

基于项相关图的协同过滤算法

王丽萍

(光电子应用安徽省工程技术研究中心 铜陵 244000) (合肥工业大学管理学院 合肥 230009)

摘 要 在电子商务中,准确的推荐结果可以提高商品的成交量,从而给企业带来更多的收益。为了提高推荐算法的准确性,提出了一种基于项相关图的协同过滤算法。以商品作为顶点,以共同购买两个商品的用户数作为连接这两个顶点的边,将用户-项评分矩阵转换成一个项相关图。根据项相关图确定项与项之间的相似性,设计了一种综合了项相关图相似性和平均值相似性的推荐算法。实验结果表明,提出的推荐算法具有很好的预测准确性,明显优于相关的项推荐算法。

关键词 推荐系统,协同过滤,项相关图,随机游走

中图分类号 TP311 **文献标识码** A

Item Correlation Graph Based Collaborative Filtering Algorithm

WANG Li-ping

(Anhui Province Engineering Technology Research Center of Optoelectronic Technology Appliance, Tongling 244000, China)

(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract In e-commerce, accurate recommendation can improve the trading volume, and thus bring more profit for enterprises. In order to improve the accuracy of recommender system, this paper proposed an item correlation graph based collaborative filtering algorithm. This paper assumed the items as nodes, the number of people buying two commodities as the weight of the edge, and constructed an item correlation graph. According to the item correlation graph, this paper proposed a recommender algorithm that takes both item correlation graph based similarity and average similarity into consideration. Experiments show that the proposed algorithm has better prediction accuracy, and is more effective than related algorithms.

Keywords Recommender system, Collaborative filtering, Item correlation graph, Random walk

在电子商务中,系统中含有大量的用户、商品,以及用户对商品的历史购买记录,或者用户对商品的评价。推荐系统根据用户的历史购买记录或评价,对用户进行个性化的商品推荐,将那些用户在未来最可能购买的商品推荐给用户。这种推荐服务可以从大量的商品中提取用户最感兴趣的,并推荐给用户,从而解决了用户在面对海量商品信息时的困惑。根据用户对商品的评价矩阵,这种推荐服务既有益于购买方,又有益于销售方。购买方可以很容易地发现他们所喜欢的商品,销售方可以发现商品的潜在用户,并进行精准的广告投放。针对不同的商品类型,目前已经有了一些相应的推荐系统,如针对电影评分的 MovieLens^[1]、针对新闻的 GroupLens^[2]、针对音乐的 Ringo^[3]、针对笑话的 Jester^[4]。

推荐系统根据用户与系统的显示的或隐式的交互构建用户的档案。用户档案是对用户进行商品推荐的依据。构建用户档案的方法很多,最简单的方法包括用户在评价商品时考虑的商品特征,以及用户已经购买的商品。在这种情况下,用户档案是一个含有参数的模型,并且参数取决于购买者的行为。在真实的电子商务应用中,系统中往往含有千万级甚至亿级的用户数量,这些用户在商品选择时往往具有一定的关联性。本文根据用户的购买倾向建立一个商品间的相互关联

图,然后利用该图进行商品的评估。

1 相关工作

在推荐系统中,学者们提出了不同的方法用于提供推荐服务,如贝叶斯网络^[5]、支持向量机^[6]、因式分析^[7]和协同过滤^[1,3]等。在这些方法中,基于协同过滤的推荐方法由于简单高效被广泛采用。

协同过滤技术^[8]是推荐系统广泛采用的技术。协同过滤技术可以分为基于用户的协同过滤^[9]和基于项(商品)的协同过滤^[10]。当评估用户对某商品的评分时,基于项的协同过滤技术通过评价矩阵选取那些和待估商品相似的商品,然后利用他们的评价值对待估商品进行评估预测。在协同过滤方法中,推荐系统又可以分为 memory-based^[11]和 model-based^[12]两种。

