ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 566－576
DOI:10. 16337/j. 1004-9037. 2019. 03. 020
© 2019 by Journal of Data Acquisition and Processing
http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

混合分层抽样与协同过滤的旅游景点推荐模型研究

李广丽[1] 朱 涛[1] 袁 天[1] 滑 瑾[1] 张红斌[2,3]

(1. 华东交通大学信息工程学院，南昌，330013;2. 华东交通大学软件学院，南昌，330013;3. 武汉大学计算机学院，武汉，430072)

摘 要：采用问卷调查与自动抓取相结合的方式，采集用户信息、用户评分等旅游数据，对数据做分层抽样，生成包含用户旅游喜好信息的“智慧旅游”数据集。围绕该数据集，预处理用户评分并执行基于用户聚类的协同过滤算法，以计算目标用户与聚类中心的相似性。结合分层抽样模型生成的旅游喜好信息，输出混合推荐列表。实验结果表明：相比基线，混合分层抽样与协同过滤的推荐模型对评分预测的均方根误差(Root mean square error，RMSE)和平均绝对误差(Mean absolute error，MAE)分别降低11.5%～64.9%和18.8%～47.7%。混合推荐的准确率和召回率相比基线也有较大程度提升，旅游景点推荐效果良好。

关键词：分层抽样；聚类；协同过滤；旅游景点；推荐模型

中图分类号：TP391　　**文献标志码**：A

Recommendation Model of Tourist Attractions by Fusing Hierarchical Sampling and Collaborative Filtering

Li Guangli[1], Zhu Tao[1], Yuan Tian[1], Hua Jin[1], Zhang Hongbin[2,3]

(1. School of Information Engineering, East China Jiaotong University, Nanchang, 330013, China;2. Software School, East China Jiaotong University, Nanchang, 330013, China;3. Computer School, Wuhan University, Wuhan, 430072, China)

Abstract: By combining the method of questionnaire survey and automatic crawling, a lot of useful tourist information such as users' personal information, users' ratings of tourist attractions and other tourism data are obtained. Based on the crawled tourism data, a hierarchical sampling method is applied in turn to generate the "Smart Travel" dataset which contains the important demographic information. Then a user clustering-based collaborative filtering algorithm is implemented to compute the semantic similarity between target user and each clustering center after the users' ratings of tourist attractions in the "Smart Travel" dataset is preprocessed. Finally, a hybrid recommendation list is generated by absorbing the demographic information obtained by the hierarchical sampling model. Experimental results show that compared with the traditional method, two evaluating indicators like the root mean square error (RMSE) and the mean absolute error (MAE) of the presented algorithm reduce 11.5%—64.9% and 18.8%—47.7%, respectively. Meanwhile, compared with the main baselines, the recommendation precision gets a large improvements as well as the recall rate and better recommendation results are obtained ultimately.

基金项目：国家自然科学基金(61762038,61861016)资助项目；江西省科技厅自然科学基金(20171BAB202023)资助项目；教育部人文社会科学研究规划基金(17YJAZH117,16YJAZH029)资助项目；江西省科技厅重点研发计划(20171BBG70093)资助项目；江西省社会科学规划项目(16TQ02)资助项目。
收稿日期：2018-01-18；**修订日期**：2019-04-09

Key words: hierarchical sampling; clustering; collaborative filtering; tourist attractions; recommendation model

引　言

随着互联网的不断发展,网络上的数据快速增长,人们正逐渐从信息匮乏的时代步入“信息过载”时代,此时,无论是信息消费者(网站用户)还是信息生产者(网站管理者)都面临很大的挑战。基于互联网搜索旅游信息已成为人们在出游前获取信息的最主要渠道之一。然而,伴随大量旅游网站的出现,人们常常被淹没于海量信息的搜索之中,却无法获取有价值的信息。推荐模型(系统)是解决“信息过载”问题,进而提升信息价值的有效方法。据统计,约有四分之三的旅游者在出游前都会搜索并查看旅游评论信息,以更好地规划他们的行程。

推荐模型可追溯到认知科学[1]、近似理论[2]以及信息检索[3]等领域的扩展研究。在20世纪90年代中期,推荐模型作为一门独立的学科被广泛关注,研究者主要从事具有显式评分的推荐问题研究,而推荐问题则转化为用户对未知物品的评分预测问题:基于用户历史评分,获取对物品的预测评分,进而向用户推荐评分最高的物品。目前,常用的推荐模型分三类:基于内容的推荐、协同过滤推荐和融合过滤。其中,协同过滤推荐的优势主要在于其能处理复杂的非结构化对象,且不需要领域知识就可以发现用户的新兴趣,同时推荐的个性化程度也较高。

“协同过滤”的概念由Goldberg等[4]提出,并应用于Tapestry系统,该系统仅适用于较小用户群,且对用户有较高要求(如用户要显式地给出评价)。作为协同过滤推荐模型的雏形,Tapestry展示了一种新的推荐思想,但它并不适合Internet环境。此后,出现了基于评分的协同过滤推荐模型。Resnick等[5]提出基于评分的协同过滤推荐模型GroupLens,向用户推荐新闻和电影。GroupLens的基本思想:分析用户偏好以形成推荐。它采集用户评分,评分值为1~5的整数,分值越大表明用户的偏好度越高。它通过计算用户间的评分相似性,选出相似性较高的一组用户来预测新用户对新物品的偏好。Konstan[6],Miller等[7]扩展GroupLens,使其成为一个基于开放式架构的分布式系统。自GroupLens之后,协同过滤理论取得较快发展,国际顶级会议、期刊发表的相关论文也逐年增加:Kim等[8]将社会网络分析(Social network analysis,SNA)与聚类技术相结合,通过反射隐藏的、用户社会群体的信息来提高推荐模型的预测精度;Nilashi等[9]利用分类与回归树(Classfication and regression tree,CART)和期望最大化(Expectation maximum,EM)等算法提出一种新的推荐方法。目前,大量Web网站也开始应用协同过滤算法向用户推荐个性化信息,如Amazon,Netflix和Last.Fm等,其中,Amazon对协同过滤推荐模型的研究已有十余年,借助推荐模型,Amazon收获了巨大的经济效益。此外,Video Recommender[10]和Ringo[11]也被认为是第一批能够进行自动预测的协同过滤推荐模型。

