计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):168−172 [doi: 10.15888/j.cnki.csa.005914]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于时间与区域粒度的农资协同过滤算法①

董　恒[1], 孙丙宇[2], 张　颖[1]

[1](中国科学院大学 电子电气与通信工程学院, 北京 100049)
[2](中国科学院 合肥智能机械研究所, 合肥 230031)

摘　要: 相似度计算是基于用户的协同过滤算法中的一个关键步骤, 随着用户数的增加, 相似度的计算空间会越来越庞大, 同时在将其运用到农资领域个性化推荐时准确度较低. 针对这些问题, 结合农资受季节和地理位置影响强的特点对原有相似度计算方法进行改进, 提出了基于时间与区域粒度的农资协同过滤算法-TA-ACF(Agricultural collaborative filtering algorithm based on both time and area)核心思想是根据已有的农资需求调研结果, 建立时间与区域粒度矩阵, 据此构造此时间与区域粒度内的用户评分矩阵. 实验结果表明, 与基于用户的协同过滤推荐算法相比, TA-ACF 能够在保证时间效率的前提下, 较好的提高推荐的质量.

关键词: 个性化服务; 协同过滤算法; 区域与时间粒度; 相似度

引用格式: 董恒,孙丙宇,张颖.基于时间与区域粒度的农资协同过滤算法.计算机系统应用,2017,26(8):168–172. http://www.c-s-a.org.cn/1003-3254/5914.html

Agricultural Collaborative Filtering Algorithm Based on Both Time and Area

DONG Heng[1], SUN Bing-Yu[2], ZHANG Ying[1]

[1](School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China)

Abstract: Similarity calculation is a key step in the user-based collaborative filtering algorithm. As the number of users increases, the similarity computing space will become increasingly tremendous. At the same time, the accuracy is relatively low when it is applied to the agricultural personalized recommendation system. According to the feature of agricultural materials which are strongly influenced by seasons and locations, TA-ACF(Agricultural collaborative filtering algorithm based on both time and area) is proposed based on time and area size, which improves the original similarity calculation method. In this way, these above-mentioned problems could be solved. The main idea is to establish the matrix of time and size according to the existing research of the agricultural demands, and establish a rating matrix within the time and size. As the result shows, compared to the user-based collaborative filtering algorithm, TA-ACF is able to improve the quality of recommendations without losing time efficiency.

Key words: personal service; collaborative filtering; time and area based; similarity

随着电子商务技术的快速发展, 电子商务平台上的商品信息成爆炸式递增, 信息过载越来越严重, 尤其是在用户普遍信息素养不高的农资电子商务领域尤为突出. 因此, 如何解决信息过载, 挖掘农资用户需求, 为农资用户提供精细, 准确的服务成为农资领域的一个热点. 个性化服务[1]技术通过使用数据挖掘和信息过滤技术将产品、服务、信息推荐给潜在消费者, 有效的解决了互联网领域日益加剧的信息过载问题. 目前, 个

① 基金项目: 国家科技支撑计划项目(2014BAD10B08)
收稿时间: 2016-11-30; 采用时间: 2017-01-05

性化服务常用的技术包括基于规则的技术[2], 信息过滤技术[3], 其中, 基于用户的协同过滤算法[4]因其出现早, 使用简单, 准确度相对较高, 成为目前个性化服务中目前使用最广泛, 最成功的技术之一, 而稀疏性问题和冷启动问题[5]一直是困扰基于用户的协同过滤算法性能的关键问题.

