混合协同过滤算法在推荐系统中的应用

沈 鹏,李 涛
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要: 推荐系统主要由两个方法组成,即基于内容属性相似的推荐和基于协同过滤算法的推荐,这两种方法可以提供有意义的推荐。其中基于内容属性相似的推荐只是单纯地依赖物品之间的属性相似来构建推荐关系;而协同过滤算法推荐作为推荐系统领域的经典,它不会去研究物品的本身属性,正如名字描述的一样,严重依靠于用户的行为及其周边用户的协同行为。文中使用了一种改进的混合方法,充分考虑和利用了协同过滤算法和内容属性过滤的优点。讨论的算法与该领域以前的方法是不同的,因为它包括一个新颖的方法来找到两个事物之间的相似内容。包含了一个分析用以证明这个新的方法,并且阐述了它是怎样提供实用性的推荐的。与其他两种常用的方法进行比较,即纯协同过滤(Pure CF)和奇异值分解(SVD),结果表明,该方法经过现有的用户和目标数据的测试,产生了有所改进的结果。

关键词: 推荐系统;协同过滤算法;内容属性相似;纯粹的协同过滤;奇异值分解
中图分类号: TP302 **文献标识码:** A **文章编号:** 1673-629X(2019)03-0069-03
doi:10.3969/j.issn.1673-629X.2019.03.014

Application of Hybrid Collaborative Filtering Algorithm in Recommendation System

SHEN Peng,LI Tao
(School of Communication and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract: The recommendation system is mainly composed of two methods,namely,the recommendation based on the similarity of content attributes and based on a collaborative filtering algorithm,which can help in providing meaningful recommendations. Among them,the former is simply relying on attribute similarity between items to build recommendation relationships,while the latter is recommended as a classic of the field of recommendation systems,not studying the attributes of items itself,just as its name described,relying heavily on the users' behavior and the collaborative behavior of surrounding users. In this article,we use an improved hybrid approach that fully considers and exploits the advantages of collaborative filtering algorithms and content attribute filtering. The algorithm discussed in this article is different from the previous method in this field because it includes a novel method to find similar content between two items. An analysis is contained to demonstrate this new method,discussing how it provides practical recommendations. Compared with the other two commonly used methods,Pure CF and SVD,it shows that the method in this paper results in improved results by passing the test of existing users and target data.

Key words: recommendation system;collaborative filtering algorithm;similarity of content attributes;Pure collaborative filtering;singular value decomposition

0 引 言

推荐系统是向用户推荐或建议适当的事物的机器软件。推荐系统主要由三个重要阶段组成,分别是目标数据收集,相似性判定和预测计算[1]。另外,市场上的推荐系统[2]主要基于三大方法[3]。基于内容[4]的方法充分利用物品的甚至是用户的内容(属性)。而文中方法使用了题材和标签。因此,使用这种方法可以发现一部电影的内容与用户喜欢的其他电影的内容之间的相似性[5]。为了预测目标用户的偏好,协同过滤也考虑了目标用户的近邻,用来发现邻居和目标用户

收稿日期: 2018-04-01 **修回日期:** 2018-08-02 **网络出版时间:** 2018-12-19
基金项目: 国家自然科学基金(61572260)
作者简介: 沈 鹏(1993-),男,硕士研究生,研究方向为大数据存储与计算;李 涛,硕士,副教授,研究方向为信号与信息处理在新型网络中的应用。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181219.1511.032.html

之间的相似性,以便选择最相似的邻居并且将他们的评分和偏好推荐给目标用户[6]。因此,用户的偏好推荐将取决于在活跃用户的邻居中存在的其他用户。此外,协同过滤的领域依赖特性可能使其易受稀疏性和冷启动的影响。这种类型的推荐系统可以分为基于记忆的,基于模型的[7]以及两者的混合[8]。由于协同过滤在很大程度上取决于用户的评分,因此,如果域中的用户数量与事物相对比较低,则可能导致冷启动[9]。文中提出的混合方法是基于内容属性和协同过滤算法的[10]。在后面的章节中,将讨论混合协同过滤算法如何优于纯协同过滤算法[11]和基于内容属性的算法[12]。

