

大数据中基于混合协同过滤的动态用户个性化推荐

刘珊珊

(广州华南商贸职业学院, 广东 广州 510650)

摘要: 为了提高大数据中动态用户个性化推荐的准确性和效率, 采用基于混合协同过滤的方法来完成用户感兴趣数据的筛选, 从而实现个性化推荐。先将用户数据及项目数据通过协同过滤算法来完成建模并评分, 然后结合XGBoost模型的树形结构和正则学习的特点进行预测评分, 接着将两种算法混合来求解最优目标函数, 得到候选的推荐数据集。最后通过实例仿真, 混合算法精确度高, 在大数据平台有较强的适用性。

关键词: 大数据; 协同过滤; XGBoost; 个性化推荐; 准确率

中图分类号: TP399 **文献标识码:** A

Personalized Recommendation for Dynamic Users Based on Hybrid Collaborative Filtering in Big Data

LIU Shanshan

(South China Vocational College of Commerce and Trade, Guangzhou 510650, China)

Abstract: In order to improve the accuracy and efficiency of personalized recommendation for dynamic users in big data, a hybrid collaborative filtering method is used to filter the data of interest to achieve personalized recommendation. Firstly, user data and project data are modeled and scored by collaborative filtering algorithm, then predicted and scored by combining the tree structure of XGBoost model and the characteristics of regular learning. Then the two algorithms are mixed to solve the optimal objective function, and candidate recommended data set is obtained. Experiments show that the hybrid collaborative filtering recommendation algorithm has high accuracy and strong applicability in big data platforms.

Keywords: big data; collaborative filtering; XGBoost; personalized recommendation; accuracy

1 引言(Introduction)

大数据作为互联网高速发展的产物, 其服务器可以来自于不同用户端的数据进行融合, 并有效利用这些数据进行数据挖掘与数据分析, 以便根据用户使用习惯为不同用户提供差异化的推荐服务。对于大数据平台的使用者来说, 能够动态地接受自己感兴趣内容, 可以有效节省在平台进行资源搜索的时间, 提高用户访问大数据平台的舒适度, 也反映了大数据平台的智能性。当前这种动态用户个性化推荐的服务在很多平台上都得到了有效应用。比如社交平台的个性化好

友推荐、资讯平台的个性化新闻推荐, 电子商务平台的商品个性化推荐, 服务类平台的个性化服务推荐等^[1,2]。

2 大数据平台的动态用户个性化推荐研究现状 (Research status of personalized recommendation for dynamic users in big data platform)

当前, 大数据平台的动态用户个性化推荐方法主要有两种模式, 一种是以大数据提供的项目为参照点, 以项目作为数据主关键字, 将所有访问过该项目的用户主要数据记录到该项目的数据表中; 另外一种是以访问大数据平台的用户为

参照点,以用户作为数据作为主关键字,将该用户访问所有项目的数据记录到该用户的数据表中。两种方法各有优劣,前者的优点是数据结构较简单,耗费的存储资源少,后者的优点是用户分类更清晰,推荐服务更细化。在大数据平台的个性化推荐服务中,根据用户量的需求和用户对推荐服务的精细化程度,两种推荐模式各有优势。

当前采用协同过滤算法完成动态用户的个性化推荐服务研究较多。比如Lim等人将该算法运用于旅游推荐系统^[3],根据用户的偏好提供有效的旅游路径和旅游景点,并以用户形象标签作为有效推荐的辅助^[4]。王重仁等人利用XGBoost算法对互联网客户流失情况的预测分析^[5],这两种算法在预测推荐系统中均有应用,但将两种算法结合的相关研究较少,本文将两种算法混合,将XGBoost算法的效率和协同过滤算法的有效性相结合,完成大数据平台的动态用户个性化推荐。

3 大数据中基于混合协同过滤的动态用户个性化推荐(Personalized recommendation for dynamic users based on hybrid collaborative filtering in big data)

3.1 混合协同过滤

为了提高协同过滤结果的有效性及精准度,提高个性化推荐系统的适用性,将传统的协同过滤算法的评分值与XGBoost推荐算法的评分值进行权重分配,根据两者共同的评分结果高低作为大数据中动态用户个性化推荐的有限次序。XGBoost算法是一种基于树形结构的学习系统。相比常用的高级算法,比如蚁群算法、鱼群算法等,XGBoost算法具有很好的伸缩性和扩展性^[5],在大数据的分布式计算中,面对数千万级别的用户数据量对应关系,在相似度计算的耗时环节,较好地解决了内存限制和算法速度慢等问题。下面对用户数据进行评分建模,具体方法如下:

