文章编号:1000-2367(2020)04-0007-05 DOI:10.16366/j.cnki.1000-2367.2020.04.002

融合儿童成长信息的协同过滤推荐算法

刘行兵,刘孝飞,司思,张震

(河南师范大学 计算机与信息工程学院,河南 新乡 453007)

摘 要:在互联网母婴领域中,由于育婴网络自身的特殊性,推荐算法不仅与用户以及项目的信息有关还与儿童的数据信息有关,而传统的用户相似度计算并未考虑儿童的数据信息.针对此问题,重新定义用户相似度计算方法,将儿童的数据信息通过加权融合的方法融入用户相似度计算中,并提出一种融合儿童成长信息的协同过滤算法,实验结果表明,该算法的准确率与召回率都优于传统算法,推荐系统的推荐质量也有所提高.

关键词:协同过滤;用户相似度;母婴

中图分类号:TP391 **文献标志码**:A

随着互联网与信息技术的飞速发展,各个行业向互联网转型,并取得较大的进展.在众多的行业中,母婴领域凭借近些年国家政策以及家庭对育婴的重视,更是得到前所未有的繁荣.作为父母的人们可以通过互联网来了解儿童成长状况,寻找志同道合的朋友或育婴达人,共同探讨或学习育婴经验.互联网母婴的发展为用户提供了极大的便利,但随着网络信息的增加,母婴用户将不可避免地面对大量的垃圾信息和无意义数据,即所谓的信息过载问题.特别是随着育龄妇女越来越年轻化,互联网已成为大多数父母生活的一部分,他们也更倾向于使用互联网来帮助自己了解孩子的成长状况,得到更加科学可靠且符合自己需求的信息,因此信息过载问题也就更加突出.

协同过滤技术被视为解决信息过载问题最有效的方式之一[1],但随着网络中信息资源的增加,协同过滤技术仍有着巨大的挑战,如数据稀疏问题、系统的可扩展性等[2],为此,研究者们提出各种解决方案.如文献[3]提出了一种多级协同过滤推荐方法,提高推荐效率,改善用户整体体验.文献[4]首先对目标用户与新用户之间的社会关系进行预测,然后利用社会关系计算相似度,最后给用户进行推荐,这一种方法并未解决矩阵稀疏问题.也有研究将时间衰减函数用于预处理用户的评级,不仅可以有效地解决数据稀疏和新项目的问题,而且算法的推荐精度也明显提高[5].文献[6]将用户的属性信息与互动信息充分融合,提出一种融合用户属性和互动信息的推荐算法,提高推荐质量.文献[7]提出一种改进的基于元路径的加权异构信息网络协同过滤算法使得推荐结果在准确性方面更加准确.文献[8]认为人口统计属性在一定程度上所表现出来的兴趣爱好可能更加符合实际情况,因此充分挖掘人口统计属性,并将其应用于推荐算法中,以提高推荐的准确率.文献[9]将用户相似度与信任度融合,不仅可提高推荐精度还能有效的解决冷启动问题.

在互联网母婴中,用户间的相似度不仅仅与用户对项目的评分以及用户自身属性有关,还与儿童的属性信息以及成长信息有关,当两个用户的小孩属性信息以及成长信息相似时,可以认为这两个用户会因小孩之间的相似性而产生共同兴趣,具有较高的相似性.例如当小孩成长未达标时,目标用户则倾向于与具有同样情况的用户交流,因此目标用户小孩的属性信息应该与其邻居用户的小孩信息具有一定的相似性,所以将儿童属性以及成长的相似性与传统的用户相似性计算方法相结合,能够提高推荐系统的推荐精度.

通常利用用户的属性信息来提高推荐精度主要有两种方法,加权融合求最终相似度和将用户属性作为推荐算法的一部分或一个步骤再结合其他方法来得到新的算法.例如文献[10]通过加权融合方法,并综合考

收稿日期:2019-04-19;**修回日期**:2019-05-26.

基金项目:国家自然科学基金(U1804164,61902112);河南省教育厅自然科学项目(17A520039).

作者简介(通信作者):刘行兵(1973—),男,河南信阳人,河南师范大学副教授,博士,研究方向为无线网络、机器学习,E-mail:shanhe18@126.com.

虑用户的社会行为、背景信息以及评分信息,提出一种基于用户动态社会行为和背景信息的协同过滤算法.文献[11]则先对用户进行聚类,再根据用户属性信息以及聚类结果构建决策树预测.

