

基于改进型协同过滤算法的研究

徐志超,单剑锋

(南京邮电大学 电子与光学工程、微电子学院,江苏 南京 210046)

摘要:个性化推荐一直是互联网商品的重要特点,精准的个性化推荐一方面能够准确地定位市场,另一方面能够带来更好的用户体验。尽管基于不同的应用场景下的推荐算法种类越来越多,但是推荐算法的智能性、精准性、稳定性还有待提高。针对个性化的精准推荐需求,提出了一种基于用户的改进型协同过滤算法。该算法主要解决由于不同用户存在不同的评价体系造成的评分偏差以及用户由于本身的特征属性(年龄、兴趣、性别)的不同造成的评分偏差,进而造成余弦相似度计算偏差变大的问题。针对该问题,提出了一种融合型的余弦相似度计算方法,该方法包括一个相似度修正参数 α 和一个用户特征属性向量 $\vec{\beta}$,前者主要解决不同用户评价体系带来的偏差问题,后者是为了解决用户自身的特征属性不同产生的偏差问题。根据协同过滤算法应用在电影评分推荐实验上的分析表明,改进型协同过滤算法大大提高了实验效率和推荐准确率。

关键词:数据挖掘;个性化推荐;相似度修正参数;用户特征属性向量

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2019)10-0196-05

doi:10.3969/j.issn.1673-629X.2019.10.038

Research on Improved Collaborative Filtering Algorithm

XU Zhi-chao, SHAN Jian-feng

(School of Electronics and Optical Engineering and Microelectronics, NJUPT, Nanjing 210046, China)

Abstract: Personalized recommendation has always been an important feature of Internet products. On the one hand, accurate personalized recommendations can accurately locate the market, and on the other hand, it can bring a better user experience. Although there are more and more types of recommendation algorithms based on different application scenarios, the intelligence, accuracy, and stability of the recommended algorithms still need to be improved. Aiming at the demand of personalized accurate recommendation, we propose a user-based improved collaborative filtering algorithm, which mainly solves the problem of score bias caused by different evaluation systems of different users and the user's own characteristic attributes (age, interest, gender), thus resulting in larger calculation bias of cosine similarity. A fusion cosine similarity calculation method is proposed for this problem, which includes a similarity correction parameter α and a user feature attribute vector $\vec{\beta}$. The former mainly solves the bias caused by different evaluation systems of different users. The latter is to solve deviation caused by the user's own characteristic attributes. According to the analysis of the collaborative filtering algorithm applied to the film scoring recommendation experiment, the improved collaborative filtering algorithm greatly improves the experimental efficiency and recommendation accuracy.

Key words: data mining; personalized recommendation; similarity correction parameter; user feature attribute vector

0 引言

互联网社会中,人在社会生产生活中产生了源源不断的历史数据。文献[1]中较为全面地论述了推荐算法在电子商务中的应用,并且分析了推荐策略,同时指出了当前策略的优缺点和未来的研究方向。如何从大量的、复杂、冗余的历史数据中挖掘出有价值的数

据,分析这些数据的规律,对人的生产和生活做一些预测和相应的建议,这有利于生产力的提高和为人们提供更好的服务。相比于传统的软件工程,大数据、人工智能越来越被大众所熟悉^[2]。大数据的应用场景在生活中比比皆是,天猫、淘宝等购物网站通过消费者的历史消费数据对用户进行商品个性化推荐;学校图书馆中,根据不同读者的借书数据和读者的角色进行个性化推荐;淘淘网根据用户的影评和用户的历史观看数

收稿日期:2018-11-23

修回日期:2019-03-20

网络出版时间:2019-06-26

基金项目:江苏省第七批教育改革发展战略性与政策性研究课题(JG04018JX02)

作者简介:徐志超(1993-),男,硕士研究生,研究方向为数据挖掘和调制识别;单剑锋,副教授,研究方向为调制识别、数据挖掘、故障诊断。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190626.0829.032.html>

