

基于用户历史行为的协同过滤推荐算法

王志虎^{1,2}, 黄曼莹²

(1 武汉理工大学 信息工程学院, 湖北 武汉 430070;
2 广西经济管理干部学院 计算机系, 广西 南宁 530007)

摘 要: 协同过滤推荐是数据挖掘一个重要方向, 传统协同过滤推荐算法受到数据稀疏性和冷启动制约, 难以获得理想的推荐结果, 为了改善协同过滤推荐的准确性, 提出了基于用户历史行为的协同过滤推荐算法. 首先根据用户的历史行为预测用户对每一个项目的偏好程度, 并采用标签描述用户对项目的偏好程度, 建立相应的特征向量, 然后根据特征向量计算项目相似度实现个性化推荐, 最后采用多个经典数据进行了仿真测试, 以验证算法的优越性. 测试结果表明, 该算法大幅度降低了推荐的误差, 提高了协同过滤推荐的准确率, 克服了传统协同过滤推荐算法存在的局限性, 而且可以加快推荐速度, 具有更高的实际价值.

关键词: 数据挖掘; 协同过滤; 用户偏好; 项目相似度; 个性化推荐

中图分类号: TP31 文献标识码: A 文章编号: 1000-7180(2017)05-0132-05

Collaborative Filtering Recommendation Algorithm Based on User's Historical Behavior

WANG Zhi-hu^{1,2}, HUANG Man-ying²

(1 College of Information Engineering, Wuhan University of Technology, Wuhan 430070, China;
2 School of Computer, Guangxi Economic Management Cadre College, Nanning 530007, China)

Abstract: Collaborative filtering is an important direction of data mining, the traditional collaborative filtering recommendation algorithm is sensitive for by data sparsity and cold start control, so recommendation result is difficult to achieve the ideal. In order to improve the accuracy of collaborative filtering recommendation, a collaborative filtering recommendation algorithm based on user behavior prediction is proposed in this paper. Preference for resources is predicted of according to the user's historical behavior, label is used to describe users Preference for resources and establish feature vector, and secondly similarity of resources are computed according to the feature vector and achieve personalized recommendation, finally a number of classic data are adopted to carry out the simulation test to verify the superiority. The test results show that the proposed algorithm greatly reduces the error of collaborative filtering recommendation and improve the accuracy to overcome the limitations of traditional collaborative filtering algorithms, but also can accelerate the recommended speed, so it has higher practical value.

Key words: data mining; collaborative filtering; user preference; resource similarity; personalized recommendation

1 引言

随着 Internet 的迅速发展以及信息化水平的不断提高, 人们进行了电子商务时代. 每一天网络上产生大量的数据, 用户从海量数据中找到自己真正需要的数据, 好像大海捞针, 极其费时, 在该背景下,

推荐系统应运而生^[1-2]. 推荐系统可以根据用户需求实现智能化推荐, 其中协同过滤推荐算法是使用最多、最成功的个性化推荐算法, 如何实现性能优良的协同过滤推荐算法一直是数据挖掘中的热点问题^[3].

针对协同过滤推荐问题, 研究人员进行了不懈

努力的研究,主要根据目标用户和其他用户之间目标项目的评价相似度实现目标项目的推荐^[4].迄今为止,算法主要分为两类:基于用户评分的协同过滤推荐算法和基于项目评分的协同过滤推荐算法^[5].在实际应用中,两类协同过滤推荐算法均有各的应用范围,但它们均受到数据稀疏性和冷启动制约,推荐结果不稳定^[6-7].为了解决该难题,研究人员对协同过滤推荐算法进行相应的改进和完善,如有学者提出聚类分析和最近邻相融合的协同过滤算法^[8];有学者设计了基于信任模型填充的协同过滤推荐算法,它们一定程度上改善协同过滤推荐效果^[9],但是数据稀疏性与冷启动仍然没有得到完全解决,推荐质量有待提高^[10-11].

为了改善协同过滤推荐的准确性,提出了基于用户历史行为的协同过滤推荐算法,并采用多个经典数据验证其的性能.测试结果表明,该算法大幅度降低了推荐的误差,提高了推荐准确性,而且加快了推荐速度.

2 个性化推荐系统的原型

个性化推荐系统的对象包括用户和项目两类,其中用户是系统使用者,项目是被推荐对象,基本原理如图1所示.在图1中,输入模块用于收集和更新用户的相关信息,如用户的历史行为、用户属性等;推荐引擎主要设计合理推荐算法;输出模块主要输出商品推荐列表,其中设计合理推荐算法最为关键,直接影响个性化推荐系统的推荐质量^[12].