本文围绕旅游景点推荐这一热点问题展开研究,采用问卷调查与自动抓取相结合的方式,采集并制作涵盖若干属性维度的、全新的“智慧旅游”数据集。继而对其作分层抽样统计,获取用户的旅游喜好信息。最后,根据用户评分,设计基于用户聚类的协同过滤推荐算法,并融合旅游喜好信息,生成高质量的混合推荐结果。本文模型简单、有效,具备较高实用价值,它对于旅游景点推荐模型的开发与应用具有重要的借鉴意义。

1　旅游景点推荐模型设计

1.1　基本原理

协同过滤推荐模型指:分析其他用户已评分的物品(本文指旅游景点)来预测目标用户对某类物品的偏

好程度(兴趣)。因此,景点z对用户u的效用$ef(u,z)$取决于景点z对与u评分行为相似的其他用户$u_i \in U$(用户集合)的相关效用$ef(u_i,z)$。在旅游景点推荐过程中,协同过滤算法先要找出那些与用户u在旅游景点上兴趣相同的用户(即对同一旅游景点评分相似的用户)。然后,将这些用户喜好的景点推荐给用户u。故用户相似性度量是协同过滤算法的关键所在。综上分析,混合分层抽样与协同过滤的旅游景点推荐模型包括用户数据采集、分层抽样、用户聚类以及协同过滤并产生推荐等几大核心部件,系统框架如图1所示。

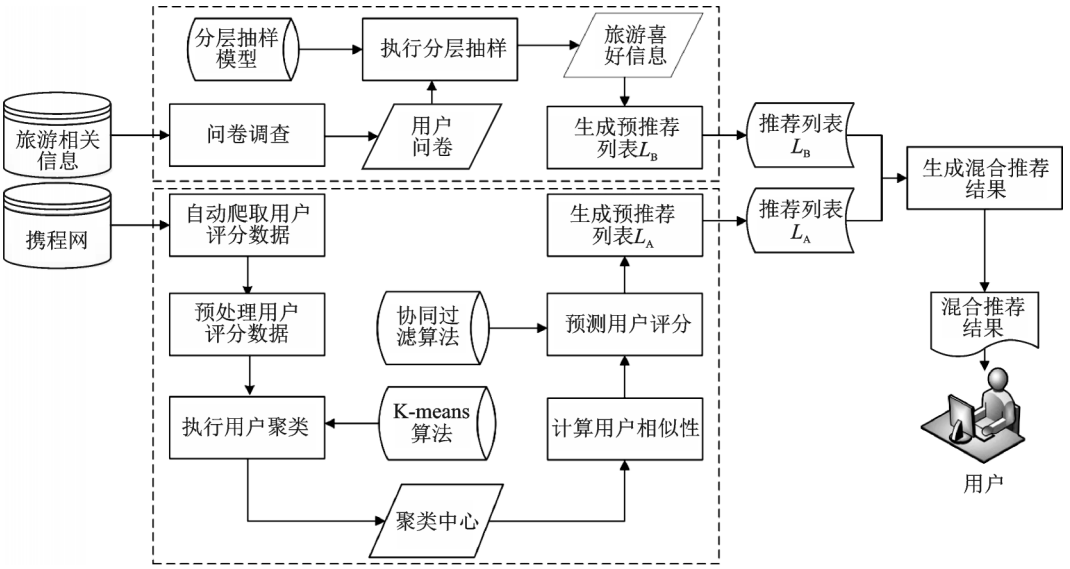


图1 混合分层抽样与协同过滤的旅游景点推荐模型

Fig.1 Recommendation model of tourist attractions based on hierarchical sampling and collaborative filtering

首先,设计调查问卷以收集人们的旅游喜好信息,并对其做分层抽样统计;其次,设计抓取规则,在“携程网”自动采集用户对不同旅游景点的评分,并对数据做预处理:用0~5表示用户对旅游景点的满意度(0分最低,5分最高);再次,采用K-means聚类算法处理用户数据,产生k个聚类中心,以刻画用户兴趣所属类别;基于协同过滤算法计算目标用户与各聚类中心的语义相似性,预测用户评分,形成预推荐列表L_A。最后,混合分层抽样结果输出的预推荐列表L_B,生成旅游景点的混合推荐列表“$L_A + L_B$”。

1.2 “智慧旅游”数据集建立

“智慧旅游”数据集的建立分3步:问卷调查、自动抓取和数据汇总。问卷采用“问卷星网”[12]制作,主要用于收集用户的旅游喜好信息,问卷中包括:用户性别、地区、年龄、学历、工作性质和月收入等基本信息及“出游季节”“兴趣类别”“出游方式”等旅游喜好信息;自动抓取通过规则抓取“携程网”上的景点图像、用户评分等信息:进入旅游景点用户点评页面,抓取用户对旅游景点的评分及景点图像,制作出涵盖“用户评分”“用户统计信息”“景点图像”等在内的数据集。最后,汇总调查问卷结果与自动抓取的数据,生成“智慧旅游”数据集。