在 memory-based 推荐系统中,我们根据一些启发式规则计算用户之间的相似性,然后选取一定数量的相似用户,并根据这些用户对商品的评价来评估待估商品的评价值^[13]。其中相似用户的选取可以是一个固定的数目,也可以设定一个门限值,即将超过门限值的用户作为相似用户。在衡量用户间的相似性时,常用的相似性度量方法有皮尔森相关系

数^[14]、余弦相似性、欧几里得距离等。memory-based 协同过滤算法的优点是设计简单,执行效率高。然而,其启发式函数的设计往往与具体的应用相关,需针对不同的应用场景设计不同的启发式函数;此外,相似用户数目的选择也需要根据实际情况来确定。

在 model-based 推荐系统中,假设用户根据对相似商品的喜好构成了一个个簇,然后通过相同簇内的用户对商品的喜好对待估商品进行评价。首先,对已知的用户评价矩阵进行建模,然后根据该模型对未知的用户-商品评价进行预测。在模型的建立中,可以采用聚类方法^[15]、分类方法^[16]、矩阵分解^[17]等方法。在 model-based 推荐系统中,建立的模型可能包含几个参数,在不同的应用背景下对参数调优是一项复杂的工作;此外,model-based 推荐算法只对推荐矩阵进行建模,没有很好地利用用户间的社会网络等其它有用的信息。

2 基于项相关图的推荐算法

2.1 问题定义

在一个推荐系统中,用户的集合为 $U = \{u_i | 1 \leq i \leq n\}$,项(商品)的集合为 $I = \{i_j | 1 \leq j \leq m\}$,用户对项的评价是一个 $n \times m$ 的矩阵 R ,其中 r_{ij} 表示用户 u_i 对项 i_j 的评价值。如果用户 u_i 没有对项 i_j 进行评价,那么 $r_{ij} = \text{null}$ 。由于系统中含有大量的商品,并且每个用户只对部分商品进行了评价,因此矩阵 R 是稀疏的。本文的目标是根据 R 中的已知信息,对未知数据进行预测。在给定某用户 u_i 时,我们对 u_i 尚未评分的商品进行预测评分,并对评分结果进行排名,将排名靠前的商品推荐给该用户。

2.2 项相关图

在 R 中,每一个 r_{ij} 表示用户 u_i 对项 i_j 的喜好程度,而项和项之间没有明显的关联关系。然而,由于不同的项可能吸引相同的用户群体,因此项之间有一种潜在的相关性。本文用 $U_{i,j}$ 来表示同时关注了项 i_i 和 i_j 的用户群体, $U_{i,j} \subseteq U$,于是有

$$U_{i,j} = \begin{cases} \{u_k | (r_{ki} \neq \text{null}) \wedge (r_{kj} \neq \text{null})\}, & i \neq j \\ \emptyset, & i = j \end{cases} \quad (2)$$

用 $c_{i,j}$ 来表示同时关注了项 i_i 和 i_j 的用户的数目,即 $c_{i,j} = |U_{i,j}|$,那么 C 是一个 $m \times m$ 的矩阵。如果令项集作为节点结合,用矩阵 C 作为邻接矩阵,那么就构成了一个项相关图,仍然用 C 来表示。该矩阵是对称的,在已知的训练集合中,每个用户对若干个项进行了评价,用 L_i 表示用户 u_i 已经评价的项的集合,并且反映了项两两之间的相互联系。 C 是一个项与项之间的相关矩阵,同时也可以看作是项与项之间的相关图。对于邻接矩阵 C ,其转移概率矩阵 P 为

$$p_{i,j} = \frac{c_{i,j}}{\sum_{j=1}^m c_{i,j}} \quad (2)$$

2.3 推荐算法

本文的思想是用项相关图预测用户对商品的偏好。矩阵 C 又可以看作一个图,在这个图中,节点表示项,边表示两个节点的共同喜好用户数。在项相关图 C 中,设计推荐算法时我们主要考虑了两个因素:扩张和衰减。首先,用户间的喜好存在着多样性,并且这种多样性会沿着网络传播。其次,在用户的喜好传播的过程中,传播的距离越远,这种影响就越小。