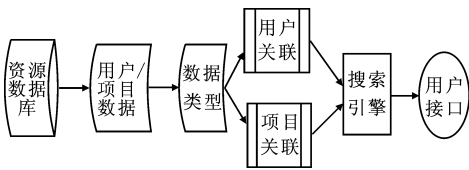


图1 协同过滤的推荐系统原型

3 基于用户历史行为预测的协同过滤推荐算法

3.1 用户特征向量

用户特征向量经常用于描述用户对项目偏好程度,而根据项目偏好程度计算项目相似度.设 N 个用户组成的集合为: $User = \{u_1, u_2, \dots, u_j, \dots, u_N\}$, p 为项目组成的项目集合为: $Item = i_1, i_2, \dots, i_j, \dots, i_p$; M 为标签组成的集合为: $Tag = \{t_1, t_2, \dots, t_j, \dots, t_M\}$,采用用户点击的项目以及项目对应词频值建立用户的特征向量,则有

$$V_{u_l} = tf_{\text{user_tag}} \times idf_{\text{user_tag}} \left\{ \frac{n_{u_l}}{n_{u_l}} \log\left(\frac{N}{N_{u_l}}\right) \dots \frac{n_{u_i}}{n_{u_i}} \log\left(\frac{N}{N_{u_i}}\right) \dots \frac{n_{u_M}}{n_{u_M}} \log\left(\frac{N}{N_{u_M}}\right) \right\} \quad (1)$$

式中, n_{u_i} 为用户使用标签 t_i 的次数; $f_{\text{user_tag}}$ 为计算标签的频率; $idf_{\text{user_tag}}$ 为标签重要程度,它们为

$$\begin{cases} tf_{\text{user_tag}} = \frac{n_{u_i}}{n_{u_i}} \\ idf_{\text{user_tag}} = \log\left(\frac{n}{n_{u_M}}\right) \end{cases} \quad (2)$$

3.2 项目的特征向量

项目特征向量主要描述标签对项目的重要程度,计算公式为

$$V_{u_j} = tf_{\text{item_tag}} \times idf_{\text{item_tag}} \left\{ \frac{n_{u_l}}{n_{u_l}} \log\left(\frac{P}{N_{u_l}}\right) \dots \frac{n_{u_i}}{n_{u_i}} \log\left(\frac{P}{N_{u_i}}\right) \dots \frac{n_{u_M}}{n_{u_M}} \log\left(\frac{P}{N_{u_M}}\right) \right\} \quad (3)$$

式中, n_{u_j} 为标签 t_j 标记项目 i 的次数; n_{u_i} 为项目 i 的总的标签数; $f_{\text{item_tag}}$ 表示项目使用的频率; $idf_{\text{item_tag}}$ 表示标签对项目的重要程度,它们为

$$\begin{cases} tf_{\text{item_tag}} = \frac{n_{u_j}}{n_{u_i}} \\ idf_{\text{item_tag}} = \log\left(\frac{P}{N_{u_M}}\right) \end{cases} \quad (4)$$

3.3 用户-项目的偏好矩阵

用户对项目偏好矩阵的计算公式为

$$V_{u_j i_k} = \sum_{t=1}^M V_{u_j t} \times V_{i_k t} \quad (5)$$

式中, $u_j \in U, j = 1, 2, \dots, N; i_k \in I, k = 1, 2, \dots, P$.

用户 u_j 的项目偏好特征向量可以描述为

$$V_{u_j} = (V_{u_j i_1}, V_{u_j i_2}, \dots, V_{u_j i_k}, \dots, V_{u_j i_N}) \quad (6)$$

式中, $V_{u_j i_k}$ 为 u_j 对 i_k 的偏好程度.

用户-项目偏好矩阵描述了反映用户对各项目的喜爱程度,其可以表示为

$$V_{N \times P} = \begin{bmatrix} V_{u_1 i_1} & \dots & V_{u_1 i_k} & \dots & V_{u_1 i_p} \\ \vdots & \ddots & \vdots & & \vdots \\ V_{u_j i_1} & & V_{u_j i_k} & & V_{u_j i_p} \\ \vdots & & \vdots & \ddots & \vdots \\ V_{u_N i_1} & & V_{u_N i_k} & & V_{u_N i_p} \end{bmatrix} \quad (7)$$

