文章编号:1671-5896(2018)04-0470-05

基于项目属性分类的协同过滤算法研究

吴佳婧[1],贺嘉楠[2],王越群[2],董立岩[2]
(1. 哈尔滨工程大学 计算机科学与技术学院,哈尔滨 150001; 2. 吉林大学 计算机科学与技术学院,长春 130012)

摘要: 用户对项目的评分数据是传统协同过滤算法进行项目或用户推荐的唯一依据,项目或用户本身的属性特征并未进行过多考虑。为此,在计算项目之间的相似度时融合了项目标签属性,提高了项目推荐的准确率。具体方法是首先通过创建项目属性分类表,得到项目属性之间的差异度,然后将项目属性差异度融入pearson相关系数公式中,计算项目之间的相似度。通过实验验证,改进后的方法比传统的基于项目的协同过滤算法的推荐结果平均偏差小,命中率高,推荐结果更加准确。

关键词: 协同过滤; 用户项目评分; 项目属性; pearson相关系数
中图分类号: TP391.3 **文献标识码:** A
DOI:10.19292/j.cnki.jdxxp.2018.04.018

Research on Collaborative Filtering Algorithm Based on Items´ Attribute Categories

WU Jiajing[1], HE Jianan[2], WANG Yuequn[2], DONG Liyan[2]
(1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China;
2. College of Computer Science and Technology, Jilin University, Changchun 130012, China)

Abstract: The algorithm of collaborative filtering is only considered to analyze the user-item evaluation matrix traditionally. The properties of the item or the user are often ignored. In order to solve this problem and improve the accuracy of the algorithm of collaborative filtering recommendation, we apply the attribute categories of the items into the formula for calculating the similarity of items. The specific method is as follows: firstly, get the degree of difference between the item properties by creating the items´ attribute categories table. Secondly, apply the degree of difference between the item properties into the pearson correlation formula and calculate the similarity between items. The experiment results show that the recommended MAE of the improved method is smaller and the hit rate is higher compared to the traditional collaborative filtering algorithm.

Key words: collaborative filtering; user program rating data; item properties; pearson correlation coefficient

0 引 言

随着计算机技术的飞速发展,人类社会已经逐步进入到电子信息时代。电子信息技术虽然改变了人们的生活方式,提高了人们的工作效率,但同时也给人们带来了新的问题。随着信息量的飞速增长,使人们在信息检索过程中很难找到正在需要的信息。推荐系统可以在一定程度上解决信息过载的问题,因为其实用性,推荐系统被国内外大多数的电子商务平台所青睐。因此对推荐系统的研究和改进得到了

收稿日期: 2018-04-17
基金项目: 国家自然科学基金资助项目(61272209)
作者简介: 吴佳婧(1996—),女,长春人,哈尔滨工程大学本科生,主要从事数据挖掘研究,(Tel)86-15567870997(E-mail)451587391@qq.com; 董立岩(1966—),男,长春人,吉林大学教授,博士生导师,主要从事数据挖掘研究,(Tel)86-15943013891(E-mail)dongly@jlu.edu.cn。

越来越多研究者的关注，而推荐算法则是推荐系统的核心。根据推荐的生成方式，推荐算法大致可以分为基于内容的推荐系统以及基于协同过滤的推荐系统。

协同过滤[1]推荐算法主要根据用户或项目的相似度进行推荐，其首要条件是需要找到目标用户[2]的相似用户或目标用户喜好项目的近邻项目，进而可以为目标用户推荐其兴趣值较高的未产生评分行为的项目[3,4]。由此可见用户间相似性的计算以及项目间相似性的计算对推荐系统的推荐准确率起到了重要的作用。对于传统的相似性计算[5,6]并没有考虑用户以及项目自身的属性，仅仅将用户项目评分矩阵[7-9]作为唯一的参考因素。

传统的相似性计算过程中只关注于用户项目评分矩阵，而忽略了用户或项目本身的属性特征，笔者将项目自身的属性标签[10]结合到相似性计算公式中，给出了基于权重调节的矩阵补全协同过滤推荐算法。

1　基于权重调节的矩阵补全协同过滤

1.1　算法思想

项目的标签不只有一种，可以有很多种类，对于一部电影而言，既可以属于喜剧，也可以属于情感类型，所以在进行相似度计算时，需要考虑每个项目的所有标签类别[11,12]，项目的标签重合度与项目的相似性成正比。