1.3 基于分层抽样模型的用户喜好信息分析

分层抽样模型按规定比例从不同层中随机抽取样品。它按照一定规则将目标分成num个互不相交的子集。然后,在每个子集中独立抽样。每个子集称为层$(E_1,E_2,\cdots,E_{num})$,num个层合起来就是总体分布

$$E=\sum_{i=1}^{num}E_i \tag{1}$$

第1步:引入目标随机变量,以反映不同人群旅游兴趣的差异,目标随机变量有多个,如“出游季节”、“出游目的地”和“出游方式”等;

第2步:目标总体分层,把影响因素λ(如性别、地区等)作为分层原则,将旅游人群E根据其属性分为num层,第i($i=1,2,\cdots,num$)层中有E_i个人;

第3步:确定各层人数,若总体抽样人数为M,分num层抽样,第i层的人数为E_i,对于第i层的抽样人数:$X_i=M*E_i/E$,其中$i=0,1,2,\cdots,num$。

分层抽样模型根据目标总体的特征分布对其作层次化分类,以降低层内差异并增大层间差异,从而提高分类精度,更准确、客观地捕获用户的喜好信息。

1.4 基于K-means算法的用户聚类

聚类分析是把数据对象划分成若干子集的过程。每个子集是一个簇,簇中对象彼此相似,而与其他簇中对象不相似。簇的集合称作一个聚类。常用聚类算法可分为层次聚类和非层次聚类。K-means算法属于非层次聚类,它给定一组观测值($\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n$),每个观测值是一个$d$维向量,聚类算法将$n$个观测值划分成$t$($\leqslant n$)个集合$\boldsymbol{S}=(\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_t)$,以减少集合内的平方和(方差),即

$$\underset{S}{\operatorname{argmin}}\sum_{i=1}^{t}\sum_{a\in S_i}\left\|a-\mu_i\right\|^2=\underset{S}{\operatorname{argmin}}\sum_{i=1}^{t}\left|\boldsymbol{S}_i\right|\operatorname{Var}\boldsymbol{S}_i \tag{2}$$

式中:μ_i是$\boldsymbol{S}_i$中点的平均值,$\operatorname{Var}\boldsymbol{S}_i$表示$\boldsymbol{S}_i$的方差。

令用户集合是$\boldsymbol{U}=(u_1,u_2,\cdots,u_m)$的矩阵,评分景点集合是$\boldsymbol{Z}=(z_1,z_2,\cdots,z_n)$的矩阵,用户-景点评分矩阵是$\boldsymbol{G}$,$\boldsymbol{G}$中$d_{ij}$表示用户集合$\boldsymbol{U}$中第$i$个用户对景点集合$\boldsymbol{Z}$中第$j$个景点的评分,故用户基本信息矩阵为$\boldsymbol{S}=(\boldsymbol{U},\boldsymbol{Z},\boldsymbol{G})$。执行K-means聚类的前提:确定聚类数$k$和选定初始聚类中心。在收集用户数据时,根据景点性质将60个评分景点划分为k(5$\leqslant k\leqslant$10)个类别,每种类别各选1个典型用户(共k个)作为初始聚类中心,以完成K-means聚类。

算法1 基于K-means算法的用户聚类

输入:用户聚类数目k、用户基本信息数据源$\boldsymbol{S}=(\boldsymbol{U},\boldsymbol{Z},\boldsymbol{G})$、迭代次数$iter_num$

输出:用户聚类中心矩阵$\boldsymbol{D}$

1. 从k个类别用户中各选择1个典型用户作为初始聚类中心
2. repeat
3. for $i=1:m$ {
4. 计算用户u_i与聚类中心之间的用户相似性,见式(3)
5. 计算用户相似性中的最大值
6. 若相似性最大值出现在第j组,则用户u_i属于该类别
7. }
8. 同一类别中所有用户的评分均值作为新的聚类中心,更新聚类中心矩阵$\boldsymbol{D}$
9. until 各聚类中心不再发生变化或迭代次数$iter_num$到达

用户聚类中心$\boldsymbol{D}$是k行n列的矩阵,k代表k个用户聚类中心,n表示景点数。故$\boldsymbol{D}$中第i行、第j列元素D_{ij}表示第i个聚类中心中所有用户对景点j的评分均值。

采用式(3)计算用户之间的相似性。即基于皮尔逊相关系数度量用户之间的语义相关性

$$\operatorname{sim}(x,y)=\frac{\sum_{s\in S_{xy}}(r_{x,s}-\overline{r_x})(r_{y,s}-\overline{r_y})}{\sqrt{\sum_{s\in S_{xy}}(r_{x,s}-\overline{r_x})^2\sum s\in S_{xy}(r_{y,s}-\overline{r_y})^2}} \tag{3}$$

式中：r表示用户评分；$\overline{r_x}$和$\overline{r_y}$分别表示用户x和用户y对景点的平均评分；$\boldsymbol{S}_{xy}$表示用户x和y共同评分的景点集合，即

$$\boldsymbol{S}_{xy}=\{s\in\boldsymbol{S}|r_{x,s}\neq\varnothing\&r_{y,s}\neq\varnothing\} \tag{4}$$

1.5 基于用户聚类的协同过滤推荐算法

聚类算法输出用户聚类中心，即可计算目标用户与各聚类中心的相似性，从而得出目标用户的最近邻居，并对景点进行评分预测，最终输出预推荐列表。令给定目标用户c及最近邻居集合$\boldsymbol{KS}_c$时，对新旅游景点的评分预测[13]计算公式为

$$r_{cj}=\overline{r_c}+\frac{\sum\limits_{i\in KS_c}\mathrm{sim}(c,i)\times(r_{ij}-\overline{r_i})}{\sum\limits_{i\in KS_c}|\mathrm{sim}(c,i)|} \tag{5}$$

式中：$\overline{r_c}$，$\overline{r_i}$表示用户c和用户i对景点的评分均值，$\mathrm{sim}(c,i)$即式(3)计算的用户c与用户i之间的语义相关性。

