基于用户偏好和项目特征的协同过滤推荐算法

张应辉,司彩霞
(东北大学 计算机科学与工程学院,辽宁 沈阳 110000)

摘 要:采用对项目属性和用户行为的分析,为用户提供了一个有效的推荐资源解决方案(通过用户的兴趣偏好和项目的属性进行推荐)。对于用户而言,根据对用户注册时的显示属性和用户的历史行为记录(对项目资源的浏览、观看、下载、分享等操作)的分析,以及对用户历史行为的量化将用户划分为不同的近邻;对于项目而言,对项目也进行相似的操作即通过项目本身具有的属性和用户对项目的评价来将项目聚类分成不同的资源类型。以此对协同过滤算法进行改进,来改善推荐结果单一、评分矩阵数据不多、推荐准确性不高以及对新用户和新项目存在的冷启动问题。实现推荐资源随用户行为、兴趣的改变而动态改变,以满足用户需求,达到个性推荐的目的,避免用户在海量资源中为搜索资源而浪费时间。

关键词:协同过滤;推荐系统;用户属性;项目属性
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)01-0016-04
doi:10.3969/j.issn.1673-629X.2017.01.004

A Collaborative Filtering Algorithm Based on Interest of User and Attributes of Item

ZHANG Ying-hui, SI Cai-xia
(School of Computer Science and Engineering, Northeastern University, Shenyang 110000, China)

Abstract: Using the analysis of project properties and user behavior provides the user with a valid solution recommended resources. It is recommended by the interest of the user and item's attributes. To the user, according to analysis of the user registration display attributes and the user's behavior data (for browsing, viewing, downloading and sharing of project resources), as well as the quantization of the history of user behavior, the users could be divided into different neighbors. For the project, the clustering of project could divided into different resource types by its attributes and the evaluation of user to project. Therefore, the collaborative filtering algorithm is improved to solve the problems of single recommended results, little evaluation matrix data, low accuracy of recommendation as well as cold start for new users and new project. Recommended resources is achieved to change dynamically along with behavior and interest of users, to meet their requirements, achieving the purpose of the personalized recommendation, avoiding the waste of time for user to search resources in huge amounts of resources.

Key words: collaborative filtering; recommendation system; user attribute; item attribute

0 引 言

随着国际化以及国内经济的快速发展,人们的需求呈现出多样化的形式[1],如何让用户快速有效地选择自己所需要的资源是一项非常重要的工作。推荐技术因此应运而生[2]。推荐技术根据用户的兴趣爱好进行推荐,满足用户的不同需求。为了实现个性化推荐,研究人员不断研究改善,提出了许多的推荐算法,常用的有:基于关键字的检索[3],用户来选择自己所需资源的搜索,这种方式虽然在一定程度上满足了用户的需求,但是不能挖掘出用户的潜在感兴趣的资源;基于内容的推荐[4],虽然易于实现但由于现在文本形式的多样性,受到了一定的限制;基于协同过滤的推荐[5],是使用最广的推荐算法,很容易挖掘出目标用户潜在的兴趣,但存在一些问题,如评分矩阵数据中零元素较多造成的稀疏问题。

为了解决协同过滤算法存在的问题,提出了基于用户兴趣偏好和项目特征混合的算法(UIA-CF)。对于新用户和新项目根据其显示属性进行近邻的分类,

收稿日期:2016-06-03 **修回日期**:2016-09-14 **网络出版时间**:2017-01-04
基金项目:国家自然科学基金资助项目(61262058)
作者简介:张应辉(1972-),男,副教授,硕士生导师,研究方向为计算机图像处理、机器学习;司彩霞(1991-),女,硕士研究生,研究方向为数据挖掘技术。

再根据近邻的潜在兴趣对目标用户进行推荐,从而解决了新用户和新项目的冷启动问题,且对评分矩阵进行矩阵分解从而解决了稀疏问题。

1　用户分类及用户相似度的计算

协同过滤算法没有考虑用户属性,在此部分介绍用户属性及潜在兴趣的获得。用户显示属性的获得比较容易,由用户在注册时直接生成(如用户的年龄等信息),这些信息构成了用户的显式信息。

根据用户的在线时间,用户收藏,用户对项目的评分、评价等信息来计算用户的隐式属性[6-7],在用户事务数据库中记录了用户的行为。设用户浏览项目 I_j 的时间为 $\mathrm{Time}(I_j)$,其中在 User-Interest 中记录的是 $\frac{\mathrm{Time}(I_j)}{\mathrm{Time}}$(Time 为用户的在线时间),用户收藏浏览项目,则用户对浏览项目的兴趣度为 $\mathrm{Interest}(I_j=1)$,否则为 $\mathrm{Interest}(I_j=0)$。用户对项目的评分可以在 User-item 评分矩阵中获得,而分析用户对所有浏览项目的评价信息可以获得用户潜在喜欢的类型,由此获得用户的兴趣爱好。