本文则根据互联网母婴的自身特点,使用加权融合的方法,将儿童的信息相似度融合到传统的用户相似度计算中,从而得到最终的用户相似度,然后根据计算得到的最终相似度得到用户的邻居集,最后根据邻居集来实现对目标用户的推荐.

1 融合儿童成长信息的协同过滤推荐算法

1.1 用户相似度计算

用户的属性信息在某种程度上揭示了用户的兴趣偏好,充分利用用户的属性信息可以提高用户相似度的准确性.用户的属性信息一般包括两种数据,数值型和文本型.数值型数据如年龄、身高、收入等,文本型数据如签名、昵称等.对于用户的属性相似度计算通常先计算两个用户之间每一个属性的相似度,然后对不同的属性设置合适的权重值,最终得到用户属性的总相似度.这里对于用户属性信息相似度的计算采用文献[6]中用户相似度计算公式,公式如(1)式所示:

$$D=\sum_{i=1}^{n} w_i d_i, \tag{1}$$

其中,w 为每一个数值型属性的权重值,n 表示共有 n 个数值型属性.为方便计算采用(2) 式进行转换,使其值域为[0,1],其中 $S(u,v)_{\text{user}}$ 表示用户相似度的值.

$$S(u,v)_{\text{user}}=\frac{1}{1+D}. \tag{2}$$

根据数据集中实际情况,用户有 3 维属性,包括用户位置、用户年龄、用户性别,在实验中需将数据进行预处理,具体处理方式如下.(1)性别信息,1:男性;0:女性.(2)年龄信息,1:(18,23]岁; 2:(23,30]岁; 3:(30,35]岁;4:(35,42]岁;5:(42,48]岁;6:48 岁以上.(3)位置信息,将其位置信息根据其对应的邮政编码转化为数字编码.

1.2 儿童相似度计算

儿童的属性信息包含性别、年龄、出生身长、出生体重等,若将儿童的属性作为向量的元素,则属性信息就可以表示为 $U_n=(I_1,I_2,\cdots,I_n)$,利用儿童信息中的性别、年龄等属性的特征向量建立儿童特征属性矩阵,如表 1 所示.

表 1 儿童特征属性表

Tab.1 Child characteristics attribute table

特征属性	I_1	I_2	I_3	…	I_m
U_1	A_{11}	A_{12}	A_{13}	…	A_{1m}
U_2	A_{21}	A_{22}	A_{23}	…	A_{2m}
⋮	⋮	⋮	⋮		⋮
U_n	A_{n1}	A_{n2}	A_{n3}	…	A_{nm}

本文根据爬虫所得到的数据,选择儿童年龄、儿童性别以及儿童成长状况三个属性计算儿童之间的属性相似度.其中成长状况则是根据中国 0～18 岁儿童青少年身高、体质量的标准化生长曲线来判定[12],当其身高在预测值的上下一个标准差内时,成长状况为正常用数值 1 表示,当处于两个标准差时用数值 2 表示,当处于三个标准差时用数值 3 表示,当处于大于三个标准差时用数值 4 表示.对于年龄属性,从数据集中可了解到儿童的年龄分布为 0～14,考虑到儿童年龄越接近其相似度越高,因此可将年龄进行量化,具体量化方法如表 2 所示.

表 2 儿童年龄量化方法表

Tab.2 Child age quantification method table

具体年龄	(0,1]	(1,4]	(4,7]	(7,10]	(10,13]	13<
量化后的值	1	2	3	4	5	6

对于儿童的性别来说,仍与用户性别表示相同,将男孩表示为 1,女孩表示为 0.用户间的相似性计算方法,常用的有皮尔森相关系数、欧几里得距离、余弦相似度等[13],本文则采用欧几里得距离来计算儿童的相似度,计算公式如下:

$$S(u,v)_{\text{child}}=\frac{1}{1+\sqrt{\sum(r_u-r_v)^2}},\tag{3}$$

其中,r_u,r_v 表示儿童 u 和儿童 v 的特征向量,$S(u,v)_{\text{child}}$ 表示儿童相似度的值.欧几里得的取值范围为[0,1],其值越小相似度越大,因此,在(3)式中 $S(u,v)_{\text{child}}$ 值越大,儿童相似度越高.