如表 1 所示，《女儿国》和《前任 3》都属于喜剧、爱情类型，与《红海行动》类型相比，这两部电影的类型更为相似，因此可以断定这两部电影的相似度更高。

表 1　基于项目属性的协同过滤推荐算法基本原理

Tab. 1　The basic principles of the Item-based collaborative filtering recommendation algorithm

	喜剧	爱情	动作	奇幻		喜剧	爱情	动作	奇幻		喜剧	爱情	动作	奇幻
红海行动	0	1	1	0	前任 3	1	1	0	0	女儿国	1	1	0	1

定义 1　项目属性值 $I_{i,x}$ 表示项目 i 是否具有属性 x。如果项目 i 具有属性 x，则项目属性值 $I_{i,x}$ 的值为 1，如果项目 i 不具有属性 x，则 $I_{i,x}$ 置为 0。

定义 2　$C_{i,j}$ 表示项目 i 与项目 j 之间的属性差异度，即项目 i 和项目 j 所具有的不同属性个数。可通过项目 i 中的每个属性值减去项目 j 中与之对应的每个属性值获得。计算方法如下

$$C_{i,j} = \sum_{x=1}^{n} |I_{i,x} - I_{j,x}| \tag{1}$$

其中 n 值由项目属性种类个数决定。项目差异度 $C_{i,j}$ 的值与项目之间的相同属性标签个数成反比。项目差异度的值越小，代表两个项目之间相同的属性标签越多，进而可以推测出两个项目越相似。因此可以得出项目差异度与项目之间的相似性成反比。

项目属性如表 2 所示，假设具有 A、B、C 3 个项目，并给出 S_1 到 S_5 等 5 个属性。每个项目都具有其相应的属性值。可以通过项目属性统计表获得项目 A、B、C 之间的属性差异度。

表 2　项目属性统计表

Tab. 2　The item properties statistics

属性	S_1	S_2	S_3	S_4	S_5
项目 A	1	0	1	1	0
项目 B	1	1	0	1	1
项目 C	1	1	1	1	0

3 个项目之间的项目属性差异度 $C_{A,B}$、$C_{A,C}$、$C_{B,C}$ 可由式(1)计算

$$C_{A,B} = |1-1| + |0-1| + |1-0| + |1-1| + |0-1| = 3$$

$$C_{A,C} = |1-1| + |0-1| + |1-1| + |1-1| + |0-0| = 1$$

$$C_{B,C} = |1-1| + |1-1| + |0-1| + |1-1| + |1-0| = 2$$

由计算可知 $C_{A,C} < C_{B,C} < C_{A,B}$，因此通过推理可得出，项目 A 与项目 C 之间的共同属性标签较多，相似性较大。

由于项目属性差异度与项目之间的相似性有相关性，所以将项目属性差异度应用到计算相似度的

person 相关系数[8] 中。由于项目属性差异度的取值均为大于等于 0 的整数，所以需要对权值进行修正，故引入指数函数 $y = e^{-x}$，由指数函数图像(见图 1)可知，当自变量取值为非负时，函数的值域区间为0 ~ 1。

将项目属性差异度与指数函数相结合，作为属性差异度的指数函数，并在计算两个项目的相似度时，引入属性差异度的指数函数 $y = e^{-C_{i,j}}$，并将差异度指数函数融入到 person 相关系数中，进行相似度计算的优化。改进后的 person 相关系数公式如下

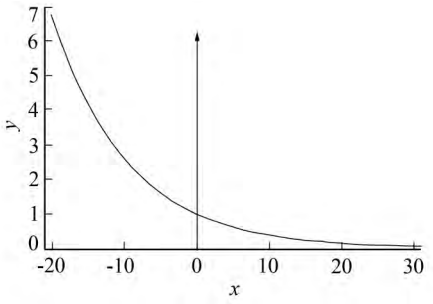

图 1 指数函数 $y = e^{-x}$ 的图像

Fig. 1 The image of the exponential function

$$S_{\text{sim}}(i,j) = \frac{\sum_{u\in I_{i,j}} (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u\in I_{i,j}} (R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u\in I_{i,j}} (R_{u,j} - \bar{R}_j)^2}} \tag{2}$$

1.2 算法流程

步骤 1 获得目标用户 u 的已评分项目集合 I_{Rated}，通过与所有项目集进行差运算，获得目标用户 u 的未评分但将要进行评分的项目集合 I_{unRated}。为目标用户 u 所推荐的项目将在 I_{unRated} 中产生，所以 I_{unRated} 也是目标项目集合。

表 3 项目用户统计表

Tab. 3 The item user statistics

UserID	$i \in I_{\text{Rated}}$	$j \in I_{\text{unRated}}$
1	$R_{1,i}$	$R_{1,j}$
…	…	…
n	$R_{n,i}$	$R_{n,j}$