据对用户进行推荐;内容搜索行业的今日头条能够根据用户的历史浏览轨迹实时推荐用户感兴趣的内容,这大大提升了头条的用户流量。

目前,主流的推荐算法主要有三种,第一种是基于协同过滤的推荐算法,主要包括基于内存和基于模型两种。基于内存有包括基于用户或商品;基于模型主要利用统计学、机器学习、数据挖掘来研究^[3]。文献[4]提出的基于协同过滤算法的高校图书馆推荐系统,主要利用专业、角色、学历等多维特征构建读者模型,结合基于商品评分的系统过滤算法,相比于单一的基于商品评分协同推荐算法,该算法的有效性、实用性大大提高。第二种是基于内容的推荐算法,主要包括基于 TF-IDF 文本的推荐算法和基于潜在语义分析的推荐算法,但它们都只能基于历史的文本信息进行挖掘。文献[5]针对特征高维问题,提出一种基于中心词扩展的 TF-IDF 特征提取算法,增加了特征节点的表达能力,实现了特征降维。第三种是基于图结构的推荐算法,文献[6]提出了一种基于随机森林修正的加权二部图推荐算法。算法经过改进和融合后,提高了推荐的准确度,解决了基于二部图网络结构的算法中仅考虑用户与商品之间关系、忽略兴趣偏好影响的问题,从而增强了推荐的可解释性。另外,协同过滤算法也可以和其他经典算法相结合,文献[7]很好地将遗传算法与协同过滤算法进行有效结合。

在计算仿真平台和工具的选择上,以 MapReduce 为主的 Hadoop 体系和基于内存计算的 Spark 体系在计算上变得越来越重要。

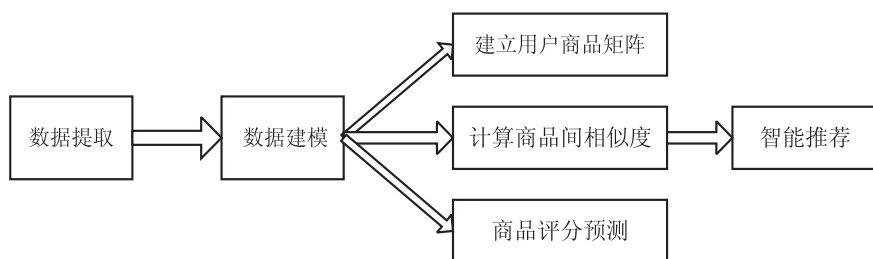


图 1 视频网站电影智能推荐流程

余弦相似度^[13]计算公式如下:

$$\text{sim}(x, y) = \frac{\sum_{i \in I_{x,y}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_{x,y}} r_{x,i}^2 \sum_{i \in I_{x,y}} r_{y,i}^2}} \quad (2)$$

相比于余弦相似度计算公式,皮尔森相似度计算公式考虑了不同用户评分平均分不同的情况。

1.2 评分估计方法

设 $U = \{u_1, u_2, \dots, u_n\}$ 为用户集合, $V = \{v_1, v_2, \dots, v_m\}$ 为商品集合, $r_{u,v}$ 表示用户 u 对商品 v 的评分估计。常见的评分估计方法^[13]如下:

针对传统的单机集中式计算已无法满足推荐系统的实时性和扩展性要求的问题,基于主流的大数据平台 Spark 在迭代计算以及内存计算方面的优势,设计了基于项目的协同过滤算法在 Spark 上的并行化方案^[8-9]。文献[10-11]都提出了基于 Hadoop 的协同过滤算法,成功提高了算法的运行速度,扩大了算法输入数据的规模。由此可见计算工具的提升,有助于算法性能的提升。

1 协同过滤算法

协同过滤算法主要是将与目标用户具有相似特征的用户的商品推荐给目标用户^[12];或者根据目标用户历史消费的商品,推荐相类似的商品。前者属于基于用户的推荐算法,后者属于基于商品的推荐算法。协同过滤算法大致思路见图 1。