1.3　用户相似度与儿童相似度的融合

用户最终相似度是将用户相似度和儿童相似度通过加权融合的方法所得到.在互联网母婴中,用户安全意识的提高,为防止隐私泄露,很多用户的信息并不完善,并且还存在大量虚假信息,因此用户信息以及儿童信息都会产生一定的缺失.为了将信息缺失问题降到最低,提高算法的推荐性能,本文通过在训练集中进行训练来选择合适的权重,此时的权重更加符合实际.若假设 α 与 β 分别为用户和儿童信息在相似度计算中所占的比值,则使用加权融合方式得到用户的总相似度计算公式,如(4)式所示.

$$S(u,v)=\alpha\cdot S(u,v)_{\text{user}}+\beta\cdot S(u,v)_{\text{child}},\tag{4}$$

其中,α 与 β 满足 $\alpha+\beta=1$.$S(u,v)$ 为用户的总相似度的值,其值越大,用户 u 与 v 之间的相似度越大.

1.4　算法描述

融合儿童成长信息的协同过滤推荐算法首先将儿童成长信息与用户属性信息加权融合计算用户的总相似度,相似度值越大说明用户之间越相似,然后根据最终相似度选择用户的邻居集,邻居集的选择最终也将影响算法的准确度,因为用户数据的稀疏程度差异较大,若对每个用户都选择相同的邻居数则存在不合理性,因此选择一个阈值来克服这个缺点,当用户的相似度值大于这个阈值,就将用户加入到候选集,并将候选集中的用户按相似度大小降序排列,最后使用 Top-k 法为用户实现推荐,具体算法如下:

输入:目标用户 U_u,推荐个数 K.

输出:给目标用户 U_u 推荐的其他用户.

Begin:

1)用户和儿童相似度的计算

利用(2)式和(3)式分别计算用户和儿童的相似度.

2)最终相似度的计算

利用(4)式计算用户总相似度.

3)产生邻居用户

根据步骤 2)中的用户总相似度,选取相似度值大于 0.5 的用户集,作为目标用户 u 的邻居用户集合 N.

4)产生推荐集

将步骤 3)中的邻居用户集合 N 中的用户按照相似度降序排列,选取前 K 个用户给目标用户进行推荐.

End

2　实验结果与分析

2.1　数据集

本文在育儿网(http://blog.ci123.com/)上进行网络爬虫获取数据,共获取 12 017 条有效数据,包括用户位置、用户性别、用户年龄、儿童性别、儿童年龄等,然后对其进行预处理,实验中,随机选取 20%作为测试集,80%作为训练集.

2.2　评价指标

由于推荐结果为好友故评分预测无法实现,因此为确保实验结果可信度,取推荐总数 K 的初始值为 10,以 10 的幅度逐步递增,主要从综合准确率、召回率度量值来进行结果的评价.综合准确率和召回率计算公式如下:$P=M/I$,$R=M/L$,其中,P 为准确率,R 为召回率,M 为推荐出的已成为好友的用户数量,I 表示推荐的好友总数,L 则表示测试集中用户好友总数.

2.3　结果与分析

本文提出一种融合儿童成长信息的协同过滤算法,在计算用户间相似度时不再只考虑用户的属性特征

等因素,而是根据互联网母婴的特殊属性,将儿童的数据信息融合到传统的用户相似性度量中去,然后,通过比较本文算法与基于用户的协同过滤算法的准确率和召回率,分析本算法的推荐性能.

首先,在训练集中,通过实验分析(4)式中的权重值,权重值的确定需要经过多次实验,但由于 $\alpha+\beta=1$,故在实验中只需确定 α 的值即可.为保证权重值的有效性,这里选择 $K=10$、20、30、40、50,分析不同 K 值下 α 和 β 的最佳取值.实验时,设 α 的初始值设为 0,终值为 1,增长幅度为 0.1,从而实现对 α 取不同值的效果,观察 α 在不同取值情况下推荐系统的准确率,结果如图 1 所示.