算法2 基于用户聚类的协同过滤推荐算法

输入：目标用户的景点评分矩阵$\boldsymbol{R}$，用户聚类中心矩阵$\boldsymbol{D}$

输出：用户预测评分矩阵$\boldsymbol{D}'$、预推荐列表$\boldsymbol{L}_{\mathrm{A}}$

1. 根据目标用户的景点评分$\boldsymbol{R}$和用户聚类中心矩阵$\boldsymbol{D}$，采用式(3)计算目标用户与各聚类中心的相似性

2. 假设当前用户为u，初始化使其最近邻集合为空，即$\boldsymbol{KS}(u)=\varnothing$

3. 选择聚类中心内与用户u相似性最高的用户，并放入最近邻集合$\boldsymbol{KS}(u)$中，此时$\boldsymbol{KS}(u)\neq\varnothing$

4. 重复第3步，直到所有的相关用户都被加入到$\boldsymbol{KS}(u)$中

5. 根据$\boldsymbol{KS}(u)$中用户相似性大小，对集合中所有用户作降序排序，并将排序中的前l项作为用户u的最近邻$\boldsymbol{KS}(u)$

6. 根据式(5)计算评分预测矩阵$\boldsymbol{D}'$，并按照降序排列，取前v项生成预推荐列表$\boldsymbol{L}_{\mathrm{A}}$

7. 计算模型的RMSE值和MAE值，以评估推荐性能

2 实验结果及分析

2.1 数据集

“智慧旅游”数据集中的评论信息采集自“携程网”，共抓取了60个景点、5 000个用户的评分数据，可将60个景点分为8个类别(实验中也进行了验证)，分别是：海滨海岛、世界遗产、祈福拜佛、邮轮之旅、古镇游玩、亲子游、养生休闲和民俗体验。随机选取4 000个用户(80%)数据作为训练集，剩余1 000个用户(20%)数据作为测试集。数据集详见网址：https://drive.google.com/drive/mydrive?tdsourcetag=s_pctim_aiomsg。此外，编写调查问卷，基于Web吸引2 170位游客随机完成调查，收集用户基本信息，以完善“智慧旅游”数据集。最后，对评分数据进行预处理：将用户评分中的“很满意”“满意”“一般”“不满意”“很不满意”分别用“5”“4”“3”“2”“1”5个离散值表示。此外，用户的性别、地区、年龄、学历、工作性质和月收入等基本属性(详见表1)也进行相似的数字化处理，以便于后续实验。

2.2 分层抽样结果分析

采用1.3节的分层抽样模型对2 170位游客的调查答卷进行统计分析：从游客的性别、地区、年龄、学历、工作性质和月收入等方面对游客旅游兴趣进行分层抽样，各用户属性中随机抽出1 000名游客作为抽样代表，得到游客旅游兴趣的分层抽样统计表，如表1所示(用户偏好最大值用黑体加下划线给

表1 游客的旅游兴趣分层抽样统计表

Tab. 1 Hierarchical sampling statistics of tourist interest

%

类别		出游季节				兴趣类别								出游方式	
		春季	夏季	秋季	冬季	海滨海岛	世界遗产	祈福拜佛	邮轮之旅	古镇游玩	亲子游	养生休闲	民俗体验	个人出游	家庭或组团出游
性别	男	**67.39**	43.48	45.65	23.91	**67.39**	39.13	6.52	17.39	52.17	26.09	34.78	34.78	**67.39**	52.17
	女	64.81	31.48	**66.67**	18.52	**72.22**	51.85	20.37	25.93	61.11	20.37	38.89	44.44	37.04	**81.48**
地区	东部	**55.56**	33.33	7.80	20.00	**60.00**	31.11	17.78	13.33	55.56	13.33	31.11	31.11	40.00	**75.56**
	西部	**68.90**	43.84	52.45	33.98	43.21	20.21	59.84	37.36	43.21	33.58	56.45	**62.90**	55.40	**57.10**
	南部	**64.71**	35.29	52.94	11.76	52.94	35.29	23.53	**64.71**	35.29	29.41	23.53	35.29	47.06	**82.35**
	北部	50.00	50.00	**75.00**	37.50	37.50	25.00	12.50	25.00	**75.00**	12.50	25.00	37.50	37.50	**100.00**
	中部	**77.78**	59.26	70.37	37.04	**85.19**	48.15	14.81	25.93	70.37	14.81	51.85	48.15	**70.37**	66.67
年龄	20岁以下	42.11	**55.26**	50.00	18.42	**71.05**	44.74	10.53	15.79	50.00	15.79	23.68	31.58	34.21	**81.58**
	20～30岁	**65.12**	41.86	51.16	25.58	27.91	39.53	9.30	27.91	65.12	6.98	41.86	**74.42**	**60.47**	58.14
	30～40岁	**66.67**	33.33	16.67	33.33	33.33	**50.00**	16.67	33.33	0.00	0.00	0.00	33.33	**73.68**	57.10
	40～50岁	**50.00**	40.00	30.00	10.00	40.00	30.00	20.00	10.00	20.00	**70.00**	40.00	50.00	20.00	**90.00**
	50岁以上	33.33	33.33	**66.67**	0.00	57.10	15.70	9.05	39.68	42.62	12.52	**80.43**	42.62	0.00	**100.00**
学历	高中及以下	**83.33**	27.78	50.00	16.67	55.56	38.89	27.78	27.78	**61.11**	27.78	27.78	55.56	33.33	**100.00**
	专科	50.00	47.22	**52.78**	19.44	**63.89**	47.22	8.33	11.11	36.11	11.11	16.67	27.78	33.33	**80.56**
	本科	**73.91**	39.13	56.52	26.09	73.91	43.48	17.39	17.39	**82.61**	17.39	21.74	47.83	47.83	**73.91**
	硕士及以上	**56.52**	43.48	39.13	39.13	**69.57**	52.17	13.04	30.43	60.87	4.35	30.43	39.13	**60.87**	47.83
工作性质	企事业单位人员	50.00	28.57	**75.00**	14.29	**87.81**	35.71	0.00	0.00	42.86	0.00	28.57	14.29	35.71	**78.57**
	学生	48.61	**62.50**	61.11	50.00	**69.44**	50.00	19.44	31.94	66.67	20.83	37.50	40.28	50.00	**73.61**
	自由职业者	**45.68**	21.43	35.71	7.14	42.86	21.43	42.86	14.29	35.71	28.57	35.71	**50.00**	35.71	**85.71**
月收入	2 000元以下	**68.33**	45.00	60.00	26.67	**80.00**	43.33	13.33	30.00	65.00	20.00	31.67	33.33	53.33	**70.00**
	2 000～3 000元	**66.67**	41.67	58.33	0.00	58.33	25.00	33.33	16.67	33.33	16.67	41.67	**58.33**	25.00	**83.33**
	3 000～4 000元	70.00	30.00	**80.00**	30.00	**70.00**	60.00	0.00	20.00	50.00	0.00	20.00	30.00	40.00	**90.00**
	4 000～5 000元	40.00	**80.00**	20.00	0.00	60.00	40.00	40.00	20.00	**60.00**	0.00	0.00	40.00	20.00	**100.00**
	5 000元以上	38.46	30.77	53.85	**61.54**	**84.62**	30.77	30.77	15.38	69.23	23.08	30.77	38.46	**69.23**	61.54