3.4 项目相似度计算

协同过滤描述算法通常根据项目的相似度描述不同项目之间的相似程度,在传统算法的基础上,本文算法在项目的相似度计算过程中引入了标签信息,那么项目的特征信息为

$$\mathbf{I}_k = (n_{k1}, n_{k2}, \dots, n_{ki}, \dots, n_{kL}) \quad (8)$$

式中, $k = 1, 2, \dots, N$; $I = 1, 2, \dots, P$; n_{ki} 为归一化后的值。

项目的特征向量矩阵 $\mathbf{I}_{k \times k}$ 可以描述为

$$\mathbf{I}_{k \times k} = \begin{bmatrix} n_{11} & \dots & n_{1k} \\ \vdots & \ddots & \vdots \\ n_{k1} & \dots & n_{kk} \end{bmatrix} \quad (9)$$

计算项目间的余弦相似度, 项目 i_i 和 i_j 的相似度为

$$\text{sim}(i_j, i_k) = \cos(\mathbf{I}_j, \mathbf{I}_k) = \frac{\mathbf{I}_j \cdot \mathbf{I}_k}{|\mathbf{I}_j| \times |\mathbf{I}_k|} \quad (10)$$

根据式(10)的计算结果建立项目相似度矩阵 $\mathbf{S}_{P \times P}$, 即

$$\mathbf{S}_{P \times P} = \begin{bmatrix} 1 & \dots & s_{1j} & \dots & s_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{j1} & & 1 & & s_{jp} \\ \vdots & & & \ddots & \vdots \\ s_{p1} & \dots & s_{pj} & \dots & s_{pp} \end{bmatrix} \quad (11)$$

3.5 用户偏好的预测

对用户历史行为进行分析, 得到用户 u_j 历史使用项目 i_k 的偏好值 $p_{u_j i_k}$, 计算项目 i_i 和 i_j 之间的相似度 $s_{i_k i_j}$, 然后预测用户 u 使用项目 i_j 的偏好值, 得到:

$$pp_{u_{i_j}} = \sum_{k=1}^N p_{u_{i_k}} \times s_{i_k i_j} \quad (12)$$

3.6 用户历史行为的协同过滤推荐算法工作步骤

(1) 建立用户-标签矩阵, 并根据用户-标签矩阵估计使用标签的用户数。

(2) 建立项目-标签矩阵, 并根据项目-标签矩阵估计项目被不同标签标记的次数。

(3) 计算用户和项目特征向量和项目的标签特征向量, 结合用户对项目的喜欢程度建立用户-项目偏好矩阵。

(4) 计算每一个项目之间的相似度, 并建立项目的相似度矩阵。

(5) 根据用户的历史行为, 结合项目之间相似度, 预测用户对每一个项目的偏好度值。

(6) 对偏好度的预测结果进行排序列, 选择前 N 个项目推荐给用户, 从而得到相应的推荐结果。

基于用户历史行为的协同过滤推荐流程如图 2 所示。

4 仿真测试

4.1 数据集及仿真环境

为了分析基于用户历史行为的协同过滤推荐算法的有效性, 选择 Intel Core i3 6100(3.7 GHz/L3 3

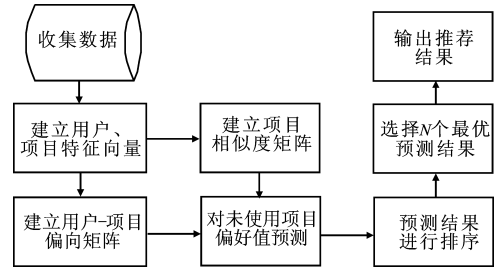


图 2 协同过滤推荐算法的工作流程

MB CPU, 16GB RAM, Linux 操作系统作为仿真平台, 采用 Matlab 2014 作为仿真工具, 实现协同过滤推荐算法的编程和测试实验。采用标准数据集采用 Book-Crossing 作为研究对象, 包括 278 858 个用户, 271 379 项目, 1 149 780 条记录, 由于数据规模大, 为此, 本文选择 10% 进行仿真实验, 并实验数据中选择 80% 作为训练样本, 测试样本为 20%。同时为了测试用户历史行为的协同过滤推荐算法的优越性, 采用当前典型的协同过滤推荐算法^[12-13]进行对比测试, 选择 F-measure 作为协同过滤推荐算法的性能评价指, 其计算公式如下:

$$\text{Precision} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \quad (13)$$

$$\text{recall} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \quad (14)$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