首先,建立用户user和项目item列表,每一个用户u对应一条数据库记录,这条记录中存储了用户u在大数据平台中访问项目的情况,在这条记录中,可以挖掘出用户u对某些项目i的访问次数,根据访问次数来了解用户偏好,在大数据平台中,当两个用户访问项目i的次数达到设定的阈值,就判定为这两个用户为相邻用户,在大数据平台中,将所有项目根据相似度进行分类,根据用户在上一时间段内访问项目i的情况,动态地为下一时间段用户访问平台推荐类似于项目i的内容。

推荐系统中用户u对项目i的预测评分是根据用户u的k个相邻用户对项目i的评分求均值得到的^[6,7],若用户u的相邻用户对项目i未曾有过任何评分,则将该用户剔除相邻用户。

$$P_{u,i} = \frac{1}{k} \sum_{v \in N_u \cap S(i)} M_{v,i} \quad (1)$$

其中, $S(i)$ 表示所有访问过项目i的用户, $M_{v,i}$ 表示用户v对项目i的预测评分。

$$P_{u,i} = \frac{\sum_{v \in N_u \cap S(i)} \text{sim}(u,v) M_{v,i}}{\sum_{v \in N_u \cap S(i)} |\text{sim}(u,v)|} \quad (2)$$

其中, $\text{sim}(u,v)$ 表示目标用户u和相邻用户集合 N_u 中用户v的相似度^[8],通过相似度计算方法得到。

$$P_{u,i} = \bar{M}_u \cdot \frac{\sum_{v \in N_u \cap S(i)} \text{sim}(u,v) \cdot (M_{v,i} - \bar{M}_v)}{\sum_{v \in N_u \cap S(i)} |\text{sim}(u,v)|} \quad (3)$$

$$\text{sim}(i,j) = \frac{i \cdot j}{\|i\| \|j\|} \quad (4)$$

$\bar{M}_u$ 和 $\bar{M}_v$ 分别表示用户u和v的评分均值,计算方法如公式(5)所示。

$$\bar{M}_u = \frac{1}{N(I(u))} \sum_{i \in I_u} M_{u,i} \quad (5)$$

取前N项预测评分较高的项目推荐给用户^[9,10]。设协同过滤算法推荐为 T_1 ,采用XGBoost模型的推荐算法为 T_2 ,则混合推荐算法T的评分结果计算方法为:

$$T = aT_1 + (1-a)T_2 \quad (6)$$

设用户u对项目i的评价,协同过滤算法权重为 α_{ui} ,XGBoost模型的推荐算法权重为 β_{ui} ,设干扰为e,则有:

$$P_{ui} = \alpha_{ui} M_{ui} + \beta_{ui} M'_{ui} + e \quad (7)$$

由 α_{ui} 和 β_{ui} 组成的数据集合是稀疏集合,大部分数据值为0,在进行算法求解数据拟合过程中,可以将 α_{ui} 和 β_{ui} 写为 α_u 和 β_u 。

则式(7)可写为式(8):

$$P_{ui} = \alpha_u M_{ui} + \beta_u M'_{ui} + e \quad (8)$$

根据式(6),可知 $\alpha_u + \beta_u = 1$ 。

最后求解动态用户个性化推荐评分最高值,变为求解实际评分值与预测评分均值的最小差值,计算方法如公式(9)所示。

$$\min_{\alpha_u, \beta_u} = \sum_{R \in S} (M_{u,j} - \bar{M}_u)^2 \quad (9)$$

衡量推荐算法的标准主要有四个,即,准确率(Precision)、召回率(Recall)、平均绝对误差(MAE)和均方根误差(RMSE)^[11-13]。其计算方法分别如公式(10)—公式(13)所示。

$$\text{Precision} = \frac{|S_u \cap S'_u|}{|S'_u|} \quad (10)$$

$$\text{Recall} = \frac{|S_u \cap S'_u|}{|S'_u|} \quad (11)$$

其中,公式(9)和公式(10)中 S_u 表示采用混合算法评分预测值最

高的集合， S'_u 表示在训练样本里对于用户 u 评分最高的项目集合。

$$MAE = \frac{1}{|count(S)|} \sum_{R \in S} |R_{u,j} - M_{u,j}| \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{R \in S} (R_{u,j} - M_{u,j})^2}{count(S)}} \tag{13}$$