注:因调查问卷中各旅游兴趣选项为多选题,所以表1中结果的百分比之和大于100%。

出),同时根据表中结果生成预推荐列表L_B。

由表1可以发现很多有趣的结论,这些结论与人们的客观认知较吻合:

(1) 大部分受访人群更偏爱于春、秋季出行,这应该缘于中国大部分地区在春、秋两季的风景非常吸引人;

(2) 大部分人群更注重于“家庭或组团出游”,这源于中国人强烈的家庭观念,即以家庭为单位出行,其乐融融;

(3) 中老年人更偏爱于“养生休闲”类景点,这类景点对个人体力没有过高要求,即休闲、旅游两不误,非常适合中老年人;

(4) 学生团体更愿意夏季出游,因为,中国的7,8月份是暑假,故各类围绕大、中、小学生的旅游活动、暑期项目非常普遍;

(5) 中、东部地区游客更喜欢“海滨海岛”,南方地区游客更青睐“邮轮之旅”,西部地区游客乐于“民俗体验”,而北方游客则倾向于“古镇游玩”,这与各地区游客的生活习惯、地理环境等都密不可分;

(6) 男性受访者倾向于“个人出游”,而女性受访者则更喜欢“家庭或组团出游”。

2.3 聚类实验结果分析

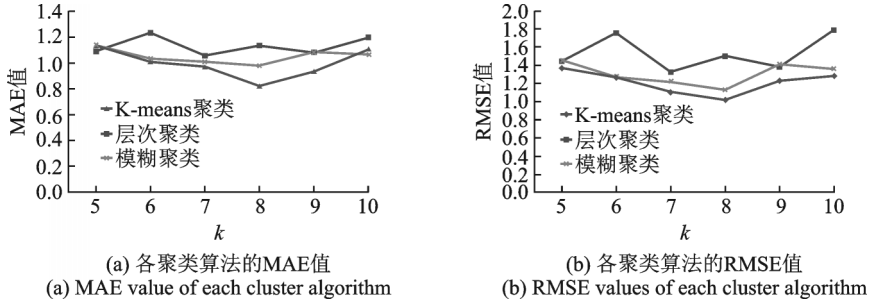
用户聚类是本文模型的核心,故需先评判聚类算法对推荐性能的影响。人们常通过预测用户对被推荐物品的评分来评判推荐性能。均方根误差(Root mean square error,RMSE)、平均绝对误差(Mean absolute error,MAE)是目前应用最广泛的评判评分精度的指标。本文利用RMSE,MAE对推荐模型作评测。若给定测试数据集T、用户-景点组合(u, z)、用户真实评分r_{uz}及推荐模型的预测评分$\hat{r}_{uz}$。预测评分和真实评分之间的RMSE和MAE分别为

$$\mathrm{RMSE}=\sqrt{\frac{1}{|T|}\sum_{(u,z)\in T}\left(\hat{r}_{uz}-r_{uz}\right)^2} \tag{6}$$

$$\mathrm{MAE}=\frac{1}{|T|}\sum_{(u,z)\in T}\left|\hat{r}_{uz}-r_{uz}\right| \tag{7}$$

由式(6,7)可知:RMSE,MAE分别计算预测评分和真实评分之间的均方根误差、平均绝对误差。RMSE、MAE值越小,推荐性能越好。实验时把60个旅游景点划分成$k(5\leqslant k\leqslant 10)$类,然后分别执行K-means聚类、层次聚类和模糊聚类等多种算法。接着,结合协同过滤算法完成旅游景点推荐,计算各算法在选取不同k值时所获取的RMSE和MAE值,实验结果如图2所示。

由图2可知:RMSE和MAE的变化趋势非常相似,这表明RMSE,MAE的评判本质很接近,它们可以同时使用,以综合评判推荐性能。然而,由于惩罚力度更大,RMSE结果的震荡幅度相对也更大;显然,K-means算法的性能优于其他两类算法。综合RMSE和MAE值,当聚类数目k=8时,K-means