本文采用基于项相似性的推荐算法。在最近邻集合的选择中,我们采用两种相似性方法。给定用户 u_i 和项 i_j ,在对 $c_{i,j}$ 进行评价时,我们既考虑 u_i 的偏好项集 L_{u_i} ,同时又考虑偏好在项相关图上的传播。由于 PageRank^[18] 满足扩张性和衰减性,本文应用 PageRank 作为图上节点相似性的度量。

在项相似图 C 中,我们以项 i_j 作为出发点,然后在图 C 上随机游走。当随机游走的步数趋于无穷时,系统处于稳态,用随机游走停在某点的概率作为该点与出发点之间的相似性。该随机游走相似性向量 π_j 可通过如下公式计算:

$$\pi_j = (1 - \epsilon) P \pi_j + \epsilon \delta_j \quad (3)$$

其中, ϵ 为传输变量,表示随机游走以概率 ϵ 返回到出发点; π_j 为项集 I 的 m 维列向量,每一项 $\pi_j(k)$ 为项 i_k 与项 i_j 的相似性; δ_j 是一个 m 维列向量,并且有

$$\delta_j(k) = \begin{cases} 1, & k=j \\ 0, & k \neq j \end{cases} \quad (4)$$

在计算出其它项与项 i_j 之间的相似性 π_j 后,由于项 i_j 与项 i_j 的相似性为 $\pi_j(j)$,且 $0 < \pi_j(j) < 1$,我们利用 $\pi_j(j)$ 对向量 π_j 进行归一化,即

$$\pi_j(k) = \begin{cases} 1, & k=j \\ \frac{\pi_j(k)}{\pi_j(j)}, & k \neq j \end{cases} \quad (5)$$

由于用户 u_i 已经评价的项集为 L_i ,我们利用 L_i 对项之间的相似性进行规范化,可得项 i_j 与项 i_k 之间的相似性为

$$s'_{j,k} = \frac{\pi_j(k)}{\sum_{k=1}^m \pi_j(k) c_{j,k}} \quad (6)$$

我们利用项之间的相似性 π_j 对待估项 i_j 进行评价,可以得到用户 u_i 对项 i_j 的评估值

$$r'_{i,j} = \sum_{k=1}^m s'_{j,k} \cdot r_{i,k} \quad (7)$$

此外,对于用户 u_i 而言,其喜好也可以从已经评价的项集 L_i 中体现出来,即根据用户已经评价的项对未评价的项进行评估。该评价公式为

$$r''_{i,j} = \frac{1}{|L_i|} \sum_{k \in L_i} r_{i,k} \quad (8)$$

其中, $|L_i|$ 为集合 L_i 的势,且 $|L_i| = \sum_{j=1}^m c_{i,j}$ 。

综上所述,我们在计算用户 u_i 对项 i_j 的评估值时,既考虑了项之间的喜好关系传播(式(7)),又考虑了用户个人的喜好(式(8)),可得 $r_{i,j}$

$$\hat{r}_{i,j} = \alpha r'_{i,j} + (1 - \alpha) r''_{i,j} \quad (9)$$

其中, $\alpha (0 \leq \alpha \leq 1)$ 为偏好选择系数,其具体参数值可以根据具体的数据集进行选择。

3 实验

3.1 评价标准

本实验选用均方根误差(Root Mean Squared Error, RMSE)、准确率(Precision)和召回率(Recall)作为衡量预测准确性的标准。均方根误差的计算公式如下:

$$\sqrt{\frac{1}{k} \sum_{i,j} (\hat{r}_{i,j} - r_{i,j})^2} \quad (10)$$

其中, k 为评估的样本总数。

在信息检索中,我们通常用准确率和召回率评价算法的性能。由于系统中含有大量的项,推荐系统通常将算法计算

结果中排名靠前的 k 项返回给用户。然而在系统中,其真实的前 k 项可能并不是算法返回的这 k 个结果,它们之间的关系可以用表 1 表示。

表 1 算法准确性分析表

k (算法返回的前 k 项)		$I-k$
k' (真实结果中的前 k 项)	k_1	k_3
$I-k'$	k_2	k_4

准确率的计算公式如下

$$Precision = \frac{k_1}{k_1 + k_2} \quad (11)$$

召回率的计算公式如下

$$Recall = \frac{k_1}{k_1 + k_3} \quad (12)$$

3.2 实验设计

本实验采用的数据集为 MovieLens 100k 数据集,其中包含 943 个用户和 1682 个项。评分项为 100000 个,其中每个用户至少为 20 个项进行了评分。实验平台为一台个人 PC,操作系统为 Ubuntu10.04,编程语言为 Java,应用的推荐系统框架为 Mahout 0.7^[19]。

我们将本文提出的算法称为 ItemRWR。为了验证本文提出的算法的有效性,我们将 ItemRWR 与 5 种基于项的推荐算法进行了对比。这 5 种算法分别是皮尔森相关系数(Pearson)、欧几里得距离(Euclidean)、余弦相似性(Cosine)、斯皮尔曼相似性(Spearman)和 Tanimoto 系数。

3.3 实验结果

本小节首先确定算法在特定场合下的系数选择,然后应用选择好的参数对数据集进行性能评估。

(1) 参数选择

由于本文提出的 ItemRWR 算法含有偏好选择系数 α ,因此在进行推荐时需要首先确定算法的参数。我们分别选择数据集的 70%,80%和 90%作为训练集合,然后分析不同参数对评价准确性的影响。从图 1 可以看出,参数 α 在训练数据为 70%,80%和 90% 3 种情况下都是一个中间向下凹的曲线。当 $\alpha=0$ 时,算法采用的是完全的基于项相似性的预测方法;当 $\alpha=1$ 时,算法采用的是基于项相似图的预测算法。从图中可以看出,在综合两种相似性的情况下,算法的预测准确性进一步提高,并且在 α 在介于(0.7,0.8)之间时,算法的误差最小。此外,当完全采用项相似图时,算法的准确性高于完全的基于项相似性的预测方法。

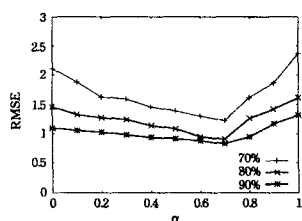


图 1 RMSE vs α

(2) 结果分析

本实验采用 $\alpha=0.75$,对比了本文提出的算法 ItemRWR 与 Pearson, Euclidean, Cosine, Spearman 和 Tanimoto 之间的性能。为了测试算法在不同训练集合下的性能,我们选择的数据集训练比例为 50%和 90%。首先,我们对比了这 6 种算

法的 RMSE,其结果见图 2 和图 3。从这两个图中可以看出,本文提出的 ItemRWR 算法的 RMSE 低于其它的算法。此外,当训练数据集很少(50%)时,ItemRWR 算法的 RMSE 波动不大,因此其能较好地处理推荐系统的冷启动问题(Cold start)。

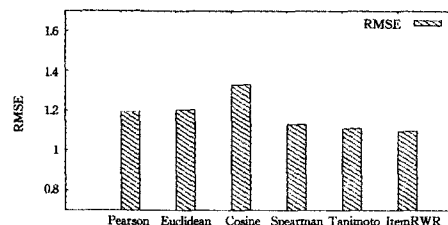


图 2 RMSE 对比图,训练数据集为 90%

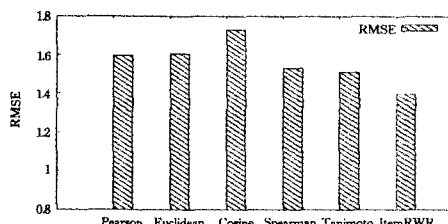


图 3 RMSE 对比图,训练数据集为 50%

其次,本实验评价了各种算法的 Precision 和 Recall,训练数据集仍然选择 50%和 90%的训练比例。在对用户进行产品推荐时,我们返回排名靠前的 10 个商品,即 top-10,并根据式(11)和式(12)计算了相应的准确率和召回率,实验结果见图 4 和图 5。图中结果表明,本文提出的 ItemRWR 算法与基于 Tanimoto 系数的推荐算法在准确率和召回率上结果相近,并且明显优于其它的相似性算法。