1.1 协同过滤算法的相似度

相似度的计算方法主要有两种:皮尔森相似度计算和余弦相似度计算。

皮尔森相似度^[13]计算公式如下:

$$\text{sim}(x, y) = \frac{\sum_{i \in I_{x,y}} (r_{x,i} - \bar{r}_x)(r_{y,i} - \bar{r}_y)}{\sqrt{\sum_{i \in I_{x,y}} (r_{x,i} - \bar{r}_x)^2 \sum_{i \in I_{x,y}} (r_{y,i} - \bar{r}_y)^2}} \quad (1)$$

其中, $\text{sim}(x, y)$ 为用户 x 和用户 y 之间的相似度; $I_{x,y}$ 为用户 x, y 共同评过分的商品; $r_{x,i}$ 为用户 x 对商品 i 的评分; $r_{y,i}$ 为用户 y 对商品 i 的评分; \bar{r}_x, \bar{r}_y 分别为用户 x 、用户 y 的评价平均分。

$$r_{ui} = \frac{1}{n} \sum_{u \in U} r_{ui} \quad (3)$$

$$r_{ui} = \frac{\sum_{u \in U} w(u, u') r_{u'i}}{\sum_{u' \in U} w(u, u')} \quad (4)$$

$$r_{ui} = \frac{\sum_{u \in U} w(u, u') (r_{u'i} - \bar{r}_{u'})}{\sum_{u' \in U} w(u, u')} + \bar{r}_u \quad (5)$$

其中, U 表示与目标用户相似的用户集合; u' 表示 u 的相似用户; $r_{u'i}$ 表示相似用户对商品 i 的评分; $w(u, u')$ 表示目标用户和相似用户之间的权重; \bar{r}_u 表

示用户 u 对商品评价的平均分。

1.3 基于用户的内存推荐算法

基于用户的内存推荐算法主要先计算用户的相似度,然后根据相应的算法求出目标用户对目标商品的估计评分。表 1 中的数据来自某电商网站的用户评价商品的部分数据。表中 U 表示用户, V 表示商品,评分范围为 1 至 10。表中“-”表示未评价,“?”表示待评价。

表 1 不同用户对已购商品的评分

	V_1	V_2	V_3	V_4	V_5	V_6
U_1	6	7	-	4	7	8
U_2	5	8	7	-	8	-
U_3	9	9	8	7	9	7
U_4	8	10	6	9	10	6
U_5	5	?	10	-	7	9

基于用户的协同过滤算法,可以计算出目标用户 U_i 对商品 V_j 的评分结果 R_{ij} 。以表 1 中的数据为例,可以计算评估用户 U_5 对商品 V_2 的评分。具体步骤如下:第一,分别计算 U_5 和 U_i ($i=1,2,3,4$) 的相似复杂度;第二,根据除目标用户以外的其他用户对目标商品的评分和相似度,计算目标用户 U_5 对目标商品 V_2 的相似度 $r_{5,2}$ 。文中使用余弦计算用户之间的复杂度, U_5 和 U_i ($i=1,2,3,4$) 的相似复杂度如下:

$$\text{sim}(U_5, U_1) = \frac{6 \times 5 + 7 \times 8 + 7 \times 8}{\sqrt{(6^2 + 7^2 + 8^2)(5^2 + 7^2 + 9^2)}} = 0.993 \tag{6}$$

$$\text{sim}(U_5, U_2) = 0.974 \tag{7}$$

$$\text{sim}(U_5, U_3) = 0.947 \tag{8}$$

$$\text{sim}(U_5, U_4) = 0.914 \tag{9}$$

$$r_{52} = [\text{sim}(U_5, U_1)r_{12} + \text{sim}(U_5, U_2)r_{22} + \text{sim}(U_5, U_3)r_{32} + \text{sim}(U_5, U_4)r_{42}] / [\text{sim}(U_5, U_1) + \text{sim}(U_5, U_2) + \text{sim}(U_5, U_3) + \text{sim}(U_5, U_4)] = 8.43 \tag{10}$$

