

doi:10.3969/j.issn.1672-4348.2017.06.016

融合位置相似性度量的快消费品电商网站推荐算法

王晨阳, 刘垣, 郭李华, 肖琳

(福建工程学院 信息科学与工程学院, 福建 福州 350118)

摘要: 提出一种融合位置相似性度量的协同过滤推荐算法(CF-FLSM)。算法融合位置相似性度量进行加权计算用户间的兴趣相似度,从而为目标用户产生推荐结果。将 CF-FLSM 应用于一个具体的快消费品电商网站,得出的推荐结果与传统使用余弦相似性的协同过滤推荐算法(CF)相比,精确率和召回率分别提高了 3.74%和 3.91%。

关键词: 推荐系统; 协同过滤; 位置服务; 电子商务; 快消费品

中图分类号: TP391

文献标志码: A

文章编号: 1672-4348(2017)06-0586-05

An FMCG e-commerce website recommendation algorithm with the fusion of location similarity measurement

Wang Chenyang, Liu Yuan, Guo Lihua, Xiao Lin

(School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China)

Abstract: A collaborative filtering recommendation algorithm with the fusion of location similarity measurement (CG-FLSM) was proposed. This algorithm integrated location similarity measurement to compute preference similarity between users, thus generating recommendation results for the target user. Finally the CF-FLSM was applied to a specific FMCG e-commerce website, and experimental results showed that, compared with the traditional collaborative filtering recommendation algorithm (CF) using cosine similarity, its precision rate and recall rate were increased by 3.74% and 3.91% respectively.

Keywords: recommendation system; collaborative filtering; location-based service; e-commerce; fast moving consumer goods(FMCG)

近年来,互联网快消费品市场迅速发展,涌现出一大批快消费品电商网站,如阿里 1688 零售通、京东掌柜宝、掌合天下、惠民网、易酒批、便利宝、进货宝等。快消费品电商网站主要为广大下游便利店提供一站式的采购进货服务,所以它的用户通常就是便利店。面对海量的信息,高效的推荐算法无疑将对快消费品电商网站的成功起到至关重要的作用,它能够在信息过载的环境中帮助用户发现他们感兴趣的物品^[1]。协同过滤推荐系统是一

个已经在电子商务系统中广泛应用的一种个性化推荐算法^[2]。该方法首先找到和目标用户兴趣相似的用户集合,然后根据这个集合的评分数据向目标用户推荐商品。算法的关键是用户的兴趣相似性度量准则。对用户相似性度量准则的研究有很多,如文献[1-2]在计算用户相似性度量中引入了位置信息,但是位置信息只是作为商品的一个属性,对推荐的作用较弱;文献[3]将位置信息聚类为位置簇,虽然强化了位置信息的推荐作

收稿日期: 2017-08-03

基金项目: 福建省中青年教师教育科研项目(JA15342)

通讯作者: 王晨阳(1984-),男,福建莆田人,讲师,硕士,研究方向:大数据技术、电商推荐系统、搜索引擎技术。

用,但位置之间的距离所产生的推荐意义没有融入到推荐模型中。在快消品电商领域,因为快消品的流通具有区域的相似性,因此地理位置较近的便利店,他们的兴趣相似度也会相应更高。本文提出一种融合位置相似性度量的协同过滤推荐算法,根据用户的位置以及位置之间的距离算出用户的位置相似度,再与用户的兴趣相似性加权得到一个新的用户相似度,最后在一个真实的快消品电商网站数据集上验证了该算法的可行性。

1 传统的用户兴趣相似性度量方法

基于用户的协同过滤推荐算法是推荐系统中最为古老的算法,包括2个步骤:1)找到与目标用户兴趣相似的最近邻居集合;2)根据最近邻居集合中用户对商品的评分向目标用户进行推荐。其中步骤1)的关键是用户相似性的计算方法。最常用的有 Jaccard 相似性、余弦相似性、修正余弦相似性和相关相似性^[4-6]。

1.1 Jaccard 相似性

给定用户 u 和 v ,令 $N(u)$ 表示用户 u 有过评分的商品集合, $N(v)$ 表示用户 v 有过评分的商品集合。利用 Jaccard 公式计算用户 u 和 v 之间的相似度,Jaccard 系数定义为 $N(u)$ 与 $N(v)$ 交集的大小与 $N(u)$ 、 $N(v)$ 并集的大小比值,定义如下。

$$\text{sim}(u, v)^{\text{Jaccard}} = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}. \quad (1)$$