其中，公式(12)和公式(13)中 $R_{u,j}$ 表示用户 u 对项目 j 的实际评分值， $M_{u,j}$ 表示采用推荐算法后用户 u 对项目 j 的评分值， $count(S)$ 表示用户评分集合的总个数。

3.2 实例仿真

为了验证本文算法的性能，采用Matlab对推荐算法进行实例仿真。实验仿真平台硬件采用Intel Core i5处理器，内存为8GB，操作系统为Windows 7旗舰版，Matlab版本为R2010b。

实验数据样本集合相邻用户数量最大为50个， S'_u 的个数为100，用户评分集合总数为1000个，项目总数为10000个，为了验证混合算法的有效性，分别将协同过滤算法和基于XGBoost模型的推荐算法及两种算法的混合算法进行性能比较，比较它们的Precesion、Recall和MAE，考虑到MAE和RMSE有类似之处，在此只对MAE进行仿真。

从图1可以看出，随着相邻用户数量的增加，准确率逐渐上升，表明算法对大数据的适用程度高，在海量数据中能找到对于目标用户评分高的项目集合，且有较高的准确度。三种算法中基于XGBoost模型的推荐算法准确率相比较低，本文算法的准确率最高，当相邻用户数量达到15个时，准确率迅速攀升，当相邻用户数量达到47个时，推荐准确率达到了90%以上。具体准确率指标见表1所示。

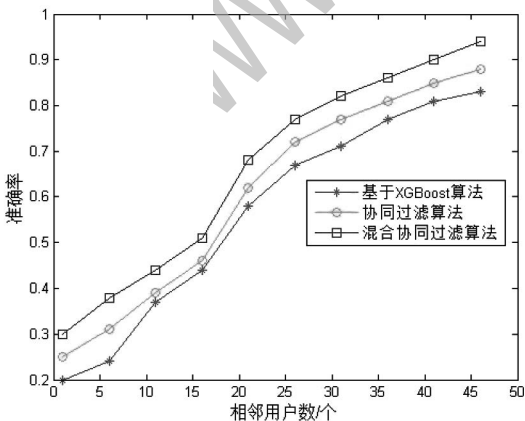


图1 准确率比较

Fig.1 Accuracy comparison

从表1可以看出，在相邻用户数量够多的情况下，三种算法均能取得较好的准确率数值，对于大数据平台的个性化推荐服务均有一定的应用价值，特别是混合协同过滤算法，实验中的准确率最高达到了94%；当相邻用户数量较少时，三种推荐算法准确率都较差，这表明三种算法的实现均是基于大数据平台下大数据量的条件下训练而来的，若某用户在大数据平台下的关联数据少，这无法完成用户个性化数据的有效推荐，这也符合数据挖掘的特点，当用户相关特征数据较少时，难于挖掘出针对该用户有价值的信息，更加无法根据用户偏好和习惯实现有效推荐。

表1 三种算法的准确率统计表

Tab.1 Statistical table of the accuracy of the three algorithms

相邻用户数量	准确率		
	XGBoost	协同过滤	混合算法
2	0.2	0.25	0.30
7	0.24	0.31	0.38
12	0.37	0.39	0.44
17	0.44	0.46	0.51
22	0.58	0.62	0.68
27	0.67	0.72	0.77
32	0.71	0.77	0.82
37	0.77	0.81	0.86
42	0.81	0.85	0.90
47	0.83	0.88	0.94

从图2可以看出，随着相邻用户数量的增加，召回率逐渐增加，表明算法对大数据用户的个性化推荐是有效的，三种算法比较，本文算法的性能突出，在相邻用户数量达到45时，推荐的有效性能达到75%以上，准确率和召回率相互制约，因此在算法训练时，选择的相邻用户数量应适当，保证两个性能能够均衡。大部分情况下，协同过滤算法的召回率性能要优于XGBoost算法，当相邻用户数量为5时，基于XGBoost算法相比与协同过滤算法性能更优，当相邻用户数量为30时，两者召回率相同，这表明两种算法在召回率方面，优劣并不明显。