2 推荐系统评价准则

在协同过滤算法中,不论是采用基于哪一种推荐算法用于用户估计商品的评分,或者是用于对用户推荐一个商品的列表,都需要对估计的评分和推荐的列表进行评价,检验实际评分值和估计评分值之间的误差。误差越小,说明评分估计越准确,则实际推荐商品越准确。文献[14]对现有的推荐系统评价指标进行了系统回顾,总结了推荐系统评价指标的最新研究进展,从准确度、多样性、新颖性及覆盖率等方面进行多角度阐述,并对各自的优缺点以及适用环境进行了深入分析。特别讨论了基于排序加权的指标,强调了推

荐列表中商品排序对推荐评价的影响。一般地,对于用户对商品评分结果的检验可用平均绝对误差(MAE)或均方根误差(RMSE)评估评分的误差程度。

2.1 平均绝对误差

MAE^[14]用于度量用户估计评分和真实值之间的误差,其表达式为:

$$E_{MAE} = \frac{1}{n} \sum_{i \in U, j \in I} |p_{ij} - r_{ij}| \tag{11}$$

2.2 均方根误差

表达式为:

$$E_{RMSE} = \sqrt{\frac{1}{n} \sum_{i \in U, j \in I} (p_{ij} - r_{ij})^2} \tag{12}$$

其中, U 和 I 分别为用户集合和商品集合; p_{ij} 为真实值; r_{ij} 为估计评分值。

3 改进型的基于用户推荐算法

传统的基于用户推荐算法中,只是根据用户对商品的评分来估计对其他商品的评分以及将评分高的商品推荐给用户,单一的评分尺度往往无法挖掘用户深层次的需求。同时考虑到用户因生活背景、消费习惯等各种因素的不同,带来评分差异的不同。在有些情况下,由于这种评分差异巨大使得在使用余弦相似度计算用户间的相似度时,出现了极大的偏差^[15]。例如用户 U_1 、 U_2 、 U_3 对 4 种样品进行评分,见表 2。

表 2 不同用户对于不同商品的差异评分

	V_1	V_2	V_3	V_4
U_1	1	1	1	1
U_2	5	5	5	5
U_3	1	1	2	1

根据计算可知,用户 U_1 和用户 U_2 的相似度为 1,明显大于 U_1 和 U_3 的相似度。出现这种相似度计算结果偏差极大的原因,一方面是不同用户自身的评价差异性大;另一方面是余弦相似度只是根据不同用户的共同评分商品计算的,没有考虑到所选商品对不同用户的影响,也就是说,所选参与计算复杂度的商品有可能偏向用户 U_2 , U_1 和 U_3 虽然参与评分,但不一定真正感兴趣。另外也有可能用户 U_1 和 U_3 相对于用户 U_2 来说,更加理性,评分更加严格,也就是说不同用户的评分体系有可能不一致。

根据这种情况,从相似度本身出发,提出一种融合型的相似度计算公式,它由两部分组成。

3.1 余弦相似度修正型参数 α

余弦相似度修正型参数 α 主要是针对用户的评价体系不同而造成相似度计算偏差大的问题。其修正后的相似度表达式为:

$$\text{sim}_\alpha(x,y)=\alpha*\frac{\sum_{i\in I_{x,y}}r_{x,i}r_{y,i}}{\sqrt{\sum_{i\in I_{x,y}}r_{x,i}^2\sum_{i\in I_{x,y}}r_{y,i}^2}}\tag{13}$$
$$\alpha=1-\frac{1}{1+e^{\frac{\sum_{i\in S}(r_{u,i}-r_{v,i})^2}{|S|^2}}},\alpha\in[0,1)\tag{14}$$

其中, s 为用户 u 、 v 的共同评价商品的集合;
 $|S|$ 为用户 u 、 v 的共同评价商品数。该修正型参数 α
与不同用户对相同商品评分的差异性呈负相关,当用
户评分差异性大时, α 值偏小;反之, α 值偏大。

3.2 用户特征属性向量 $\vec{\beta}$

传统的基于用户的协同过滤算法只是考虑用户对
商品的评分,没有考虑用户自身的特征属性,这些属性
可以是用户的年龄、兴趣、性别等。比如不同年龄段、
不同兴趣的群体喜欢的电影类型肯定是有差异的。针
对该问题,从用户自身特征属性出发,提出了用户特征
属性向量 $\vec{\beta}=(\beta_1,\beta_2,\cdots,\beta_n)$ 。其中 $\beta_i=\frac{t_{u,v}}{t_{u_i}}\overline{r_{u,v}}$
表示对商品 v 进行评分的用户中含有特征 i 的用户个
数; t_{u_i} 表示对商品 v 进行过评分的所有用户个数; $\overline{r_{u,v}}$
表示具有特征 i 的用户对所有商品评分的平均分。