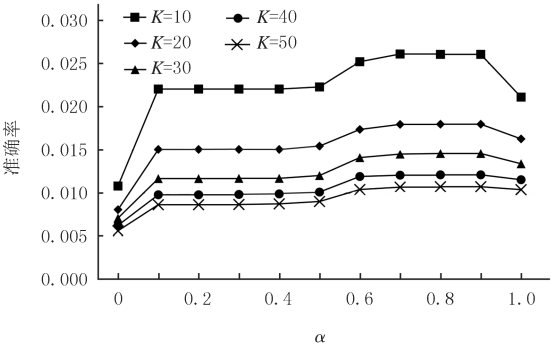

图1 α 在不同的取值情况下推荐系统的准确率

Fig.1 α Accuracy of the recommended system under different values

观察图 1 可看到,算法的准确性总体呈现先升后降的趋势,当 $\alpha=0.7$,$\beta=0.3$ 时,推荐系统综合准确率在不同的 K 值下均达到最高,即用户信息占比 0.7,儿童信息占比 0.3 时推荐系统的综合准确率最高.

为验证本推荐算法的有效性和准确性,本文对推荐结果进行统计分析,取 $\alpha=0.7$,$\beta=0.3$,在测试集中将本文推荐算法与经典推荐算法(即基于用户的协同过滤推荐算法)和只考虑儿童特征的推荐方法的结果进行对比,为确保实验结果可信度,设推荐总数 K 的初始值为 10,以 10 的幅度逐步增加,观察在不同的 K 值下,三种算法的召回率和准确率,分别如图 2 和图 3 所示.

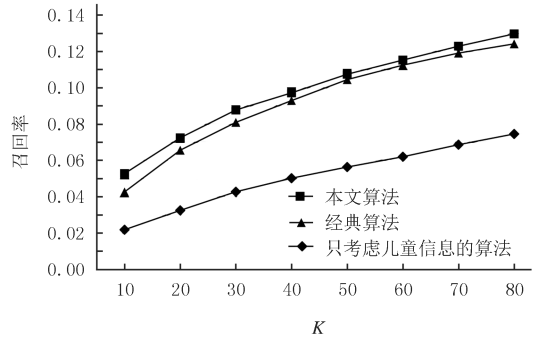

图2 各算法在不同K值下的召回率

Fig.2 Recall rate of each algorithm at different K values

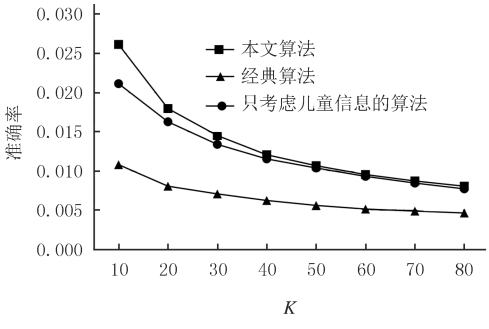

图3 各算法在不同K值下的准确率

Fig.3 Accuracy of each algorithm at different K values

在图 2 和 3 中可看到随着推荐项目 K 越大,算法准确率逐渐降低、召回率逐渐增大,符合实际情况;并且还可看出本文推荐算法的综合准确率和召回率测量值均高于经典推荐算法(基于用户的协同过滤推荐算法)和只考虑儿童信息的推荐方法,表明本文融合儿童数据信息的推荐算法的推荐效果要好,这说明通过本文的算法能够改进协同过滤算法的性能.此外,从图 3 还可以发现,三种算法准确率都较低,这是由于实验数据集是从网上爬虫所得,获取信息技术有限,用户某些较隐秘的属性信息无法得到,造成了用户以及儿童的数据信息不完善,计算相似度时的准确性降低.

3 结束语

本文针对传统的协同过滤算法仅考虑用户和项目并未考虑儿童数据信息的问题,提出一种融合儿童成长信息的协同过滤算法.该方法从互联网母婴领域的自身特点入手,通过充分挖掘互联网母婴领域中的儿童信息,达到利用儿童信息来提高推荐效果的目的.实验证明,融合儿童成长信息的协同过滤推荐算法提高了推荐质量,在准确率与召回率上都优于传统协同过滤推荐算法.但受限于信息获取技术,无法获取更全面的用户信息,导致整体准确率较低,目前,正在与育儿网取得合作,以便拿到更加全面真实的数据来对算法进行进一步的验证.

参 考 文 献

[1] 王晓东，时俊雅，李淳，等.学习资源精准推荐模型及应用研究[J].河南师范大学学报(自然科学版)，2019，47(1)：26-32.
WANG X D，SHI J Y，LI C，et al.Accurate recommendation model and application of learning resources[J].Journal of Henan Normal University(Natural Science Edition)，2019，47(1)：26-32.

[2] WEI J，HE J，CHEN K，et al.Collaborative filtering and deep learning based recommendation system for cold start items[J].Expert Systems with Applications，2017，69：29-39.