1 相关工作

梅尔维尔等提出过一种内容提升的混合协同过滤算法,将内容和协同过滤算法结合起来提供推荐。用户-事物评分矩阵的稀疏度为97.4%,伪评级矩阵是使用了借鉴用户画像的基于内容的过滤算法计算出来的。内容提升的混合协同过滤算法中使用的朴素贝叶斯文本分类器对不同电影的内容进行比较和分类,并通过协同过滤方式生成的伪评分矩阵进行预测。对于两部电影之间的相似性,作者使用了Pearson相关性[13]进行度量。而文中方法修改了基于内容的算法部分,使用了一个简单的比较器,而不是根据朴素贝叶斯进行文本匹配,它可以比较和匹配在Movielens数据集上测试的两部电影的标签和题材,并与SVD和纯CF进行比较。此外还确定,与以前的模型相比,用户项目评分矩阵中的初始稀疏度较高。

2 数据集描述

为了测试修改后的混合内容推荐方法,使用了推荐系统的标配实验数据—Movielens数据集[14]。数据集中包含用户为特定电影提供的个人评级。该数据集中总共包含100 004个评分,这些评分由671个用户给出,针对9 125部电影,评分范围从0到5。电影的题材总数为20。用户和电影组成了671* 9 125的用户-事物评分矩阵,其中行表示用户,列表示电影。Movielens数据集包含以下属性: userId,movieId,标题,评分,标签和题材。紧接着数据集被过滤,用文中算法对稀疏度为98.36%的用户评分矩阵进行了测试。数据集中最初提供的评分数量为100 004。在这些评分中,将2 000个评分分开,用于之后测试文中算法的准确性。分离这些评级后,留下了训练数据集,构成了98 004个评分,其稀疏度为98.3%。现对于进一步的读数设置,从稀疏度增加的方向随机删除训练数据集中一定比例的评分,并用对预测评分进行测试。通过这种方式,总共采集了六次读数,其中的稀疏度从98.3%到99.8%不等。

3 提出的方法

文中提出的算法考虑了数据集中指定的标签和题材,并且对基于内容的预测,应用了一组匹配比较器。该比较器返回两部电影之间公共属性的数量。这里的属性一词是指标签和题材。对于每部特定的电影,标签和题材都合并到一个集合中。这给了每部电影庞大的内容,而更多的内容则可以产生更好的预测。获取一组通用属性后,计算每部电影的权重。

一旦将权重分配给每个组,则它们将被用于使用先前比较的额定电影来提供未评级电影的评级。首先在该方法中,分配给用户的每部电影所打上的标签都要使用,并且将其转换为单个列表。每部电影的题材都附加到相同的标签列表中。该最终列表被称为特定电影的属性。将为每部有效影片设置的属性与数据集中每部其他影片的属性集进行比较,并将成功匹配的对象分配到一组。该组的长度用于预测评分,预测的公式如下所示:

$$R = M^* \ (\mathrm{Hr}/M')$$

其中,R为有效电影的评分; M为普通对象的数量; M'为数据集中任意两部电影之间匹配对象的最大数量; Hr(highest_rating) 为可以分配给一部电影的最大评分,在该例中是5。如果评级大于2.5(阈值等于最低和最高评级的平均值) ,那么可以将该电影及其计算的评级分配给相似的一部有效电影的电影集。接下来,使用数据集构建用户评分矩阵。这个矩阵的稀疏度为98.36%。使用形成的相似电影列表,减少了用户评分矩阵的稀疏性。对于用户评分矩阵中的每个非零条目,使用上述步骤中形成的列表找到与其类似的影片。一旦来自用户评分矩阵的稀疏性降低,就使用Pearson相关性来应用CF,并据此为用户生成最终的预测。