改进后的余弦相似度表达式为:

$$\text{sim}_\beta(x,y)=\frac{\sum_{i\in T}\vec{\beta}_x\times\vec{\beta}_y}{\sqrt{\sum_{i\in T}\vec{\beta}_x^2\sum_{i\in T}\vec{\beta}_y^2}}\tag{15}$$

其中, T 为特征属性的集合。
根据上述提出的两种改进型余弦相似度计算方
法,提出一种融合型的余弦相似度计算方法,即:
$$\text{sim}(x,y)=\gamma*\text{sim}_\alpha(x,y)+(1-\gamma)\text{sim}_\beta(x,y)\tag{16}$$

其中, γ 是一种平衡参数,可以看作是一种权重因
子,取值为 $[0,1]$ 。

4 实验仿真

在 TipDM-HB 平台进行视频网站的电影推荐建
模仿真步骤如下:

- (1) 导入经过简单预处理的 csv 数据,部分数据见
表 3;
- (2) 构建用户-商品矩阵,分别根据余弦相似度公
式和融合型的余弦相似度公式计算商品相似度;
- (3) 根据用户相似度和用户-商品矩阵,使用式 4
估算测试用户对不同商品的评分(步骤 2 和步骤 2 被
封装成协同过滤算法建模平台系统组件)以及将评分
高的电影推荐给用户;
- (4) 部分评分和推荐结果见表 4,分析电影推荐

结果。

表 3 导入的部分用户电影评价数据

用户 id	商品 id(电影 id)	用户对电影的评分
74913	19461	10
74913	29431	2
74913	35647	10
74914	61621	4
74914	68989	5
74914	147239	1
74915	50699	10
74916	8318	5
74916	19006	10
74196	78565	7

表 4 控制台输出的对部分用户推荐结果

用户 12		用户 74		用户 199	
电影 id	评分	电影 id	评分	电影 id	评分
11334	7.565	25	9.165	44771	6.565
22335	7.697	1155	8.667	14535	8.127
12456	7.561			6609	8.031

根据表 4 的推荐结果分析,用户 12、用户 74 和用
户 199 所推荐电影的评分都高于 7 分,这样的推荐结
果是有意义的,会产生较好的用户体验。改进型的基
于用户协同过滤算法有效解决了由于新项目冷启动导
致的用户推荐不准确问题,提高了推荐的精准度,进而
影响平台的受欢迎程度。

由 MAE 误差曲线(见图 2)可知,对比传统的基于
用户的协同过滤算法和改进型的基于用户的协同过滤
算法的平均绝对误差,在相同的用户邻居个数的条件
下,改进型的协同过滤算法的 MAE 明显小于传统的基
于用户的协同过滤算法的 MAE。

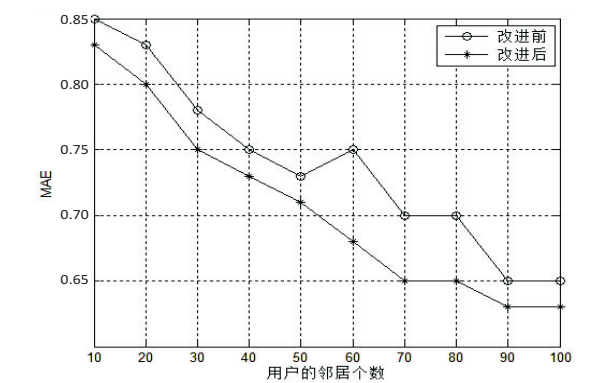


图 2 改进前和改进后的 MAE 误差曲线

5 结束语

由于传统的余弦复杂度计算公式只是从商品评分
本身出发,没有考虑到用户的评价体系不同和用户自

身的特征属性对商品评分的影响,因此在计算用户相似度时出现了极大的偏差。文中提出了一种改进型的协同过滤算法。第一,提出了一个余弦相似度修正参数 α ,通过该参数修正后,在计算评分差异大的用户之间相似度时,能够有比较好的修正作用;第二,提出了用户特征属性向量,该向量能够考虑到用户自身的特征属性,避免在计算相似度时出现较大偏差。通过上述的融合性相似度计算公式,能够解决相似度计算偏大过大的问题。