通过以上介绍获得了用户的兴趣,得到用户兴趣矩阵 User-Interest,由 User-Interest 来计算用户间的相似度。用户 a,b 间的相似度计算[8]公式为:

$$\mathrm{sim}(a,b)=\frac{\sum_{i\in I}(R_{a,i}-\overline{R_a})(R_{b,i}-\overline{R_b})}{\sqrt{\sum_{i\in I}(R_{a,i}-\overline{R_a})^2}\sqrt{\sum_{i\in I}(R_{b,i}-\overline{R_b})^2}} \tag{1}$$

其中,I 为用户的一些属性,既包含显式属性又包含隐式属性;$\overline{R_a}$ 为用户 a 对用户属性的平均评分情况;$\overline{R_b}$ 为用户 b 对用户属性的平均评分情况。

2　项目分类及项目相似度的计算

项目的显式属性很容易获得,如项目(以电影为例)的作者、演员、播放时长、发布时间、类型等等。有时项目的显式属性可能对项目的性质描述得不太准确,则可以与所有用户对该项目的评价及评分分析出该项目的隐式属性相结合来显示项目的性质。

对项目的隐式属性的分析主要是对用户评价进行文本分析[9],采用文本分析法中的朴素贝叶斯算法来分析篇幅不是很长的用户评论。在贝叶斯算法[10-11]中用符号 $C_j\in C=\{C_1,C_2,\cdots,C_n\}$ 代表第 j 个类别,先计算先验概率 $P(W_t\mid C_j)$,即在给定类别 C_j 的条件下,每个独立词 W_t 的条件概率,计算公式如下:

$$P(W_t\mid C_j)=\frac{1+\sum_{i=1}^{|D|}N(W_t)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N(W_s,d_i)} \tag{2}$$

其中,$N(W_t)$ 为特征词 W_t 在所有用户对项目评价中出现的次数;$|D|$ 为项目评论的数量;$|V|$ 为训练集的总词汇表 $V=\langle W_1,W_2,\cdots,W_v\rangle$ 中的词汇数量。

类别 C_j 的先验概率为:

$$P(C_j)=\frac{1+\sum_{i=1}^{|D|}\mu}{|C|+|D|} \tag{3}$$

其中,$\mu\in\{0,1\}$,当 $d_i\notin C_j$ 时,$\mu=0$,否则 $\mu=1$。

当用户 a 对项目的评论 d_i 进行分类时,先计算项目相对于每一分类的后验概率,计算公式为:

$$P(C_j\mid d_i)=\frac{P(C_j)P(d_i\mid C_j)}{P(d_i)}=\frac{P(C_j)\prod_{k=1}^{d_i}P(W_{d_i}\mid C_j)}{P(d_i)} \tag{4}$$

其中,$P(d_i)$ 为常数。

式(4)可简化为:

$$P(C\mid d_i)=P(C_j)P(d_i\mid C_j)=P(C_j)\prod_{k=1}^{|d_i|}P(W_{d_i}\mid C_j) \tag{5}$$

式(2)的计算结果若有几个值比较接近且值比较大,把它们放入Item-Attribute 矩阵,则说明该项目的潜在类别不止一个而是多个,若计算结果中只有一个值比较大则说明该项目潜在类别属性只有一个。

通过以上公式的计算,可以得到项目属性矩阵 Item-Attribute,对 Item-Attribute 进行计算可以得到两个项目 i,j 的相似度计算公式:

$$\mathrm{sim}(i,j)=\frac{\sum_{t\in I}(X_{i,t}-\overline{X_i})(X_{j,t}-\overline{X_j})}{\sqrt{\sum_{t\in I}(X_{i,t}-\overline{X_i})^2}\sqrt{\sum_{t\in I}(X_{j,t}-\overline{X_j})^2}} \tag{6}$$