针对协同过滤算法存在的问题, Sarwar[6]等提出了基于物品的协同过滤算法, 该算法基于一个基本假设“能够引起使用者兴趣的物品, 必定与其之前评分高的物品类似”通过计算物品间的相似性代替计算用户间的相似性. Marko[7]等人提出一种基于物品内容的协同过滤和基于用户的协同过滤相结合的方法, 该方法通过内容分析进行来获取用户画像(需求等信息), 之后再通过比较用户画像来决定用户的相似度. 于洪[8]等提出了用户时间权重的概念, 根据时间权重值的大小判定用户对新物品的偏爱程度, 在考虑用户, 标签, 物品属性, 时间等信息的基础上, 取得评分预测值. 郭艳红[9]等通过使用基于内容的预测方法, 对新物品进行项目评分, 之后使用优化步骤过滤预测不准确的用户评分, 在此基础上产生推荐. 李改[10]等人通过运用基于 K 近邻的属性—特征映射的算法得到新用户和新项目的特征向量, 一定程度上解决了该类协同过滤算法面临的冷启动问题. 金远平[11]等人通过改进相似度度量方法, 动态调节相似度计算值, 自适应调节目标用户和目标物品的最近邻对推荐结果的影响权重给出推荐结果. Ma[12]等提出一种通过手动调节影响因子来控制目标用户和目标项目对推荐结果的影响权重的推荐算法. 上述的算方法虽然进行了各种改进, 比较适用于普通商品的推荐, 并不能很好用于农资商品推荐. 农资商品不同于普通商品, 它具有很强的 , 将现有的基于用户的协同过滤算法用于农资商品领域时, 得到的推荐结果并不理想, 准确度低于电影等推荐时的准确率. 因此, 本文在结合农资商品特点, 农业规律和上述研究的基础上, 提出了一种新的基于时间与地域粒度的农资协同过滤算法. 该算法在相似度计算时综合考虑了对农资影响最大的两个因素—时间和地域, 更好的反映出农资用户需求的相似性. 并且该算法考虑到传统协同过滤冷启动的问题, 提出新用户、新物品和新系统的解决方案. 实验结果表明该算法提高了推荐准确度, 获得了更好的推荐效果.

1 基于用户的协同过滤算法

1.1 基于用户的协同过滤算法的工作过程及问题分析

协同过滤算法通常使用相似统计的方法得到相似爱好或者兴趣的相邻使用者, 一般步骤如下:

(1) 用户行为数据收集: 即收集用户的交易记录, 评价结果等, 在进行数据清理和格式化转换后, 形成用户—商品评价表. 如表 1 所示.

表 1 用户—商品评价矩阵

R_{IJ}	I_1	I_2	...	i_n
a	3	1	2	4
b	5	3	2	2
.....	?	3	1	4
n	4	?	4	3

(2) 寻找相似用户集合: 即在数据库中找到与目标用户相似度最高的用户集合. 一般采用的皮尔森相似度公式如下:

$$\mathrm{sim}(a,b)=\frac{\sum_{i\in S_{ab}}(r_{ai}-\overline{r_a})(r_{bi}-\overline{r_b})}{\sqrt{\sum_{i\in S_{ab}}(r_{ai}-\overline{r_a})^2}\sqrt{\sum_{i\in S_{ab}}(r_{bi}-\overline{r_b})^2}} \quad (1)$$

其中 a, b 分别表示两个不同用户, $\overline{r_a}$, $\overline{r_b}$ 分别表示用户 a, 用户 b 的评分平均值, S_{ab} 表示用户 a, b 共同评分的项目的集合.

(3) 产生推荐结果集: 找出步骤(2)生成的集合中用户喜欢的, 按照一定规则推荐给用户. 常用推荐算法一般为:

$$I(a,i)=\overline{r}_a+\frac{\sum_{b\in KNS_a}sim(a,b)*(r_{bi}-\overline{r_b})}{\sum_{b\in KNS_a}sim(a,b)} \quad (2)$$

其中 $I(a, i)$表示用户 a 对物品 i 的预测评分值, $\overline{r_a}$, $\overline{r_b}$ 分别表示用户 a, 用户 b 的评分平均值, r_{bi} 表示用户 b 对物品 i 的评分, KNS_a 表示用户 a 的近邻集合.