(a) 各聚类算法的MAE值 (a) MAE value of each cluster algorithm

(b) 各聚类算法的RMSE值 (b) RMSE values of each cluster algorithm

图2 各聚类算法在选取不同聚类数时的RMSE和MAE值比较

Fig.2 RMSE and MAE values of each cluster algorithm by setting different clustering numbers

算法的推荐效果最好(这与人工设置吻合)。故选取K-means算法并设置聚类数k=8完成后续实验。

2.4 预测评分实验结果分析

本文选取基于模糊聚类的协同过滤算法、基于层次聚类的协同过滤算法实验,同时与基于用户的协同过滤(User-based collaborative filtering,UBCF)算法[14]、基于项目的协同过滤(Item-based collaborative filtering,IBCF)算法[15]以及基于杰卡德相似系数的推荐算法(LBCF)[16]等基线进行RMSE,MAE值比较,其中最近邻居值l=10。实验结果如图3所示。

由图3可知:本文算法的RMSE值与UBCF,IBCF和LBCF相比分别降低了61.0%,64.9%和59.4%,而与模糊聚类和层次聚类两算法相比分别降低11.5%和47.3%;其次,本文算法的MAE值与UBCF,IBCF和LBCF相比分别降低47.7 %,34.1%和28.3%,与模糊聚类和层次聚类两算法相比也分别降低18.8%和37.9 %。这表明:本文提出的协同过滤推荐算法能在一定程度上提高评分预测的精度,这有利于更好地生成旅游景点推荐列表,进而准确拟合用户的旅游偏好(兴趣)。主要原因分析:(1)"智慧旅游"数据集中良好的景点分类机制:8个类别、60个景点,聚类算法实验中也较好地验证了这一机制;(2)合理地选择聚类数(8个)及初始聚类中心(典型用户),并完成基于K-means的用户聚类。在聚类结果中,类内相似度较高,而类之间相似度较低;(3)皮尔逊相关系数(式(3))能较好地刻划用户之间的语义相关性。

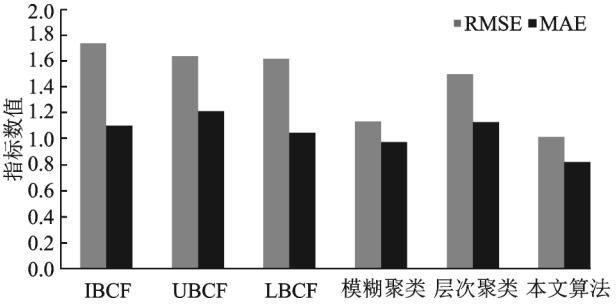


图3 不同算法的RMSE和MAE值

Fig.3 RMSE and MAE values of different algorithms

2.5 混合推荐结果分析

1.3节提出应用分层抽样模型获取用户的旅游喜好信息,将该算法命名为Our_HS。根据专家经验及交叉验证,设置各用户属性指标的相对重要性,即利用主观赋权评价法[17]对分层抽样结果设定权重。其中性别:5.1%;地区:13.75%;年龄:19.44%;学历:13.75%;工作性质:10.91%;月收入:37.06%,各项之和为100%,以建立Our_HS。此外,1.4节、1.5节提出基于用户聚类的协同过滤算法,该算法命名为Our_CF。由分层抽样结果可知(表1),用户属性不同,其旅游兴趣会有较大差异,而这些兴趣是提升推荐性能的重要依据。因此,本节混合协同过滤算法(Our_CF)输出的预推荐列表$\boldsymbol{L}_{\mathrm{A}}$与分层抽样模型(Our_HS)输出的预推荐列表$\boldsymbol{L}_{\mathrm{B}}$,生成混合推荐列表,该算法命名为Our_Mixed。

准确率(Precision)和召回率(Recall)这两个指标常被用来全面地评测推荐性能。而要计算这两个指标,先要确定被推荐的景点属于表2中的哪种情况。在表2中,"True-positive(tp)"表示被推荐且用户喜好的景点;"False-positive(fp)"表示被推荐用户却不喜好的景点;"False-negative(fn)"表示未被推荐且用户喜好的景点;"True-negative(tn)"表示未被推荐且用户不喜好的景点。

表2 被推荐景点的分类情况

Tab. 2 Classification results of each recommended item

类别	被推荐	未被推荐
用户喜好	True-positive (tp)	False-negative (fn)
用户不喜好	False-positive (fp)	True-negative (tn)

基于表2的定义,可得到准确率和召回率的计算公式为

$$\mathrm{Precision}=\frac{\sharp tp}{\sharp tp+\sharp fp} \tag{8}$$

$$\text{Recall}=\frac{\sharp tp}{\sharp tp+\sharp fn} \tag{9}$$

式中：$\sharp tp$表示被推荐且用户喜好的景点总数；$\sharp fp$表示被推荐且用户不喜好的景点总数；$\sharp fn$表示未被推荐且用户喜好的景点总数。假设给用户推荐N'个旅游景点，本节比较IBCF，UBCF，LBCF和Our_CF（协同过滤算法）、Our_HS（分层抽样模型）、模糊聚类、层次聚类以及Our_Mixed（混合推荐）等算法的准确率与召回率，并引入AP指标综合度量各算法优劣，即有

$$\text{AP}=\frac{\sum_{w=1}^{N'}\text{index}}{N'} \tag{10}$$

式中：index表示准确率/召回率值，$N'=10$。具体实验结果如表3和表4所示，每行最优值用黑体标出。

由表3可知：$N'=1$时，UBCF的准确率最高，即推荐少量旅游景点时，UBCF的效果较好。但推荐更多旅游景点时，由于用户评分的稀疏性越来越高，UBCF的效果逐渐降低。相反，Our_Mixed充分利用用户的旅游喜好信息，当$N'\geqslant 2$时，其推荐性能较优，且准确率稳步提高。Our_Mixed的AP值较同行的次优指标提升$(9.678-8.461)/8.461\approx 14.38\%$，即在混合推荐时，$L_{\text{A}}$与$L_{\text{B}}$这两个推荐列表有良好的互补性。基于AP值，所有模型的推荐性能（准确率）降序排列：Our_Mixed > Our_CF >