算法流程见图1。

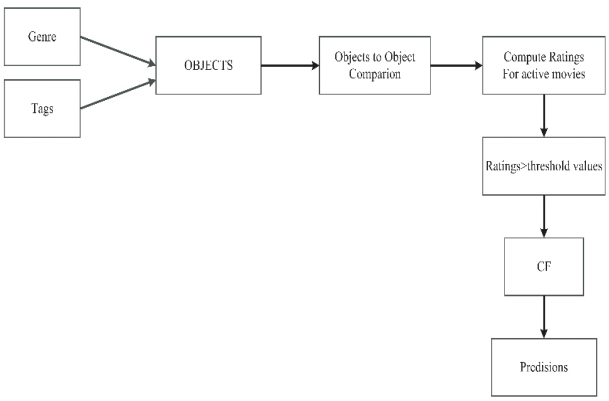

图1 改进的电影推荐混合协同过滤算法流程

4　评　估

在评估提出方法时,采用两种传统推荐方法作比较,即 Pure CF 和 SVD。用这两种方法在相同的 Movielens 数据集但不同的稀疏度上进行测试。此外,选择这两种方法进行评估和测试的原因在于,此前提出的内容提升的 CF[10] 的方法也是与这两种方法进行了比较。由于文中算法是为了比较梅尔维尔等所做的改进而设计的[10],因此,采用了前者在其工作中使用的相同评估方法进行测试。另外,凭借这种评估,可以清楚地看到提出的方法的结果和功效的差异。该模型使用平均绝对误差(MAE)进行评估。平均绝对误差是一个强大的评估模型,它是平均误差的更自然的度量。此外,维度评估和相互比较模型应该使用 MAE 作为评估指标[12],它是预测值与实际值的偏差。为了计算 MAE,考虑了预测评分和实际评分。MAE 值是在用户-事物矩阵的稀疏度级别不同上计算的,并针对所有三种算法分别进行了计算。此外,这里用于测试和评估的数据集要优于第二节提到的类似方法中使用的数据集。尽管用户总数较少,但是电影数量掩盖了这一事实,并且由于电影数量更多,因此允许算法在更稀疏的用户评分矩阵上运行,因此在下一节中提供的结果是合理的。

5　结　果

与 Pure CF 和 SVD 的方法进行比较的结果如表 1 所示。

表 1　混合 CF、纯 CF 和 SVD 的 MAE 值

Sparsity/%	Hybrid CF	Pure CF	SVD
98.3	0.773	0.769	2.520
98.5	0.774	0.773	2.592
98.7	0.778	0.775	2.761
99	0.781	0.777	2.978
99.5	0.811	0.821	3.200
99.8	0.902	0.922	3.574

与 Pure CF 相比,发现文中方法对高稀疏度的有效性比 Pure CF 更好,MAE 值比 Pure CF 产生的要稍高。原因是 Pure CF 算法取决于用户评分矩阵可用的数据,在高度稀缺的情况下,可用数据较少,因此 Pure CF 的性能表现不佳。另一方面,文中方法的主要兴趣领域是通过应用物-物比较来减少稀疏水平,因此,在这种情况下基于内容的过滤之后使用的 CF 比纯 CF 在更高的稀疏性的情况下使用效果更好。在图 2 中可以看出,在稀疏水平为 98.5% 左右时,Pure CF 和文中方法结果几乎没有差异,但在稀疏性进一步增加到 99% 左右的情况下,用 Pure CF 产生的结果的差异增加。如图 3 所示,当文中方法与 SVD 进行比较时,在稀疏度水平介于 98% 到 100% 的情况下,发现改进的混合 CF 推荐算法表现的比 SVD 更有效。从表 1 可以清楚地看出,在 98.3% 的稀疏度时,MAE 是 2.520,随着稀疏度的增加,MAE 值也随之增加,两种算法的 MAE 值的差异是巨大的。由于数据稀疏性较低,SVD 无法有效执行。表 2 显示了提出的方法是如何成功地降低了给定用户-事物评分矩阵的稀疏性的。

表 2　应用混合 CF 之后改善的稀疏性　%

Initial sparsity	Improved sparsity
98.3	97.15
98.5	97.24
98.7	97.45
99	97.81
99.5	98.37
99.8	98.94

