法5在3个维度上的平均准确率均高于方法2,而在维度 d_1 和 d_2 上的平均准确率低于方法4,说明BR-CF结合Item-CF的改进算法的推荐性能优于方法5。

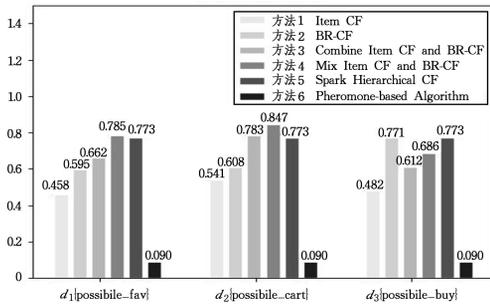


图 11 平均准确率柱状图

Fig. 11 Histogram of average precision

结束语 针对已有推荐算法大多偏向于关注用户的显性反馈或评分数据,缺少对多维度用户行为和用户行为顺序的利用,导致推荐准确率不够高及冷启动等问题,本文提出了一种基于知识图谱的行为路径协同过滤推荐算法(BR-CF)。本文使用20059条真实行为数据,根据用户行为的顺序成功创建了2266条行为路径,丰富了数据的语义信息,并且充分利用用户行为数据,从3个维度进行推荐,提高了数据的利用率和推荐的多样性。特别地,本文提出了行为路径的概念,在行为与行为之间创建关联,能够很好地表示用户行为的演化发展过程,从而更好地预测用户的喜好。在此基础上,提出了两种BR-CF与Item CF相结合的改进算法。最后,在阿里天池数据集UserBehavior上验证了该算法的有效性。该算法良好地改进了Item CF算法,并且改进算法与最新的Spark Hierarchical CF和Pheromone-based Algorithm算法相比,推荐性能更优。

由于该算法需要在用户对同一目标实体做出的行为集合中创建行为路径,因此需要一定量的行为数据作为支撑。当用户对某一目标实体操作过少时则难以创建行为路径以及实现推荐,即数据稀疏性问题,该问题有待进一步解决。未来的工作可以加入用户特征,以进一步提高推荐精度。

参考文献

[1] CHANG L, ZHANG W T, GU T L, et al. Review of recommendation systems based on knowledge graph[J]. CAAI Transactions on Intelligent Systems, 2019, 14(2): 207-216.

[2] CAI W L, ZHENG J B, PAN W K, et al. Neighborhood-Enhanced Transfer Learning for One-Class Collaborative Filtering[J]. Neurocomputing, 2019, 341(14): 80-87.

[3] PENG F, LU X, MA C, et al. Multi-level preference regression for cold-start recommendations[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(7): 1117-1130.

[4] CORTES D. Cold-start recommendations in Collective Matrix Factorization[J]. arXiv:1809.03666, 2018.

[5] WANG C D, DENG Z H, LAI J H, et al. Serendipitous Recommendation in E-Commerce Using Innovator-Based Collaborative Filtering[J]. IEEE Transactions on Cybernetics, 2018, 49(7): 2678-2692.

[6] CAI F, WANG S, RIJKE M D. Behavior-based personalization in web search[J]. Journal of the Association for Information Science & Technology, 2017, 68(4): 855-868.

[7] HUANG X Y, XIONG L Y, LI Q D. Personalized news recommendation technology based on improved collaborative filtering algorithm[J]. Journal of Sichuan University (Natural Science Edition), 2018, 55(1): 49-55.

[8] HUANG D X. Research on user dynamic interest model in recommendation system[D]. Guangzhou: South China University of Technology, 2018.

[9] SHEN D D, WANG H T, JIANG Y, et al. A next recommendation algorithm based on knowledge map and short-term preference[J]. Minicomputer System, 2020(4): 849-854.

[10] CHEN X, XU H, ZHANG Y, et al. Sequential Recommendation with User Memory Networks[C]// The Eleventh ACM International Conference. ACM, 2018.

[11] LI W J. Research on time aware recommendation algorithm[D]. Chengdu: University of Electronic Science and Technology, 2017.

[12] KANG J Y, SU F J. Long and short interest recommendation model based on generative countermeasure network[J]. Computer Technology and Development, 2020, 30(6): 35-39.

[13] ZHANG Z P, SHEN X Y. Research on User Behavior Recommendation Method Based on Deep Learning[J]. Computer Engineering and Applications, 2019, 55(4): 142-147, 158.

[14] GUI Z Y, ZHANG Y M, LI W W. Research on Learning Resource Recommendation Algorithm Based on behavior sequence analysis[J]. Computer Application Research, 2020, 37(7): 1979-1982.

[15] DUAN W Q. Prediction of online purchasing behavior based on user behavior sequence[D]. Nanchang: Jiangxi University of Finance and Economics, 2019.

[16] CHAI L, XU H F, LUO Z M, et al. A multi-source heterogeneous data analytic method for future price fluctuation prediction[J]. Neurocomputing, 2020, 418: 11-20.

[17] HUANG M X, LI M L, HAN H R. Research on entity recognition and knowledge mapping based on electronic medical record[J]. Computer Application Research, 2019, 36(12): 3735-3739.

[18] SHAO B L, LI X J, BIAN G Q. A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph[J]. Expert Systems With Applications, 2020: 113764.

[19] CHE J Q, XIE H W. Hierarchical collaborative filtering recommendation algorithm based on spark[J]. Electronic technology application, 2015, 41(9): 135-138.

[20] HAN D, SHEN X, GAN T, et al. A Dynamic Individual Recom-

mendation Method Based on Reinforcement Learning[C]//International Symposium on Parallel Architecture, Algorithm and Programming. 2017.

- [21] LI C, HU W L. Exploration on experience model of recommendation system—taking video recommendation as an example[J]. Industrial Design Research, 2018(5):81-85.
- [22] SYAEKHONI M A, LEE C, KEON Y S. Analyzing customer behavior from shopping path data using operation edit distance [J]. Applied Intelligence, 2016, 48:1912-1932.
- [23] WANG H, XIA Z Q. Research on recommendation algorithm based on ant colony algorithm and browsing path [J]. China Science and Technology Information, 2009(7):103-104.
- [24] XIONG Y R. Recommendation algorithm based on user behavior trajectory [D]. Chengdu: University of Electronic Science and Technology of China, 2013.

- [25] LEI M L. Research on shopping behavior based on Alibaba big data[J]. Internet of Things Technology, 2016, 6(5):57-60.



CHEN Yuan-yi, born in 1995, postgraduate. His main research interests include data mining and big data analysis.



FENG Wen-long, born in 1968, Ph.D., professor, Ph.D supervisor, is a professional member of China Computer Federation. His main research interests include big data and smart services.