1.2 余弦相似性

给定用户 u 和 v ,令 \mathbf{A} 和 \mathbf{B} 表示用户 u 和 v 对 n 个商品的评分向量。使用两个向量之间夹角的余弦值度量用户间的相似性。公式如下:

$$\text{sim}(u, v)^{\cos} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|} = \frac{\sum_{i=1}^n \mathbf{A}_i \times \mathbf{B}_i}{\sqrt{\sum_{i=1}^n (\mathbf{A}_i)^2} \times \sqrt{\sum_{i=1}^n (\mathbf{B}_i)^2}} \quad (2)$$

1.3 修正的余弦相似性

余弦相似性度量未考虑到用户评分尺度问题,如在评分区间 $[1-5]$ 的情况下,对用户 i 来说评分3以上就是自己喜欢的,而对于用户 j ,评分4以上才是自己喜欢的。通过减去用户对商品的平均评分,修正的余弦相似性改善了用户评分尺度偏好问题。给定用户 u 和 v ,令 I 表示所有商品

的集合, \bar{R}_u 和 \bar{R}_v 分别表示用户 u 和 v 对商品的平均评分值, $R_{u,i}$ 和 $R_{v,i}$ 分别表示用户 u 和 v 对商品 i 的评分,如果用户 u 对商品 i 没有评分,则 $R_{u,i}=0$ 。使用修正余弦相似性度量用户 u 和用户 v 之间的相似性,公式如下。

$$\text{sim}(u, v)^{\text{acos}} = \sum_{i \in I} (R_{u,i} - \bar{R}_u) (R_{v,i} - \bar{R}_v) / \left(\sqrt{\sum_{i \in I} (R_{u,i} - \bar{R}_u)^2} \times \sqrt{\sum_{i \in I} (R_{v,i} - \bar{R}_v)^2} \right) \quad (3)$$

1.4 相关相似性

给定用户 u 和 v ,令 P 表示用户 u 和 v 都有评分的商品集合, $R_{u,i}$ 和 $R_{v,i}$ 分别表示用户 u 和 v 对商品 i 的评分, \bar{R}_u 和 \bar{R}_v 分别表示用户 u 和 v 对商品的平均评分值。使用 Pearson 相关系数 (Pearson Correlation Coefficient, PCC) 度量用户 u 和用户 v 之间的相似性,公式如下。

$$\text{sim}(u, v)^{\text{pcc}} = \sum_{i \in P} (R_{u,i} - \bar{R}_u) (R_{v,i} - \bar{R}_v) / \left(\sqrt{\sum_{i \in P} (R_{u,i} - \bar{R}_u)^2} \times \sqrt{\sum_{i \in P} (R_{v,i} - \bar{R}_v)^2} \right). \quad (4)$$

2 融合位置相似性度量的推荐模型设计

2.1 模型概述

融合位置相似性度量的协同过滤推荐算法 (collaborative filtering recommendation algorithm with the fusion of location similarity measurement, CF-FLSM) 的推荐流程如图1。算法分为3步:第1步:根据目标用户的位置搜索出周边的用户集合 $\text{near}(N)$,即距离最近的用户集合,以缩小用户范围;第2步:

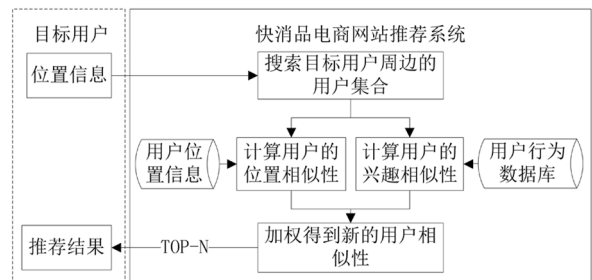


图1 CF-FLSM 的推荐流程

Fig.1 Recommendation process of CF-FLSM

在第 1 步中搜索出来的用户集合 $near(N)$ 中, 结合用户历史行为利用余弦相似性计算出用户之间的兴趣相似性, 结合用户的位置信息算出用户之间的位置相似性; 第 3 步: 按照适当的权重将用户之间的兴趣相似性和位置相似性进行合并处理, 得到一个新的用户相似性, 用于产生最终的推荐结果。

2.2 用户位置相似性度量方法

本文分析了某快消品电商网站在福州地区的注册用户(因为快消品电商网站服务的对象是便利店, 所以这里的用户就是便利店)的历史采购进货记录。图 2 是该地区注册用户的地理位置分布图。