基于以上可以看出协同过滤算法主要基于用户对已有物品的评分, 建立用户物品评分矩阵, 进而计算用户间的相似度给出推荐结果, 一个大的电商平台, 积累了大量的用户多年的交易, 评价结果. 传统的用户相似度计算往往要将所有的用户, 都纳入相似度计算, 造成用户物品评价矩阵过于庞大, 浪费了大量的计算资源, 又没能考虑用户当前的兴趣迁移. 一个新用户到来时, 我们并没有这个新用户的历史行为数据, 所以也就无法按照传统算法根据用户历史行为预测其兴趣. 当新

的物品到来时, 没有任何用户对其进行过评价, 协同过滤算法无法主动对它进行推荐, 同时在一个新开发的电商网站上, 还没有用户, 也没有用户行为, 也无法启动推荐.

2 基于时间与区域粒度的农资协同过滤算法

2.1 相似度算法的改进

同其他商品相比, 农资具有以下的特点:

(1) 物品空间小, 物品种类也少(相对于书籍和音乐).

(2) 物品的重用率高(用户用一种化肥, 可能隔年还是用这种化肥).

(3) 季节规律性强.

(4) 受地理位置影响强.

传统的用户相似度直接根据所有用户的评价矩阵计算, 忽略考虑农资商品的规律和用户兴趣的变化. 农业活动的规律性真实的反映了农资用户潜在的需求变化. 故在传统的相似性算法中结合农业规律和农资个性化特点, 能够更加反映实际中用户的需求, 提高的推荐的准确性. 本文提出了时间与区域粒度的概念—根据已有统计的农业规律将全国划分为 M 个区域, 再结合当地的农业规律的基础上将全年划为 N 个农作时间, 这样全国全年可划分为 M*N 个粒度. 对应的农资电商平台全国全年的交易数据可以划分为 M*N 个粒度, 在进行用户相似度计算时, 只需先按照原有的余弦相似度计算方法计算用户 K 个年度所在粒度的用户相似度, 然后考虑用户兴趣随时间迁移的权重, 求和计算出用户相似度. 方法步骤如下:

(1) 根据已有的农业统计规律, 建立时间与区域粒度矩阵. 以安徽地区全年为例, 粗分为皖南和皖北两个地域粒度, 皖北地区又可分为冬耕(包括种植, 养护, 收获等阶段), 夏耕两个时间粒度. 所选地域粒度较大时评分矩阵过于庞大, 数据稀疏度高, 所选粒度较小时又会存在数据不足问题. 通常情况下, 我们可以以用户注册时的市为单位作为一个区域粒度, 同一区域粒度的气候, 土壤条件, 经济状况, 农资购买力更为接近. 同样时间粒度选取时也会存在上述问题, 在农业生产中, 一项农事活动常常不会超过半年, 因此, 在本文的实验过程中, 我们将 6 个月看作一个时间粒度.

(2) 用户登录时, 根据用户所在时间与区域粒度, 所在粒度的年度交易数据, 建立 K 个年度时间区域粒度用户评分矩阵. 我们认为用户最近时间的行为数据, 对用户当前的需求影响最大, 因此如果该用户在当前时间与区域粒度内还没有行为数据, 则使用上一个时间粒度的交易行为数据构建用户评分矩阵.

(3) 在(2)的基础上分别计算出各个用户评分矩阵的用户相似度, 加权求和. 时间区域粒度用户相似度计算公式如下:

$$TA-s(a,b)=\sum_{i=1}^{K}\lambda_i \mathrm{sim(a,}b) \tag{3}$$

其中, a, b 为属于同一时间与区域粒度的两个用户, K 表示实验数据所选用的年数($k\in[1, 3]$), λ_i 表示年度权重值$\lambda\in(0, 1)$且$\sum_{i=1}^{K}\lambda_i=1$, 且$\lambda_i$ 与 K 值成反比(离当前粒度年数越长, 权重越低).