其中,t 为项目的某个属性;I 为项目的所有属性。

对以上取得的用户属性和项目属性,对这些数据进行转换、清洗、去噪等操作得到改进的用户-项目评分矩阵。

3　UIA-CF 推荐算法的实现

通过上述介绍已经对用户和项目进行了分类,然后就是对最近邻进行扫描搜索,产生推荐结果。文中采用 TOP-N[12]的推荐方式为目标用户推荐适合用户兴趣的项目资源。目标用户 a 对项目 i 的预测评分采用以下公式:

$$P_{a,i}=\delta\left(\overline{R_a}+\frac{\sum_{a'\in UN}\mathrm{sim}(a,a')(R_{a',i}-\overline{R_{a'}})}{\sum_{a'\in UN}\mathrm{sim}(a,a')}\right)+$$

$$(1-\delta)\left(\overline{R_i}+\frac{\sum_{i\in \mathrm{IN}}\mathrm{sim}(i,i')(R_{a,i'}-\overline{R_{i'}})}{\sum_{i\in \mathrm{IN}}\mathrm{sim}(i,i')}\right) \tag{7}$$

其中，UN 为用户 a 最近邻，a 为最近邻中的任何一个用户；IN 为项目 i 的最近邻，i' 为最近邻中的任何一个项目。

用户间的相似度和项目相似度分别由式(1)和式(6)计算得到，最后将预测评分的前 N 个以及目标用户没有浏览过的资源形成 TOP-N 推荐集，向目标用户推荐。

式(7)中的预测评分估计，不仅考虑了用户对项目的评分情况和用户之间的相似性，还考虑了项目之间的相似性及项目的平均评分情况，避免了只考虑用户或只考虑项目时的预测不准确的问题，提高了推荐的准确性。

根据上述介绍的各个部分的属性及计算方法，为使算法流程更加清晰，此算法的伪代码如下所示：

For each item $I_i \in I$

If(I_i not rated by U_i)

By the U_i's explicit attribute classification to the U_i, get the U_i's User-Interest

Else

By the U_i's explicit, implicit attribute and formulas(5), get the U_i's User-Interest

End if-else

By formulas(1) finding U_i's nearest neighbor UN

End for

For each User $U_i \in U$

If (I_i not rated by U_i)

By the I_i's explicit attribute classification to the I_i, get the I_i's Item-Attribute

Else

By the I_i's explicit, implicit attribute and formulas(2), get the I_i's Item-Attribute

End if-else

By formulas(6) finding I_i's nearest neighbor IN

End for

By the matrix decomposition, respectively, for the User-Interest and Item-Attribute operated into low dimension of dense matrix

By formulas(7) get the user rating of item, then user is recommended to take preliminary score first N

4 实 验

为了检验设计的 UIA-CF 算法的有效性和效率，利用开源数据集进行实验，检测其性能。

4.1 实验数据与测评准则

实验数据来自美国 Minnesota 大学 GroupLens 研究小组提供的 MovieLens[13]，并且选用最新更新(2015 年 8 月)电影评分数据集 ml-latest-small 1 MB 中等大小的数据集。将每条四元组记录转换成与初始用户-项目评分矩阵一样的形式，如表 1 所示。并且评分标准依然按照 GroupLens 的评分标准 1 ~ 5[14]，用前文提到的矩阵分解技术对初始矩阵进行分解，得到低维浓密的数据矩阵，将随机选出的用户行为数据集随机分为测试集(占 20%)和训练集(占 80%)。

表 1 用户-项目评分矩阵

	I_1	I_2	…	I_i	…	I_n
U_1	4	2	…	0	…	3
U_2	3	4	…	2	…	1
…	…	…	…	…	…	…
U_m	2	4	…	0	…	2

实验测评指标选用准确率、召回率、MAE(Mean Absolute Error)[15]。结合文中实例，准确率和召回率的计算公式为：

$$\mathrm{Precision}=\frac{\sum_{a\in A}|R(a)\cap T(a)|}{\sum_{a\in A}|R(a)|} \tag{8}$$

$$\mathrm{recall}=\frac{\sum_{a\in A}|R(a)\cap T(a)|}{\sum_{a\in A}|T(a)|} \tag{9}$$