随机选定 N (本文实验过程中 N 取 100) 个用



图 2 注册用户的地理分布图

Fig.2 Geographic distribution map of registered users

户, 分别计算出与目标用户距离(单位: km) $[0, 1]$ 、 $(1, 2]$ 、 $(2, 3]$ 、 $(3, 5]$ 、 $(5, 8]$ 、 $(8, 10]$ 内的用户有过共同评分商品的平均数量, 得出的具有共同评分的商品数量与用户之间距离的关系如图 3 所示。

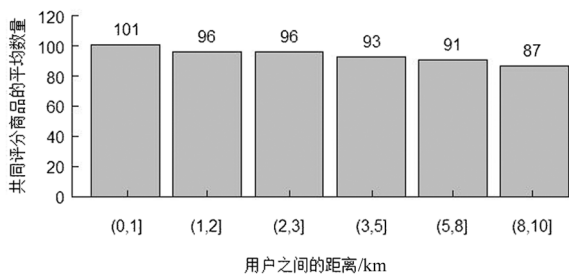


图 3 兴趣相似性与距离的关系

Fig.3 Relationship between interest similarity and distance

从图 3 可知, 用户之间共同评分的商品数量与用户之间的距离成反比。这说明在同一地区

内, 用户之间的偏好相似度对用户之间的距离是敏感的, 这也符合快消品电商市场销售的现状。因此, 本文提出如下度量方法计算用户之间的位置相似性:

$$\text{sim}(u, v)^{\text{dis}} = \frac{\overline{\text{dis}}}{\text{dis} + \text{dis}(u, v)} \tag{5}$$

其中, $\text{dis}(u, v)$ 表示用户 u 和 v 的距离, $\overline{\text{dis}}$ 表示所有用户之间距离的平均值。从该方法可以看出, $\text{sim}(u, v)^{\text{dis}}$ 的取值范围是 $(0, 1)$, 用户之间的距离越大, 则他们的位置相似性度量值越小。

2.3 加权合并处理

将用户位置相似性度量融合到传统的用户兴趣相似性度量, 两者进行加权合并处理得出一个新的用户相似性。

$$\text{sim}(u, v) = (1 - \lambda) \text{sim}(u, v)^{\text{cos}} + \lambda \text{sim}(u, v)^{\text{dis}} \tag{6}$$

其中 λ 表示 $\text{sim}(u, v)^{\text{dis}}$ 的权重, 可以根据实际情况适当调整 λ 的值。

2.4 推荐过程

算出用户之间的相似度后, 可以根据与目标用户兴趣最相似的 K 个用户向目标用户推荐商品。首先可利用以下公式预测出目标用户对与目标用户最相似的 K 个用户有过评分但目标用户未曾评分的商品的评分预测值。

$$p(u, i) = \sum_{v \in S(u, K) \cap N(i)} \text{sim}(u, v) R_{v, i} \quad (7)$$

其中, $S(u, k)$ 是根据公式 (6) 计算出的与目标用户 u 兴趣最相似的 K 个用户的集合, $N(i)$ 是对商品 i 有过评分的用户集合, $\text{sim}(u, v)$ 为用户 u 和用户 v 的相似度; $R_{v, i}$ 为用户 v 对物品 i 的评分。

最后将评分预测值最高的前 Top- N 项商品向目标用户做推荐。

3 实验

3.1 实验数据

选取某快消品电商网站在福州地区的 2016 年真实数据来验证上面提出的算法。该数据集在该地区的每个用户都维护一个固定的地位经纬度坐标, 各项数据统计值如表 1 所示。

表 1 实验数据集

Tab.1 Experimental dataset

有效用户数/个	商品数量/个	评分数/分
2 091	16 778	956 600

将数据集按照用户对商品评分的时间倒序排序选取 20% 作为测试集, 剩下的 80% 作为训练集。并随机选择 test_ U 家用户作为测试用户, 在训练集上计算用户的相似性, 最后在测试集上分别对每个测试用户进行商品推荐预测。

3.2 评测指标

网站在做商品推荐时, 通常是为用户提供一个个性化的推荐栏目, 就是推荐系统里的 Top- N 推荐。对 Top- N 推荐结果进行评测的指标通常是采用准确率和召回率。

准确率评测的是最终的推荐结果中有多少比例是正确推荐的。

$$P_{\text{precision}} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \quad (8)$$