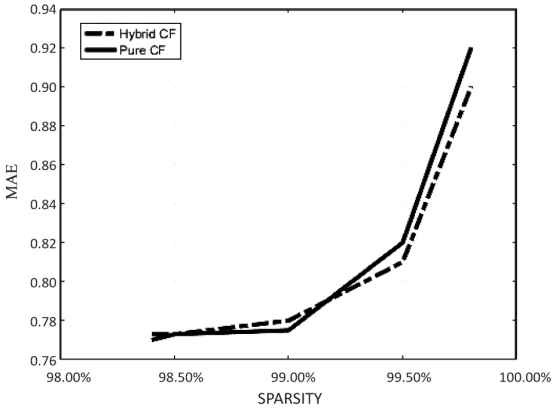

图 2　Hybrid CF 和 Pure CF 的 MAE 值与稀疏性的比较

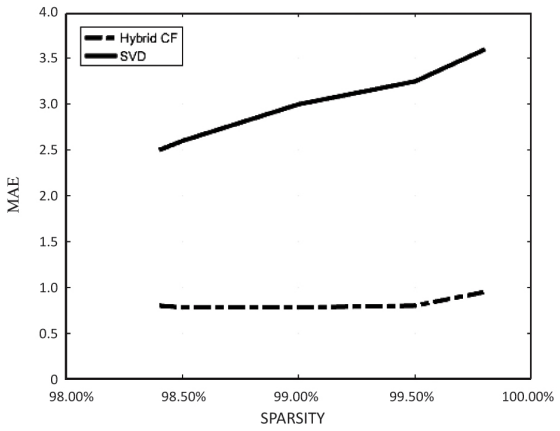

图 3　混合 CF 和 SVD 的 MAE 值与稀疏性增加的比较

6　结束语

文中方法是一种新颖的替代方法,描述了一种可以在基于内容的过滤中使用集合交集来找到两个特征之间的相关性的简单方法,该方法可以找出两个事物

(下转第 76 页)

的表现都优于传统 K-means，并且在 Iris、Seeds、Dum 不同数据集上均取得了较好的结果，说明 K-means 算法在聚类中心选择上得到了一定程度的优化，拥有更好的刻画数据分布特征的能力。同时当工作节点数量小于一定值时，加速比、扩展比可出现理想的变化，说明算法具有较好的并行化能力，但是当工作节点过多或者实验数据集过小时，算法的并行化效果就会受到影响，有待进一步完善。

参考文献：

[1] 刘 辉. 某些条件下的极大极小系统的全局最优解 [D]. 石家庄：河北师范大学，2017.

[2] 张忠林，曹志宇，李元韬. 基于加权欧式距离的 k_means 算法研究 [J]. 郑州大学学报：工学版，2010，31(1)：89-92.

[3] 贾瑞玉，管玉勇，李亚龙. 基于 MapReduce 模型的并行遗传 k-means 聚类算法 [J]. 计算机工程与设计，2014，35(2)：657-660.

[4] 冯晓云，陆建峰. 基于 Hive 的分布式 K_means 算法设计与研究 [J]. 计算机光盘软件与应用，2013(21)：62-64.

[5] 刘宝龙，苏 金. 双 MapReduce 改进的 Canopy-Kmeans 算法 [J]. 西安工业大学学报，2016，36(9)：730-737.

[6] 夏俊鸾，程 浩，邵赛赛，等. Spark 大数据处理技术 [M]. 北京：电子工业出版社，2015：67-68.

[7] 侯 玥. 基于最大最小距离聚类算法的改进多重心法选址研究 [D]. 大连：辽宁师范大学，2015.

[8] 成卫青，卢艳红. 一种基于最大最小距离和 SSE 的自适应聚类算法 [J]. 南京邮电大学学报：自然科学版，2015，35(2)：102-107.

[9] ABED H，ZAOUI L. Partitioning an image database by K_means algorithm [J]. Journal of Applied Sciences，2011，11(1)：16-25.