[3] POLATIDIS N，GGORGIADIS C K.A multi-level collaborative filtering method that improves recommendations[J].Expert Systems with Applications，2018，48：100-110.

[4] NAYAK R，ZHANG M，CHEN L.A Social Matching System for an Online Dating Network：A Preliminary Study[C]// The 10th IEEE International Conference on Data Mining Workshops.Sydney：IEEE Press，2010：352-357.

[5] LIU X J.An improved clustering-based collaborative filtering recommendation algorithm[J].Cluster Computing，2017，20(2)：1281-1288.

[6] 荣辉桂，火生旭，胡春华，等.基于用户相似度的协同过滤推荐算法[J].通信学报，2014，35(2)：16-24.
RONG H G，HUO S X，HU C H，et al.User similarity-based collaborative filtering recommendation algorithm[J].Journal on Communications，2014，35(2)：16-24.

[7] 张海霞，吕振，张传亭，等.一种引入加权异构信息的改进协同过滤推荐算法[J].电子科技大学学报，2018，47(1)：112-116.
ZHANG H X，LYU Z，ZHANG C T，et al.An Improved Collaborative Filtering Recommendation Algorithm with Weighted Heterogeneous Information[J].Journal of University of Electronic Science and Technology of China，2018，47(1)：112-116.

[8] 杨超，艾聪聪，蒋斌，等.一种融合人口统计属性的协同过滤算法[J].小型微型计算机系统，2015，36(4)：782-786.
YANG C，AI C C，JIANG B，et al.Demographic Attribute-based Collaborative Filtering Algorithm[J].Journal of Chinese Computer Systems，2015，36(4)：782-786.

[9] 徐毅，叶卫根，戴鑫，等.融合用户信任度与相似度的推荐算法研究[J].小型微型计算机系统，2018，39(1)：78-83.
XU Y，YE W G，DAI X，et al.Recommendation Algorithm Incorporating Users Trust and Users Similarity[J].Journal of Chinese Computer Systems，2018，39(1)：78-83.

[10] 蒋胜，王忠群，修宇，等.基于动态社会行为和用户背景的协同推荐方法[J].计算机科学，2015，42(3)：252-255.
JIANG S，WANG Z C，XIU Y，et al.Collaborative Filtering Recommendation Method Based on Dynamic Social Behavior and Users'Background Information[J].Computer Science，2015，42(3)：252-255.

[11] SUN D T，LI C，LUO Z G.A content-enhanced approach for cold-start problem in collaborative filtering[C].2011 2nd International Conference on Artificial Intelligence，Management Science and Electronic Commerce(AIMSEC).DengLeng：IEEE Press，2011：4501-4504.

[12] 李辉，季成叶，宗心南，等.中国 0～18 岁儿童、青少年身高、体重的标准化生长曲线[J].中华儿科杂志，2009，47(7)：487-492.
LI H，JI C Y，ZONG X N，et al.Height and weight standardized growth charts for Chinese children and adolescents aged 0 to 18 years[J].CHINESE JOURNAL OF PEDIATRICS，2009，47(7)：487-492.

[13] 韩勇，宁连举，郑小林，等.基于社交信息和物品曝光度的矩阵分解推荐[J].浙江大学学报(工学版)，2019，53(1)：89-98.
HAN Y，NING L J，ZHENG X L，et al.Matrix factorization recommendation based on social information and item exposure[J].Journal of Zhejiang University(Engineering Science)，2019，53(1)：89-98.

Collaborative filtering recommendation algorithm integrating children's growth information

Liu Xingbing，Liu Xiaofei，Si Si，Zhang Zhen

(College of Computer and Information Engineering，Henan Normal University，Xinxiang 453007，China)

Abstract：In the Internet maternal and child field，due to the particularity of the baby-care network itself，the recommendation algorithm is not only related to the user and the project information but also to the child's data information，while the traditional user similarity calculation does not consider the child's data information.Aiming at solving this problem，the user similarity calculation method is redefined，and the children's data information is integrated into the user similarity calculation by weighted fusion method.A collaborative filtering algorithm that integrates children's growth information is proposed.Experimental results show that both the accuracy rate and the recall rate are superior to the traditional algorithms，and the recommended quality of the recommended system is also improved.

Keywords：collaborative filtering；user similarity；mother and baby

[责任编校　陈留院　赵晓华]