2.2 冷启动问题解决方案

(1) 新用户问题: 根据以往该时间与区域粒度内的用户的交易数据构造用户消费热单, 将排序后最热的五个推荐给新用户.

(2) 新系统与新物品问题: 充分利用专家和农业规律, 对系统中已存在的物品和新加入的物品进行时间与区域粒度基因标注. 标注后, 根据时间与区域粒度基因随机抽取若干个物品推送个用户.

3 实验结果与分析

3.1 数据来源

本文所使用的数据来源于上农网 2013 年 1 月到 2016 年 1 月在安徽省的范围内, 328 个用户对 150 个农资物品(包括农药, 种子种苗, 肥料, 农膜, 中小农具等)所产生的 8167 条交易与评价数据, 评分值范围为 1-5. 实验中随机选用整个实验数据集的 80% 作为训练集, 剩余的作为测试集.

3.2 评价标准与实验环境

本文采用单个用户产生推荐结果时间(SRT)、评分预测的均方根误差(RMSE)和综合评价指标(F-Score)三个指标. 假设用户对 M 个物品的预测评分为$\{p_1, p_2\ldots\ldots p_m\}$, 用户对 M 个物品的真实评分为$\{r_1, r_2\ldots\ldots r_m\}$.

$$SRT=\frac{O_N}{N} \tag{4}$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{m}(r_i-p_i)^2}{M}} \tag{5}$$

$$F=\frac{2\times P\times R}{P+R} \tag{6}$$

$$P=\frac{\sum_u |R(U)\cap T(u)|}{\sum_u |R(u)|} \tag{7}$$

$$R=\frac{\sum_u |R(U)\cap T(u)|}{\sum_u |T(u)|} \tag{8}$$

其中 O_N 为算法给 N 个用户产生评分预测运行的时间, P 为推荐准确率, R 为召回率, $R(u)$为给用户 u 推荐的 M 个商品集合, $T(u)$为用户 u 真是评分过的商品集合.

本文的实验环境为: 主频: 2.5 GHZ 软件: win7, MyEclipse 2014.

3.3 与传统的基于用户的协同过滤算法(UBCF)比较

从图 1 可以看出改进后的基于时间与区域粒度的农资协同过滤算法(TA-ACF)在产生单个用户推荐结果时间上很接近传统的农资协同过滤算法, 并且随着 K 值(K 为邻居个数)得增加慢慢接近.

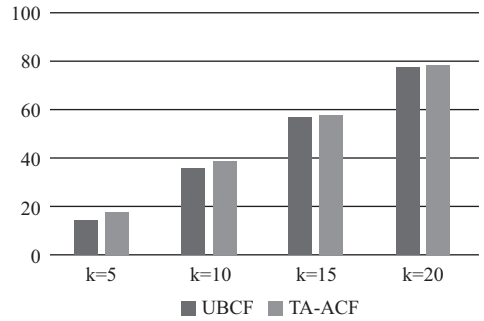

图 1 SRT 对比图(时间单位为 ms)

从图 2 可以看出 TA-ACF 算法的均方根误差明显的低于传统的 UBCF 算法, 并且随着 K 值得增大, 仍能很好的保持优势, 保证预测质量.

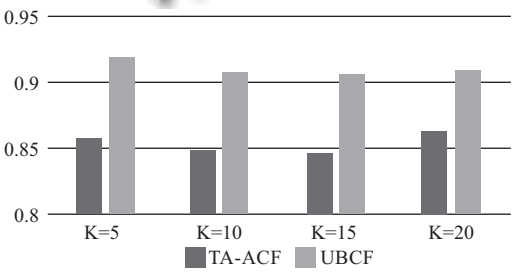

图 2 RMSE 比较图

从表 2 和表 3 看出, 随着推荐商品数量的增加, TA-ACF 算法的准确率和召回率都优于传统的协同过滤算法.