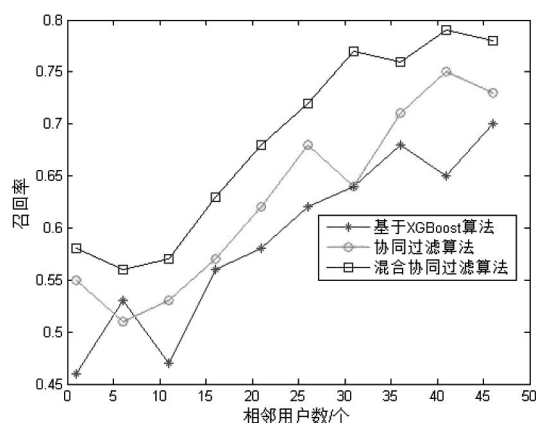


图2 召回率比较

Fig.2 Comparison of recall rate

从图3可以看出,随着相邻用户数量的增加,MAE逐渐减小,算法的预测评分值与实际评分制更接近,在相邻用户数为5—15个,MAE迅速下降,当相邻用户数量达到47时,混合协同过滤算法的MAE下降至0.4以下。

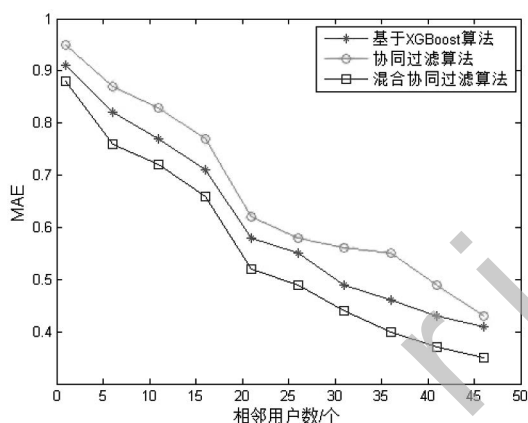


图3 MAE比较

Fig.3 MAE comparison

4 结论(Conclusion)

大数据平台的个性化推荐服务增强了用户使用体验,用户可以高效地获取平台有效数据,避免在海量数据里去搜索和寻找。这种个性化服务的良好体验是建立在精准的推荐算法基础上的,若推荐算法适用性不强,给用户推荐的服务与用户不相关,就会给用户造成垃圾数据负担,降低用户的体验感。本文将XGBoost算法和协同过滤算法相结合,提高了大数据平台个性化推荐的性能。经实验证明,本文算法在准确率和召回率等方面,表现出色,具有较强的推广应用价值。

参考文献(References)

[1] 陈杰,刘学军,李斌,等.一种基于用户动态兴趣和社交网络的微博推荐方法[J].电子学报,2017,45(4):898-905.

[2] 仲兆满,管燕,胡云,等.基于背景和内容的微博用户兴趣挖掘[J].软件学报,2017,28(2):278-291.

[3] Lim K H, Chan J, Leckie C, et al. Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency[J]. Knowledge & Information Systems, 2017, 54(2): 1-32.

[4] Jing Z, Ying Y, Qi T, et al. Personalized Social Image Recommendation Method Based on User-Image-Tag Model[J]. IEEE Transactions on Multimedia, 2017, 19(11): 2439-2449.

[5] 王重仁,韩冬梅.基于社交网络分析和XGBoost算法的互联网客户流失预测研究[J].微型机与应用,2017,36(23):58-61.

[6] Alhamid M F, Rawashdeh M, Dong H, et al. Exploring Latent Preferences for Context-Aware Personalized Recommendation Systems[J]. IEEE Transactions on Human-Machine Systems, 2017, 46(4): 615-623.

[7] 姜宇,张大方,刁祖龙.基于点击流的用户矩阵模型相似度个性化推荐[J].计算机工程,2018,44(1):219-225.

[8] Dash S, Behera R N. A Particle Swarm Optimization based Hybrid Recommendation System[J]. International Journal of Knowledge Discovery in Bioinformatics, 2017, 6(2): 1-10.

[9] Tarus J K, Niu Z, Yousif A. A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining[J]. Future Generation Computer Systems, 2017(72): 37-48.

[10] 罗娜.针对云平台协同推荐的近邻项目最优临界点优化[J].科技通报,2017(12):171-174.

[11] 陈南平.一种改进的个性化协同推荐算法研究[J].软件导刊,2017,16(3):45-47.

[12] 袁正午,陈然.基于多层次混合相似度的协同过滤推荐算法[J].计算机应用,2018,38(3):633-638.

[13] 张昊,纪宏超,张红宇.XGBoost算法在电子商务商品推荐中的应用[J].物联网技术,2017(72):102-104.

作者简介:

刘珊珊(1981-),女,硕士,讲师.研究领域:软件开发,大数据.