召回率评测的是推荐正确的商品占实际用户喜欢的商品集合的多少。

$$R_{\text{recall}} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \quad (9)$$

其中 U 为测试用户集合, 就是 3.1 节中随机选择的测试用户数 test_ U ; $R(u)$ 表示对测试用户 u 在测试集上的推荐商品集合; $T(u)$ 为测试用户 u 在测试集上实际喜欢的商品集合。

3.3 实验结果

在 CF-FLSM 和 CF 的实验过程中, 测试用户数 test_ U 取值为 50, 用户最近邻的数目 near_ N 取值为 200, 用户之间的位置相似性度量权重 λ 取值为 0.5, 在推荐商品数目 Top- N 的不同取值下, 分别算出测试用户集的平均准确率和召回率如表 2、表 3 所示。

表 2 CF-FLSM 的推荐结果

Tab.2 Recommendation results of CF-FLSM

Top- N 取值	准确率/%	召回率/%
30	0.93	9.52
35	1.05	9.27
40	1.14	8.77
45	1.24	8.48
50	1.37	8.40
55	1.46	8.16
60	1.57	8.02

表 3 CF 的推荐结果

Tab.3 Recommendation results of CF

Top- N 取值	准确率/%	召回率/%
30	0.91	9.31
35	1.01	8.92
40	1.09	8.41
45	1.19	8.16
50	1.29	7.91
55	1.42	7.94
60	1.52	7.78

两者的实验结果的比较如图 4 所示, 本文提出的 CF-FLSM 相对于传统的协同过滤算法, 准确率和召回率分别提高了 3.74% 和 3.91%。

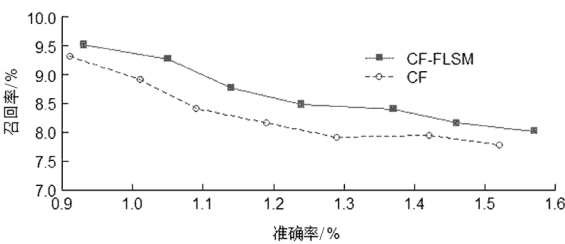


图 4 CF-FLSM 和 CF 的实验结果对比图
Fig.4 Comparison of experimental results
of CF-FLSM and CF

4 结论

调研国内众多知名快消费品电商平台,其注册用户通常是便利店,而便利店都有一个固定的地理位置。本文提出的融合位置相似性度量的协同过滤(CF-FLSM)推荐算法,能够有效利用位置以及位置之间的距离信息,从而提高快消费品电商平台商品推荐的准确率和召回率。但是 CF-FLSM 在推荐的过程中并未考虑时间的因素,而快消费品的流通活跃度通常和季节有关,比如啤酒饮料在夏季会相对比较活跃,而白酒红酒在冬季会相对比较活跃。在快消费品电商网站推荐算法中加入时间维度,是将来的研究方向。

参考文献:

[1] 申园园,余文.一种基于位置服务信息的移动推荐模型[J].计算机应用与软件,2016,33(12):203-206.

[2] 王付强,彭甫谔,丁小焕,等.基于位置的非对称相似度量的协同过滤推荐算法[J].计算机应用,2016,36(1):171-174,180.

[3] 郑慧,李冰,陈冬林,等.基于位置簇的移动生活服务个性化推荐技术[J].计算机应用,2015,35(4):1148-1153.

[4] 项亮.推荐系统实践[M].北京:人民邮电出版社,2012:45-64.

[5] 李春,朱珍民,高晓芳,等.基于邻居决策的协同过滤推荐算法[J].计算机工程,2010,36(13):34-39.

[6] 彭德巍,胡斌.一种基于用户特征和时间的协同过滤算法[J].武汉理工大学学报,2009,31(3):24-28.

[7] 马胡双,石永革,高胜保.基于特征增益与多级优化的协同过滤个性化推荐算法[J].科学技术与工程,2016,16(21):272-277.

[8] 张春永,陈群.一种基于 LBS 的移动个性化推荐系统[J].科学技术与工程,2011,11(30):7439-7442,7447.

[9] Cacheda F, Formoso V, Femandez D, et al. Comparison of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender systems[J]. ACM Transactions on the Web,2011,5(1):1-33.

[10] Liu H, Hu Z, Mian A, et al. A new user similarity model to improve the accuracy of collaborative filtering[J].Knowledge-Based Systems,2014,56(3):156-166.

(特约编辑: 黄家瑜)