式中, $R(u)$ 和 $T(u)$ 分别为训练集和测试集的推荐列表^[14]。

4.2 结果对比

4.2.1 推荐效果比较

采用所有协同过滤推荐算法对训练样本进行学习, 建立相应的协同过滤推荐模型, 并通过测试样本对建立的协同过滤推荐模型性能进行分析, 统计它们的选择 F-measure, 结果如图 3 所示。对图 3 的统计结果进行对比和详细分析, 可以知道, 随着最优项目数 N 值的增加, 协同过滤推荐算法的 F-measure 随之增加, 在相同的 N 值, 相对于经典算法, 本文方法的 F-measure 都有了不同程度的提升, 这表明本文算法借助标签描述用户的偏好度以及用户历史行为获得更优的推荐结果, 能够更好地满足用户实际查询的要求。

4.2.2 冷启动和稀疏度条件下的性能比较

在不同 N 值和数据稀疏度条件下, 模拟冷启动和数据稀疏度条件, 并统计 F-measure 值, 得到

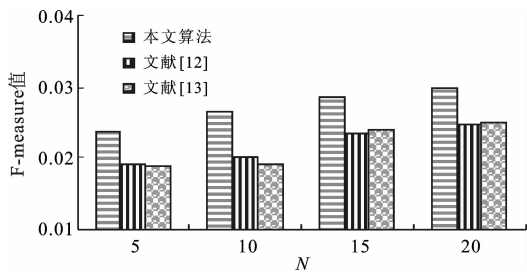


图 3 F-measure 值

实验结果如图 4 和图 5 所示,从实验结果可以清楚看出,本文算法综合利用用户历史行为,采用标签描述用户对项目偏好程度,建立了用户对项目喜的项目特征向量,得到了更优的项目相似度,从而实现精确的推荐,解决了传统算法存在的数据稀疏性和冷启动约束,大幅度降低了推荐误差,相对于对比算法,本文算法有效改善了推荐质量,而且优越性很明显。

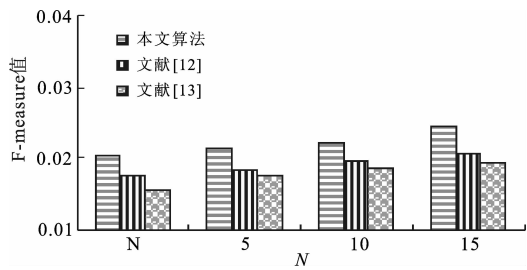


图 4 冷启动的 F-measure 值

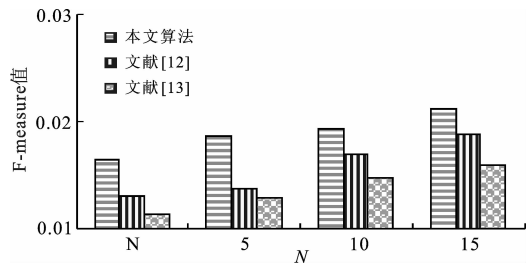


图 5 数据稀疏度的 F-measure 值

4.2.3 通用性测试

为了测试本文算法的通用性,选择当前另外两个标准数据集 MovieLens 和 Delicious 进行仿真实验,以进一步增强本文算法实验结果的说服力,得到的实验结果具体如图 6 和图 7 所示,从实验结果可以看出,采用本文算法实现协同推荐,同样可以获得高质量的推荐结果,而且用户的满意度比对比算法更高,实际结果表明,本文算法的适普性更强,实际的应用范围更加广泛。

4.2.4 算法的工作效率对比

对于大规模的数据挖掘问题,如电子商务的项

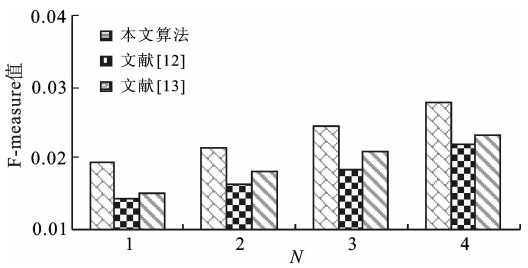


图 6 MovieLens 数据集的 F-measure 值

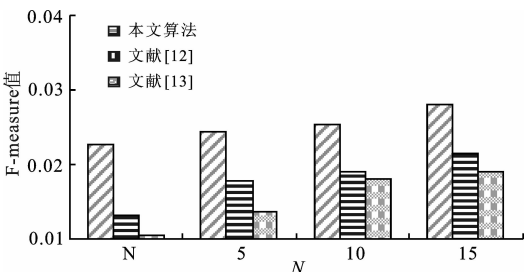


图 7 Delicious 数据集的 F-measure 值

目推荐问题,协同推荐的工作效率也是评价算法性能的关键指标,统计每一种算法的平均工作时间,实验结果具有见表 1,从表 1 可以发现,平均推荐时间最长的为文献[13]的算法,次之为文献[12]的算法,本文算法所花费的平均时间最少,这说明本文算法的推荐速度最快,有效提高了个性化推荐系统工作效率,可以应用于大规模数据的推荐中,实际应用价值更高。