表3 各算法在不同N'下的Precision值

Tab. 3 Precision values of different algorithms under different N' %

N'	IBCF	UBCF	LBCF	模糊聚类	层次聚类	Our_CF	Our_HS	Our_Mixed
1	1.896	**10.092**	4.265	4.534	4.322	5.828	5.331	6.687
2	1.896	8.028	6.398	6.918	6.764	8.243	8.167	**9.531**
3	2.504	7.645	6.319	7.256	7.013	8.321	7.811	**9.425**
4	2.725	6.766	6.754	7.457	7.196	8.646	8.622	**10.005**
5	2.655	6.330	6.445	7.503	7.276	8.696	8.384	**9.593**
6	2.844	6.040	6.872	7.369	7.208	8.417	8.773	**10.162**
7	2.911	5.701	7.114	7.853	7.564	9.053	9.082	**10.485**
8	2.784	5.447	7.300	8.035	7.836	9.264	9.012	**10.021**
9	2.791	5.097	7.233	7.832	7.545	8.848	8.512	**9.977**
10	2.701	4.954	7.323	8.141	7.648	9.290	9.080	**10.897**
AP	2.571	6.610	6.602	7.290	7.037	8.461	8.277	**9.678**

表4 各算法在不同N'下的Recall值

Tab. 4 Recall values of different algorithms under different N' %

N'	IBCF	UBCF	LBCF	模糊聚类	层次聚类	Our_CF	Our_HS	Our_Mixed
1	0.113	0.623	0.255	0.274	0.264	0.488	0.326	**1.040**
2	0.226	0.990	0.764	0.843	0.792	0.664	0.796	**1.401**
3	0.368	1.415	1.132	1.298	1.173	2.228	1.117	**2.509**
4	0.659	1.669	1.613	1.735	1.687	2.524	1.803	**3.151**
5	0.792	1.952	1.924	2.106	2.504	2.987	2.108	**3.565**
6	1.019	2.235	2.462	2.401	2.476	3.217	2.343	**4.282**
7	1.217	2.462	2.971	3.024	2.984	3.700	3.143	**4.828**
8	1.330	2.688	3.480	3.307	3.497	3.843	3.431	**5.480**
9	1.500	2.830	3.877	3.908	3.940	4.727	4.809	**6.516**
10	1.613	3.056	4.358	4.669	4.381	5.507	5.563	**6.811**
AP	0.884	1.992	2.284	2.330	2.325	2.989	2.544	**3.958**

Our_HS ＞模糊聚类＞层次聚类＞UBCF ＞ LBCF ＞ IBCF,Our_CF 优于 Our_HS,这说明:Our_HS、Our_CF 算法的性能优于传统方法,而分层抽样模型输出的用户喜好信息在旅游景点推荐中发挥了重要作用。

由表4可知:当推荐更多旅游景点时,用户之间单一的相似性度量不能完全代表用户喜好,这迫切需要融入用户的旅游喜好信息。混合推荐的效果表现更佳,且召回率稳步提高。Our_Mixed 的 AP 值较同行的次优指标提升(3.958－2.989)/2.544≈32.4%,这进一步说明 L_A 与 L_B 之间具有良好的互补性。基于 AP 值,所有模型的推荐性能(召回率)降序排列:Our_Mixed ＞ Our_CF＞Our_HS＞模糊聚类＞层次聚类＞LBCF ＞ UBCF ＞ IBCF,Our_HS 优于传统算法,这也表明分层抽样模型输出的用户喜好信息在推荐中的重要性。混合分层抽样与协同过滤的旅游景点推荐模型能获取更优的推荐性能,这有助于提升旅游网站的影响力与竞争力。

2.6 算法的时间复杂度

除上述评测方式外,时间复杂度也是评价推荐模型优劣的一个重要指标。本节分析基于用户聚类的协同过滤算法的时间复杂度,并与传统算法进行比较。

传统的UBCF算法需要查找用户共同评分的景点,假设用户 u_1 偏好景点 z_1 和景点 z_2,用户 u_2 偏好景点 z_1,则算法可能会将景点 z_2 推荐给用户 u_2。若用户数量多,需逐个查找用户共同评分的景点。若有 m 个用户、y 个景点,该算法时间复杂度为 $O(m\times y)$,m 和 y 属于同一数量级且 $m>y$,则该算法的时间复杂度近似是 $O(m^2)$。

传统的IBCF算法中,令用户集合是 $\boldsymbol{U}=(u_1,u_2,\cdots,u_m)$ 的矩阵,评分景点集合是 $\boldsymbol{Z}=(z_1,z_2,\cdots,z_n)$ 的矩阵,计算景点相似矩阵时,任意两个景点都要计算其相似性,其时间复杂度为 $O(mn^2)$。向目标用户推荐时,先根据景点相似矩阵对景点的 n 个相似度排序,找出最近邻,故其时间复杂度为 $O(n\log n)$,然后通过景点的 l 个最近邻产生推荐,其时间复杂度是 $O(mn^2)+O(n\log n)+O(l)$。

而基于用户聚类的协同过滤算法仅需计算用户与各聚类中心的语义相似性。若有 k 个聚类中心,m 个用户,该算法的时间复杂度是 $O(k\times m)$。由于 $k\ll m$,它们不属于同一数量级,则该算法的时间复杂度近似是 $O(m)$。从时间复杂度衡量,本文算法优于UBCF及IBCF。