步骤 2 在评分矩阵中查询对 I_{unRated} 项目集合中的项目进行评分过的用户集合与对 I_{Rated} 项目集合中的项目进行评分过的用户集合，二者求并集，形成用户集合 U，如表 3 所示。

步骤 3 利用

$$S_{\text{sim}}(i,j) = \frac{\sum_{u\in I_{i,j}} (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u\in I_{i,j}} (R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u\in I_{i,j}} (R_{u,j} - \bar{R}_j)^2}} \cdot e^{C_{i,j}} \tag{3}$$

即改进的 person 相关系数进行相似度计算，主要计算目标用户的目标项目$j(j \in I_{\text{unRated}})$ 与目标用户已评分项目 $i(i \in I_{\text{Rated}})$ 之间的相似度。

步骤 4 获得未评分项目 j 与已评分项目 i 的相似度后，进行为评分项目 j 的评分预测，评分预测公式如下

$$R_{u,j} = R_{u,i} S_{\text{sim}}(i,j) \tag{4}$$

所示。通过式(4)可以为目标用户进行未评分项目的评分矩阵补全。

步骤 5 通过步骤 4 获取目标用户的目标项目集中所有项目的预测评分，将评分按照由大致小的顺序进行排列。

步骤 6 从步骤 5 中的项目集合序列中，选取 Top K 个项目作为推荐结果，推送给目标用户。

2 算法实现

基于权重调节的矩阵补全协同过滤算法。

输入：用户评分矩阵 $\boldsymbol{R}_{mn}$，分类差距矩阵 $\boldsymbol{C}_{nn}$，目标用户 u，目标用户未评分项目 k。

输出：u 对 k 的预测评分 score。

```
1  ratsimTotal←0
2  simTotal←0
3  for i←1: n4      if (R_ui! =0) //用户 u 对项目 i 有评分
```

```
5        V_i←Ø, V_k←Ø
6        for j←1: m
7          if (R_ji! =0&&R_jk! =0) //用户 j 对项目 i、项目 k 均有评分
8          V_i←V_i ∪{R_ji}
9          V_k←V_k ∪{R_jk}
10        end if
11      end for
12      similarity←sim(V_i,V_k) · e^(C_ki)
13      simTotal←simTotal + similarity
14      ratSimTotal←ratSimTotal + similarity* R_ui15   end if
16    end for
17    score←ratSimTotal/simTotal
18    return score
```

3　实验结果与分析

3.1　前期准备

该实验采用的数据集来自 MovieLens ml-1M 数据集，由 6 040 个用户以及 3 900 个电影组成，并包含用户对这些电影的评分记录，每个评分包括用户评分的时间并以 Timestamp 形式表示，其中评分是由 5 个等级构成，分数越高代表用户对电影的喜爱程度越高。

以上数据记录在 3 张信息表中，user 表记录了用户的信息，movies 表记录了电影信息，ratings 表记录了评分信息。

3.2　实验结果与分析

实验 1　在同一个数据集下，比较添加了项目属性因素的协同过滤算法与传统协同过滤算法之间平均绝对值偏差 MAE[13,14] 的值，测试改进的算法是否在推荐精度上有所提高。结果如图 2 所示。

实验结果表明，在传统协同过滤算法中，考虑项目属性分类因素，将项目属性差异度指数函数融入到 person 相关系数中后获得的基于项目的(Item-based)协同过滤算法，其平均值绝对误差值更小，由此可以推断，改进的算法在推荐精度上有较大的提高。

实验 2　在同一个数据集下，比较添加了项目属性因素的协同过滤算法与传统协同过滤算法之间准确率 Precision，测试改进的算法是否在推荐准确率上有所提高，得到如图 3 所示的实验结果。