表 2 UBCF 和 TA-ACF 准确率 P(%)比较

推荐数量	10	15	20	25
UBCF	30.3	32.2	37.8	45.2
TA-ACF	45.8	47.4	53.6	54.3

表 3 UBCF 和 TA-ACF 召回率 R(%)比较

推荐数量	10	15	20	25
UBCF	0.06	0.12	0.15	0.22
TA-ACF	0.14	0.17	0.26	0.26

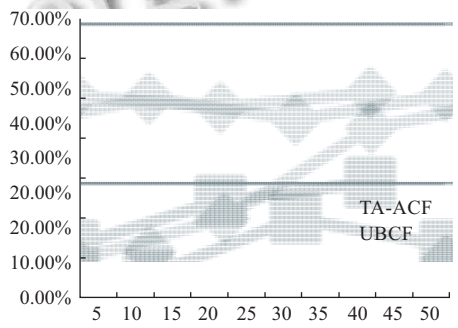

图 3 F-Score 比较图

在推荐商品数量都为 20 时, 在不同邻居数 K 的前提下, 使用 TA-ACF 和 UBCF 算法分别进行推荐, 比较两者推荐的综合指标(F-Score), 从图 3 中可以看出, 随着邻居数量 K 的增加, TA-ACF 的综合指标整体高于 UBCF 的综合指标.

4 结束语

本文针对协同过滤算法在农资电商领域的运用时准确度低的问题, 提出了基于时间与区域粒度的农资协同过滤算法. 综合的考虑了地理位置和时间因素对用户农资需求的影响, 对相似度计算方法进行了改进. 并进行了多次的实验, 实验结果证明, 该算法能够较好的提高推荐的质量. 另外本文中的年度权重系数λ_i 以及时间与区域粒度的划分是人为设定的. 因此, 下一步的研究重点是如何能够在保证推荐质量的同时, 让算法自主的维护λ_i 的值, 划分时间与区域粒度.

参考文献

1 Deng SG, Huang LT, Wu J, *et al*. Trust-based personalized service recommendation: A network perspective. Journal of Computer Science and Technology, 2014, 29(1): 69–80. [doi:

10.1007/s11390-014-1412-2]
2 Hayes-Roth F. Rule-based systems. Communications of the ACM, 1985, 28(9): 921–932. [doi: 10.1145/4284.4286]
3 Hanani U, Shapira B, Shoval P. Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction, 2001, 11(3): 203–259. [doi: 10.1023/A:1011196000674]
4 Zhao ZD, Shang MS. User-based collaborative-filtering recommendation algorithms on hadoop. Proc. of 2010 Third International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA. 2010. 478–481.
5 Schein AI, Popescul A, Ungar LH, *et al*. Methods and metrics for cold-start recommendations. Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland. 2002. 253–260.
6 Sarwar B, Karypis G, Konstan J, *et al*. Item-based collaborative filtering recommendation algorithms. Proc. of the 10th International Conference on World Wide Web. Hong Kong, China. 2001. 285–295.
7 Balabanović M, Shoham Y. Fab: Content-based, collaborative recommendation. Communications of the ACM, 1997, 40(3): 66–72. [doi: 10.1145/245108.245124]
8 于洪, 李俊华. 一种解决新项目冷启动问题的推荐算法. 软件学报, 2015, 26(6): 1395–1408.
9 郭艳红, 邓贵仕. 协同过滤系统项目冷启动的混合推荐算法. 计算机工程, 2008, 34(23): 11–13.
10 李改, 李磊. 一种解决协同过滤系统冷启动问题的新算法. 山东大学学报(工学版), 2012, 42(2): 11–17, 44.
11 黄裕洋, 金远平. 一种综合用户和项目因素的协同过滤推荐算法. 东南大学学报(自然科学版), 2010, 40(5): 917–921.
12 Ma H, King I, Lyu MR. Effective missing data prediction for collaborative filtering. Proc. of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, The Netherlands. 2007. 39–46.