3 结束语

在旅游网站上,人们常常无法快速地找到感兴趣的旅游信息。因此,本文围绕推荐模型在旅游网站中的应用这一热点问题建立全新的“智慧旅游”数据集。提出混合分层抽样与协同过滤的推荐模型,其中,设计基于用户聚类的协同过滤算法,以计算目标用户与聚类中心的相似性,完成高质量的旅游景点推荐。实验表明:混合分层抽样与协同过滤的旅游景点推荐模型能够有效地提升推荐精度,并降低算法时间复杂度。这对于面向旅游景点的推荐模型的设计及应用都具有重要的借鉴意义。

未来主要工作:(1)结合景点的图像内容对用户建模,以弥补仅依赖用户评分的聚类方法;(2)运用 Relative Attribute 模型[18]进一步分析用户旅游兴趣的程度变化,更有针对性地推送景点信息;(3)运用深度学习模型[19]抽取图像、文本特征,并刻划用户特性,以改善推荐性能。

参考文献

[1] Rich E. User modeling via stereotypes[J].Cognitive Science, 1979, 3(4): 329-354.

[2] Powell M J D. Approximation theory and methods [M].Cambridge: Cambridge University Press, 1981: 35-40.

[3] Salton G. Automatic text processing[J].Science, 1970, 168(3929): 335-343.

[4] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of

the ACM, 2013, 35(12): 61-70.

[5] Resnick P, Iacovou N, Suchak M, et al. GroupLens: An open architecture for collaborative filtering of netnews[C]//1994 ACM Conference on Computer Supported Cooperative Work. Chapel Hill, North Carolina: ACM, 1994: 175-186.

[6] Konstan J A, Miller B N, Maltz D, et al.GroupLens: Applying collaborative filtering to usenet news[J].Communications of the ACM, 1997,40(3) : 77-87.

[7] Miller B N, Riedl J T, Konstan J A.Experiences with grouplens: Making usenet useful again[C] //Annual Conference on Usenix Winter Technical Conference. Anaheim, California: ACM, 1997: 219-231.

[8] Kim K, Ahn H. Recommender systems using cluster-indexing collaborative filtering and social data analytics[J]. International Journal of Production Research, 2017,55(17): 5037-5049.

[9] DMehrbakhsh Nilashi, Mohammad Dalvi Esfahani, Morteza Zamani Roudbaraki, et al. A multi-criteria collaborative filtering recommender system using clustering and regression techniques[J]. Social Science Electronic Publishing, 2016,3(5): 24-30.

[10] Hill W, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of use[C]// The SIGCHI Conference on Human Factors in Computing Systems. Denver, Colorado: ACM, 1995: 194-201.

[11] Shardanand U, Maes P. Social information filtering: Algorithms for automating "word of mouth"[C]// The SIGCHI Conference on Human Factors in Computing Systems. Denver, Colorado: ACM, 1995: 210-217.

[12] 长沙冉星信息科技有限公司.问卷调查.[EB/OL](2018-01-02). https://www.wjx.cn/[OL].

[13] 李涛,王建东,叶飞跃,等. 一种基于用户聚类的协同过滤推荐算法[J]. 系统工程与电子技术, 2007, 29(7): 1178-1182.
Li Tao, Wang Jiandong, Ye Feiyue, et al. Collaborative filtering recommendation algorithm based on clustering basal users[J]. Systems Engineering and Electronics, 2007, 29(7): 1178-1182.

[14] 黄传飞. 基于项目的协同过滤算法的改进[D].南昌:江西师范大学, 2015.
Huang Chuanfei. An improved item based collaborative filtering algorithm[D]. Nanchang: Jiangxi Normal University, 2015.

[15] 王成,朱志刚,张玉侠,等. 基于用户的协同过滤算法的推荐效率和个性化改进[J]. 小型微型计算机系统, 2016,37(3): 428-432.
Wang Cheng, Zhu Zhigang, Zhang Yuxia, et al. Improvement in recommendation efficiency and personalized of user-based collaborative filtering algorithm[J]. Journal of Chinese Computer System, 2016, 37(3): 428-432.

[16] 张晓琳,付英姿,褚培肖.杰卡德相似系数在推荐模型中的应用[J].计算机技术与发展, 2015,25(4): 158-161,165.
Zhang Xiaolin, Fu Yingzi, Chu Peixiao. Application of Jaccard similarity coefficient in recommender system[J]. Computer Technology and Development, 2015, 25(4): 158-161,165.

[17] 俞立平,潘云涛,武夷山. 科技教育评价中主客观赋权方法比较研究[J]. 科研管理, 2009,30(4): 154-161.
Yu Liping, Pan Yuntao, Wu Yishan. Comparing objective weighting with subjective weighting in sci-tech education institute assessment[J]. Science Research Management, 2009, 30(4): 154-161.

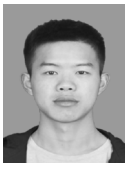
[18] Jayaraman D, Sha F, Grauman K. Decorrelating semantic visual attributes by resisting the urge to share[C] // IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio: IEEE, 2014: 1629-1936.

[19] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio: IEEE, 2016: 770-778.

作者简介:

李广丽(1977-),女,副教授,硕士生导师,研究方向:跨媒体检索、机器学习,E-mail:642908415@qq.com。

朱涛(1994-),男,硕士研究生,研究方向:推荐系统、机器学习。



袁天(1994-),男,硕士研究生,研究方向:图像理解、机器学习。

滑瑾(1995-),男,硕士生研究生,研究方向:推荐系统、机器学习。

张红斌(1979-),通信作者,男,副教授,博士,硕士生导师,研究方向:图像标注、机器学习;E-mail: zhanghongbin@whu.edu.cn。

(编辑:刘彦东)