其中，$T(a)$ 表示使用者在测试数据集上的行为列表；$R(a)$ 表示用户在训练数集上的行为给用户做出的推荐列表。

MAE 表示计算预测评分与实际评分之间的偏差的绝对值，公式为：

$$\mathrm{MAE}=\frac{\sum_{a,i\in I}|R_{a,i}-R'_{a,i}|}{n} \tag{10}$$

其中，n 表示项目的数量；$R'_{a,i}$ 表示用户对项目 i 的预测评分；$R_{a,i}$ 表示用户对项目 i 的真实评分。

4.2 实验结果及分析

改进算法中预测公式评分通过实验比较 MAE 的值来得到 δ 的值的大小，如图 1 所示。

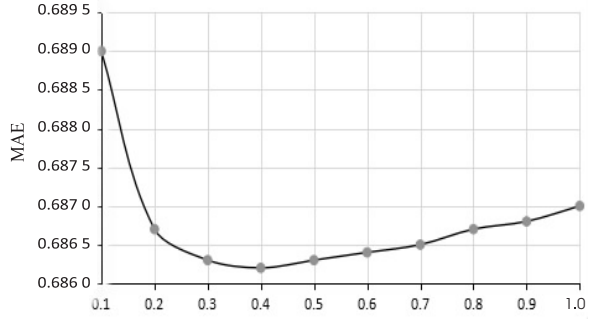

图 1 δ 的取值

改进算法与其他传统的推荐算法的 MAE 比较如图 2 所示。

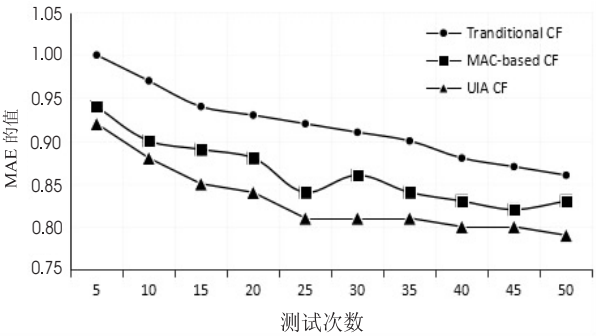

图 2　UIA-CF 与其他算法 MAE 的比较

表 2 的数据表明,UIA-CF 算法在准确率方面比协同过滤提高了 2.2%,召回率降低了 2.21%。

表 2　UIA-CF 同其他算法的比较　　%

指标	Traditional CF	MAC-based CF	UIA-CF
准确率	2.56	1.18	4.76
召回率	9.44	4.36	7.23

5　结束语

针对传统的协同过滤推荐算法中遇到的矩阵稀疏性和冷启动问题,提出了 UIA-CF 推荐算法,解决了上述问题,且实现了视频的个性化推荐,但是对视频推荐系统中的安全性没做进一步的研究,这些内容将在下一步研究中进行讨论。

参考文献:

[1] Qian Fulan, Zhang Yanping, Zhang Yuan, et al. Community-based user domain model collaborative recommendation algorithm[J]. Tsinghua Science and Technology, 2013, 18(4): 353-359.

[2] Xie F, Chen Z, Xu H, et al. TST: threshold based similarity transitivity method in collaborative filtering with cloud computing[J]. Tsinghua Science and Technology, 2013, 18(3): 318-327.

[3] 李莎莎. 面向搜索引擎的自然语言处理关键技术研究[D]. 长沙:国防科学技术大学,2011.

[4] 刘飞飞. 数字图书馆个性化信息推荐系统算法研究[J]. 情报科学,2012,30(12):1820-1823.

[5] Zhang L, Tao Q, Teng P Q. An improved collaborative filtering algorithm based on user interest[J]. Journal of Software, 2014, 9(4): 999-1106.

[6] Chatti M A, Dakova S, Thus H, et al. Tag-based collaborative filtering recommendation in personal learning environments[J]. IEEE Transactions on Learning Technologies, 2013, 6(4): 337-349.

[7] 田久乐. 基于协同过滤的电子商务个性化推荐算法分析[J]. 软件导刊,2014,13(6):36-38.

[8] 陈祎荻,秦玉平. 基于机器学习的文本分类方法综述[J]. 渤海大学学报:自然科学版,2010,31(2):201-205.

[9] 贾昱晟. 基于机器学习的中文文本分类技术研究[J]. 电脑知识与技术,2011,7(21):5194-5196.

[10] Hu Liang, Song Guochang, Xie Zhenzhen, et al. Personalized recommendation algorithm based on preference features[J]. Tsinghua Science and Technology, 2014, 19(3): 293-299.

[11] 蔡　嵩,张建明,陈继明,等. 云计算环境中基于朴素贝叶斯算法的负载均衡技术[J]. 计算机应用,2014,34(2):360-364.

[12] Salehi M, Kamalabadi I N, Ghoushchi M B G. An effective recommendation framework for personal learning environments using a learner preference tree and a GA[J]. IEEE Transactions on Learning Technologies, 2013, 6(4): 350-363.

[13] 贺桂和. 基于用户偏好挖掘的电子商务协同过滤推荐算法研究[J]. 情报科学,2013(12):38-42.

[14] Gogna A, Majumdar A. A comprehensive recommender system model: improving accuracy for both warm and cold start users[J]. Access IEEE, 2015, 3: 2803-2813.

[15] 王斌斌,周作建,过　洁,等. 基于迭代训练的 Web Service 混合协同过滤推荐模型[J]. 计算机研究与发展,2013,50:153-162.