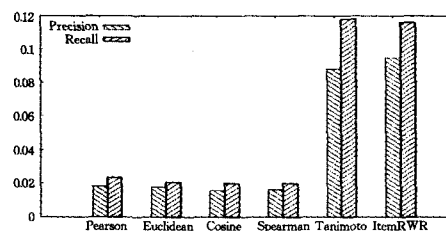


图 4 准确率和召回率对比图,训练数据集为 50%

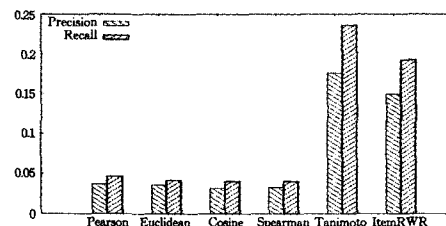


图 5 准确率和召回率对比图,训练数据集为 90%

结束语 推荐系统一直是电子商务平台中的热点研究问题,推荐结果的好坏直接影响着商家的经济收入水平。大量的用户购买记录或商品评价中往往隐含着商品之间的相关性。本文将商品看作节点,将共同购买两个商品的用户数作为这两个节点的边,建立了一个项相关图,并利用该相关图对商品的评价值进行预测。实验结果表明,本文提出的算法能有效地预测用户对未知商品的评价值,因此能很好地应用在真实的推荐系统中。

(下转封三)

在实际的农林虫害防治中,运用本文算法能够很好地识别害虫,将其中的目标区域(害虫)分离出来,从而实现仅对害虫进行药剂喷洒,防止了由于化学药剂的大面积使用而造成的环境污染问题。

本文算法的不足是目前更适合于背景颜色较少、背景色彩分布较均匀、背景和前景颜色相差较大的昆虫图片。下一步将针对各种复杂背景图片及背景与前景颜色比较接近的图片进一步优化和完善算法,以扩大算法的应用范围。

参考文献

- [1] 王江宁,纪力强. 昆虫图像分割方法及其应用[J]. 昆虫学报, 2011,54(2):211-217
- [2] 于新文,沈佐锐. 昆虫数字图像的分割技术研究[J]. 农业工程学报, 2001,17(3):137-141
- [3] 罗涛华. 基于自适应阈值的储粮害虫图像分割算法[J]. 武汉工业学院学报, 2006,25(1):5-8
- [4] 陈月华,胡晓光,张长利. 基于机器视觉的小麦害虫分割算法研究[J]. 农业工程学报, 2007,23(12):187-191
- [5] 卢亚玲. 基于迭代法和连通域的储粮图像粮虫分割算法[J]. 湖