[10] 季 赛，谭 畅. 基于 UPGMA 聚类无线传感网络的簇头选择方法 [J]. 武汉理工大学学报，2010，32(16)：139-142.

[11] 翟东海，鱼 江，高 飞，等. 最大距离法选取初始簇中心的 k-means 文本聚类算法的研究 [J]. 计算机应用研究，2014，31(3)：713-715.

[12] 于彦伟，王 沁，邝 俊，等. 一种基于密度的空间数据流在线聚类算法 [J]. 自动化学报，2012，38(6)：1051-1059.

[13] SUN Qiao. An efficient distributed database clustering algorithm for big data processing [C]//Proceedings of 2017 3rd international conference on computational systems and communications. [s. l.]：[s. n.]，2017.

[14] GUHA S，RASTOGI R，SHIM K，et al. CURE：an efficient clustering algorithm for large databases [C]//ACM SIGMOD international conference on management of data. [s. l.]：ACM，1998：73-84.

[15] CUI Chunchun. Parallel CSA-FCM clustering algorithm based on MapReduce [C]//Proceedings of 2017 international conference on sports，arts，education and management engineering. [s. l.]：[s. n.]，2017.

(上接第 71 页)

之间的相似性并使用 CF 预测它们的推荐。以前相关的方法在基于内容的算法中使用了朴素贝叶斯等文本分类器，而文中算法进一步测试并与 Pure CF 和 SVD 进行了比较。评估后生成的 MAE 值提供了成功的比较。虽然混合内容推荐可产生更好的 MAE 值，并将数据集的稀疏性提高了 1% ~2%，但使用较大的数据集进行测试时，结果可能会有所不同。

参考文献：

[1] 李 川. 实时个性化推荐系统的设计与实现 [D]. 北京：北京邮电大学，2015.

[2] 马晓迪，宣 琦，张 哲，等. 协同过滤推荐算法在豆瓣网络数据上的研究 [J]. 计算机系统应用，2014，23(8)：18-24.

[3] HAN J，ISHII M，MAKINO H. A Hadoop performance model for multi-rack clusters [C]//5th international conference on computer science and information technology. Amman，Jordan：IEEE，2013：265-274.

[4] 黄宜华. 深入理解大数据：大数据处理与编程实践 [M]. 北京：机械工业出版社，2013.

[5] WU Xindong，ZHU Xingquan，WU Gongqing，et al. Data mining with big data [J]. IEEE Transactions on Knowledge and Data Engineering，2014，26(1)：97-107.

[6] MANYIKA J，CHUI M，BROWN B，et al. Big data：the next frontier for innovation，competition，and productivity [R]. [s. l.]：[s. n.]，2011.

[7] 胡 涛，周 兵，郑明辉，等. 基于 Hadoop 的移动云存储系统研究与实现 [J]. 华中科技大学学报：自然科学版，2013，41：181-183.

[8] 潘天鸣. 基于 Hadoop 平台的决策树算法并行化研究 [D]. 上海：华东师范大学，2012.

[9] 王全民，苗 雨，何 明，等. 基于矩阵分解的协同过滤算法的并行化研究 [J]. 计算机技术与发展，2015，25(2)：55-59.

[10] 杨志伟. 基于 Spark 平台推荐系统研究 [D]. 合肥：中国科学技术大学，2015.

[11] 艾聪聪. 推荐系统中多样性和新颖性算法研究 [D]. 长沙：湖南大学，2014.

[12] CHEN Y，HARPER F M，KONSTAN J，et al. Social comparisons and contributions to online communities：a field experiment on movie lens [J]. American Economic Review，2010(4)：1358-1398.

[13] 刘建国，周 涛，郭 强，等. 个性化推荐系统评价方法综述 [J]. 复杂系统与复杂性科学，2009，6(3)：1-10.

[14] 马建威，徐 浩，陈洪辉. 信息推荐系统中的朋友关系预测算法设计 [J]. 国防科技大学学报，2013，35(1)：163-168.