表 1 平均工作时间对比				s
数据集	本文算法	文献[12]	文献[13]	
Book-Crossing	3.24	5.98	7.00	
Delicious	3.15	5.62	6.94	
MovieLens	3.39	6.10	6.99	

5 结束语

随着电子商务不断发展,协同推荐成为数据挖掘研究中的一个重要领域,为了更好提供个体化服务,获得理想的推荐结果,满足用户的实际应用要求,提出了基于用户历史行为的协同过滤推荐算法,该算法首先根据用户对项目喜好建立项目特征向量,然后根据项目特征向量估计项目相似度,并结合根据用户历史行为实现精确推荐,仿真对比实验结果表明,本文协同过滤算法提高了推荐效率,减少计算复杂度,而且获得更较高的推荐精度,推荐效果要明显优于传统协同过滤算法,实验结果证明了本文算法的有效性以及优越性。

参考文献:

- [1] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6): 734-749.
- [2] Bokde D, Girase S, Mukhopadhyay D. Matrix Factorization model in collaborative filtering algorithms: a survey [J]. Procedia Computer Science, 2015, 49(1): 136-146.
- [3] Yin F, Wang Z, Tan W, et al. Sparsity-tolerated algorithm with missing value recovering in user-based collaborative filtering recommendation [J]. 2013, 15(10): 4939-4948.
- [4] Bobadilla J, Serradilla F, Bernal J. A new collaborative filtering metric that improves the behavior of recommendersystems [J]. Knowledge-Based Systems, 2010, 23(6): 520-528.
- [5] Lin H, Yang X, Wang W, et al. A performance weighted collaborative filtering algorithm for personalized radiology education [J]. Journal of Biomedical Informatics, 2014, 51:107-113.
- [6] 施风仙,陈恩红. 结合项目区分用户兴趣度的协同过滤算法[J]. 小型微型计算机系统, 2012, 33(7): 1533-1536.
- [7] 韦素云,业宁,朱健,等. 基于项目聚类的全局最近邻的协同过滤算法[J]. 计算机科学, 2012, 39(12): 149-152.
- [8] Wu Yekui, Yao Jianrong, et al. Collaborative filtering based on multi-level item category system [J]. Journal of Convergence Information Technology, 2012, 7(7): 64-71.
- [9] 祝奇伟,陈家琪. 一种改进相似度计算方法的协同过滤推荐算法[J]. 信息技术, 2015, 22(3): 13-16.
- [10] 荣辉桂,火生旭,胡春华,等. 基于用户相似度的协同过滤推荐算法[J]. 通信学报, 2014, 35(2): 16-24.
- [11] 杨兴耀,于炯,吐尔根·依布拉音,等. 基于信任模型填充的协同过滤推荐模型[J]. 计算机工程, 2015, 41(5): 6-13.
- [12] 肖晓丽,钱娅丽,李旦江,等. 基于用户兴趣和社交信任的聚类推荐算法[J]. 计算机应用, 2016, 36(5): 1273-1278.
- [13] 冯永,陈显勇. 基于评分信息量的协同过滤算法研究[J]. 计算机工程与应用, 2013, 49(20): 198-201.
- [14] 方耀宁,郭云飞,扈红超,等. 一种基于 Sigmoid 函数的改进协同过滤推荐算法[J]. 计算机应用研究, 2013, 30(6): 1688-1691.

作者简介:

王志虎 男, (1974-), 博士研究生, 副教授. 研究方向为物联网、传感网、网络安全. E-mail: tigger_wang@163.com.

黄曼莹 女(壮族), (1978-), 副教授. 研究方向为计算机应用技术.

(上接第 131 页)

- [9] 王亚南,苏剑波. 基于图像合成的多姿态人脸图像识别方法[J]. 模式识别与人工智能, 2015, 28(9): 848-856.
- [10] 曾璐,鲁海荣,罗璐,等. 基于 RPCA 与低秩投影的有遮挡人脸识别 [J]. 计算机仿真, 2015, 32(10): 420-425.

作者简介:

张泊平 女, (1975-), 硕士, 副教授. 研究方向为计算机视觉、图像处理、虚拟现实. E-mail: iebopingzhang@163.com.