北工业大学学报, 2007,22(6):14-16

- [6] 黄世国,周明全,耿国华. 快速几何可变形彩色昆虫图像分割算法[J]. 计算机应用, 2008,28(12):3144-3146
- [7] 王媛媛,彭延军. 流域分割方法在储粮昆虫图像中的应用[J]. 山东科技大学学报:自然科学版, 2007,26(2):79-82
- [8] 党培,谭联. FCM 和 LS-SVM 相结合的粮虫图像分割算法[J]. 电脑开发与应用, 2009,22(10):7-9
- [9] 陈婷婷. 采用模糊形态学的大田害虫图像分割[J]. 内蒙古农业科技, 2007(6):45-47
- [10] 赵娟. 基于 Gabor 小波和支持向量机的储粮害虫图像分割[J]. 电脑与信息技术, 2007,15(3):37-39
- [11] 程小梅,耿国华,周明全,等. 基于多特征的 EM 算法在昆虫图像分割中的应用[J]. 计算机应用与软件, 2009,26(2):20-22
- [12] 汪荣鑫. 数理统计[M]. 西安:西安交通大学出版社, 1986
- [13] 乐宁,梁学军,翁世修. 图像过渡区算法及其改进[J]. 红外与毫米波学报, 2001,20(3):211-214
- [14] 章毓晋. 过渡区和图像分割[J]. 电子学报, 1996,24(1):12-17
- [15] Yu Yuan-hui, Chang Chin-chen. A new edge detection approach based on image context analysis[J]. Image and Vision Computing, 2006,24:1090-1102

(上接第 282 页)

参考文献

- [1] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International Conference on World Wide Web, 2011:285-295
- [2] Miller J B N, Ried T, Konstan J A. GroupLens for Usenet: Experiences in applying collaborative filtering to a social information system[M]// From Usenet to CoWebs. Springer, 2013:206-231
- [3] 赵琴琴,鲁凯,王斌. SPCF:一种基于内存的传播式协同过滤推荐算法[J]. 计算机学报, 2013,36(3):671-676
- [4] Goldberg K, Roeder T, Gupta D, et al. Eigentaste: A constant time collaborative filtering algorithm[J]. Information Retrieval, 2001,4(2):133-151
- [5] 范波,程久军. 用户间多相似度协同过滤推荐算法[J]. 计算机科学, 2012,39(1):23-26
- [6] Gr M, Ar V C, Fortuna B V Z, et al. kNN versus SVM in the collaborative filtering framework[M]// Data Science and Classification. Springer, 2006:251-260
- [7] Hofmann T. Collaborative filtering via gaussian probabilistic latent semantic analysis[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2003:259-266
- [8] Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques[J]. Advances in Artificial Intelligence, 2009, 2009:421-425
- [9] Wang J, De Vries A P, Reinders M J. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C]// Proceedings of the 29th Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval. 2006:501-508

- [10] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering [J]. Internet Computing, IEEE, 2003,7(1):76-80
- [11] 杨震,赖英旭,段立娟,等. 邮件网络协同过滤机制研究[J]. 自动化学报, 2012,38(3):399-411
- [12] Mobasher B, Burke R, Sandvig J J. Model-based collaborative filtering as a defense against profile injection attacks[C]// AAI'06 Proceedings of the 21th National Conference on Artificial Intelligence. Volume 2, 2006:1388-1393
- [13] 刘文龙,张桂芸,陈喆,等. 基于加权信息熵相似性的协同过滤算法[J]. 郑州大学学报:工学版, 2012(5):027
- [14] Benesty J, Chen J, Huang Y, et al. Pearson correlation coefficient [M]// Noise reduction in speech processing. Springer, 2009:1-4
- [15] 李华,张宇,孙俊华. 基于用户模糊聚类的协同过滤推荐研究[J]. 计算机科学, 2012,39(12):83-86
- [16] Lee J, Jun C, Lee J, et al. Classification-based collaborative filtering using market basket data[J]. Expert Systems With Applications, 2005,29(3):700-704
- [17] Yu K, Zhu S, Lafferty J, et al. Fast nonparametric matrix factorization for large-scale collaborative filtering[C]// Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009:211-218
- [18] 尹航,常桂然,王兴伟. 采用聚类算法优化的 K 近邻协同过滤算法[J]. 小型微型计算机系统, 2013,34(4):806-809
- [19] 刘东辉,彭德巍,张晖. 一种基于时间加权和用户特征的协同过滤算法[J]. 武汉理工大学学报, 2012,34(5):144-148
- [20] 傅鹤岗,王竹伟. 对基于项目的协同过滤推荐系统的改进[J]. 重庆理工大学学报:自然科学版, 2010,24(9):69-74