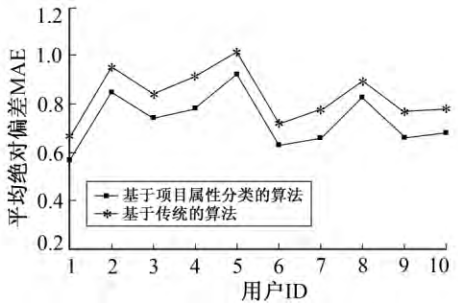

图 2　项目属性分类对 Item-based 协同过滤 MAE 的影响

Fig. 2　The item properties classified influence of Item-based CF on the MAE

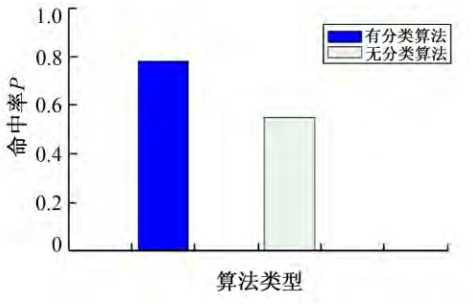

图 3　项目属性分类对 Item-based 协同过滤命中率的影响

Fig. 3　The item properties classified influence of Item-based CF on the hit rate

由图 3 可知，改进的基于项目推荐的协同过滤算法，在命中率上有很大的提高，由原 60% 的命中率提升到 80%。实验结果表明，结合了项目属性分类的基于项目的(Item-based)协同过滤算法，可以提高推荐的准确率，证明改进的算法具有一定的优势。

4 结 语

笔者针对传统的协同过滤算法进行了改进，在传统的协同过滤算法上，考虑了项目属性分类的因素。算法改进主要体现在算法的相似度计算上。通过构建项目属性统计表，对项目评分矩阵进行权重调节，并将项目属性差异度以指数函数的形式作为改进的 person 相关系数的计算参数。通过实验证明了改进算法的有效性，在提升推荐准确度上有很大的突破。

参考文献:

[1] 任磊. 一种结合评分时间特性的协同推荐算法[J]. 计算机应用与软件，2015，32(5)：112-115.
REN Lei. A Collaborative Recommendation Algorithm in Combination with Rating Time Characteristic [J]. Computer Application and Software，2015，32(5)：112-115.

[2] XI Xiaolong，CAO Lingling，WANG Xinheng. A Daptive Task Scheduling Strategy Based on Dynamic Workload Adjustment for Heterogeneous Hadoop Clusters [J]. IEEE Systems Journal，2014(99)：1-12.

[3] LUO Xin，LIU Huijun，GOU Gaopeng，et al. A Parallel Matrix Factorization. Based Recommender by Alternating Stochastic Gradient Decent [J]. Engineering Applications of Artificial Intelligence，2012，25(5)：1403-1412.

[4] BARALIS E，GARZA P. Item Selection for Associative Classification [J]. International Journal of Intelligent Systems，2012，27(3)：279-299.

[5] YANG X，STECK H，LIU Y. Circle-Based Recommendation in online Social Networks [C]//Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. [S. l.]：IEEE，2013：530-584.

[6] GUO Zhenhua，FOX G，ZHOU Mo. Investigation of Data Locality in Map Reduce [C]//Proceedings of the 12th IEEE. ACM International Symposium on Cluster，Cloud and Grid Computing. Ottawa，Canada：IEEE Computer，Society，2012：419-426.

[7] YANG B，LEI Y，LIU D，et al. Social Collaborative Filtering by Trust [C]//Proc of the 23rd International Joint Conference on Artificial Intelligence. [S. l.]：AAAI Press，2013：2747-2753.

[8] ZHANG Fuzhi，LIU Sai，LI Zhonghua，et al. Collaborative Filtering Recommendation Algorithm Incorporating User's Reviews and Contextual Information [J]. Journal of Chinese Computer Systems，2014，35(2)：228-232.

[9] YU Hong，LI Junhua. Collaborative Filtering Recommendation Algorithm Using Social and Tag Information [J]. Journal of Chinese Computer Systems，2013，34(11)：2467-2471.

[10] LIU Q，CHEN E H，XIONG H，et al. Enhancing Collaborative Filtering by User Interest Expansion via Personalized Ranking [J]. IEEE Transactions on Systems Man and Cybernetics，2012，42(1)：218-233.

[11] LIAO Zhifang，WANG Chaoqun，LI Xiaoqing. Tag Recommendation and New User Tag Recommendation Algorithms Based on Tensor Decomposition [J]. Journal of Chinese Computer Systems，2013，34(11)：2472-2476.

[12] JAMALI M，ESTER M. A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks [C]//Proc of the 4th ACM Conf on Recommender System，Vancouver，Canada：[s. n.]，2010：135-142.

[13] AN Wei，LIU Qihua，ZHANG Liyi. Review on Diversity in Personalized Recommender System [J]. Library and Information Service，2013，57(20)：127-135.

[14] GUO G. Integrating Trust and Similarity to Ameliorate the Data Sparsity and Cold Start for Recommender Systems [C]//Proc of the 7th ACM Conference on Recommender Systems. Singapore：[s. n.]，2013：451-454.

(责任编辑：刘东亮)