根据 TipDM-HB 平台的仿真数据来看,算法能够根据历史的电影评分估计出某用户未评分电影的得分,同时推送给用户评分比较高的电影。根据实验的推荐结果和 MAE 曲线可知,改进型的协同过滤算法的推荐性能有了一定的提升。

尽管该算法改善了传统的基于用户的协同过滤算法中出现的余弦相似度计算偏差的问题,但是也存在以下问题:第一,引入的用户特征向量带来了一定的计算复杂度;第二,修正参数 α 和用户特征向量能否进行扩展,用于改善皮尔森相似度计算方法。另外,根据对待计算数据的观察,发现有些数据预处理不到位,简单的预处理只是将数据格式进行调整,并没有对有些缺乏主要字段的数据进行舍弃。同时考虑到数据的复杂度,可以将数据分为合理评分记录和不合理评分记录,前者使用传统的相似度进行计算,后者使用改进型的相似度进行计算,以有效提升数据复杂度。因此,该算法有待进一步完善。

参考文献:

- [1] CAI Rui, LI Chen. Research on collaborative filtering algorithm based on MapReduce[C]//2016 9th international symposium on computational intelligence and design. Hangzhou: IEEE, 2016: 370-374.
- [2] 李国杰,程学旗. 大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中

国科学院院刊, 2012, 27(6): 647-657.

- [3] 郭 艳. 面向移动电子商务的个性化推荐策略研究[D]. 成都:成都理工大学, 2018.
- [4] 董 坤. 基于协同过滤算法的高校图书馆图书推荐系统研究[J]. 现代图书情报技术, 2011(11): 44-47.
- [5] 刘浩然, 丁 攀, 郭长江, 等. 基于贝叶斯算法的中文垃圾邮件过滤系统研究[J]. 通信学报, 2018, 39(12): 151-159.
- [6] 李 玲. 基于二部图的个性化推荐系统研究[D]. 北京:北方工业大学, 2018.
- [7] LIU Shouqiang, QI Ming, XU Qingzhen. Research and design of hybrid collaborative filtering algorithm scalability reform based on genetic algorithm optimization[C]//2016 6th international conference on digital home. Guangzhou: IEEE, 2016: 175-179.
- [8] 陆俊尧, 李玲娟. 基于 Spark 的协同过滤算法并行化研究[J]. 计算机技术与发展, 2019, 29(1): 85-89.
- [9] TIAN Hongpeng, CHEN Enjie. Research on collaborative filtering algorithm based on spark platform[C]//2017 international conference on industrial informatics - computing technology, intelligent technology, industrial information integration. Wuhan: IEEE, 2017: 33-36.
- [10] 李军华. 云计算及若干数据挖掘算法的 MapReduce 化研究[D]. 成都:电子科技大学, 2010.
- [11] ZHAO H, ZHANG H. The improved item-based clustering collaborative filtering algorithm based on Hadoop[C]//2017 IEEE 2nd advanced information technology, electronic and automation control conference. Chongqing: IEEE, 2017: 2416-2419.
- [12] 郝立燕, 王 靖. 基于项目流行度的协同过滤 TopN 推荐算法[J]. 计算机工程与设计, 2013, 34(10): 3497-3501.
- [13] 杨 博, 赵鹏飞. 推荐算法综述[J]. 山西大学学报:自然科学版, 2011, 34(3): 337-350.
- [14] 朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163-175.
- [15] 王云超, 刘 臻. 融合用户对项目和属性偏好的协同过滤算法[J]. 计算机科学, 2018, 45(S2): 412-416.