第47卷第7期 中国科学技术大学学报 Vol.47,No.7
2017年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2017

文章编号:0253-2778(2017)7-0087-08

基于标签的个性化旅游推荐

李雅美,王昌栋

(中山大学数据科学与计算机学院,广州 510006)

摘要:旅游景点数量庞大,而用户本身旅游次数有限,所以用户旅游数据非常稀疏,进而影响了推荐结果的准确度.为了解决这一问题,从海量游记中提取与旅游景点密切相关的4个因素——地域、时间、主题、类型相关的特征标签,来丰富数据信息.一方面通过基于标签内容的方法为用户推荐感兴趣的景点;另一方面,用景点特征标签描述用户兴趣特征,根据用户兴趣标签找到相似用户群,通过协同过滤的方法为用户推荐感兴趣的景点.实验结果表明,基于标签的协同过滤算法较传统的协同过滤算法推荐准确率提高了63.7%,比基于景点热度的推荐算法提高了22.5%;基于标签内容的推荐算法比基于景点热度的推荐算准确率提高了27.6%.进一步,通过线性加权的方式混合两种算法,使两种算法优势互补,从而得到更好的推荐效果.最终使得基于标签的混合算法的准确率,比基于标签的协同过滤算法提高了61.3%,比基于标签内容的推荐算法提高了54.7%.旅游景点推荐准确度的提高,将带来更好的用户体验,使在线旅游网站更加具有竞争力.

关键词:推荐系统;个性旅游;数据挖掘;基于标签;协同过滤;基于内容;混合推荐

中图分类号:TP 391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.07.010

引用格式:李雅美,王昌栋. 基于标签的个性化旅游推荐[J]. 中国科学技术大学学报,2017,47(7):323-330.
LI Yamei, WANG Changdong. Tag-based personalized travel recommendation [J]. Journal of University of Science and Technology of China, 2017,47(7):323-330.

Tag-based personalized travel recommendation

LI Yamei, WANG Changdong

(*School of Data and Computer Science, Sun Yat-Sen University, Guangzhou* 510006, *China*)

Abstract: The disparity between the huge number of tourist attraction and the limited number of trigs made by tourists has resulted in the sparseness of tourist travel data, which seriously affects the accuracy of the recommendation results. To solve this problem, four kinds of tags area, time, topic, type were extracted from a mass of travel notes to enrich the data. On the one hand, travel attractions were recommended to users by tag-content-based recommendation algorithm. On the other hand, user interest features were described with attractions feature tags. Then, similar users were found according to the interest tags of users and attractions were recommended through collaborative filtering. The tag-based collaborative filtering algorithm by 63.7% compared with the collaborative filtering recommendation algorithm and by 22.5% compared with the attraction-heat-based recommendation algorithm. Tag-content-

收稿日期:2016-08-28;**修回日期**:2016-12-08

基金项目:国家自然科学基金(61502543),广东省自然科学基金杰出青年项目(2016A030306014),广东省自然科学基金博士启动项目(2014A030310180),中央高校基本科研业务费专项(16lgzd15)资助.

作者简介:李雅美,女,1994年生,硕士生.研究方向:数据挖掘、推荐系统. E-mail: liym0923@126.com

通讯作者:王昌栋,博士/副教授. E-mail: changdongwang@hotmail.com

based recommendation algorithm can improve the accuracy by 27.6% compared with the attraction-heat-based recommendation algorithm. The two algorithms were further combined with linear weight so that the two algorithms complement each other, resulting in better recommendation results. Our tag-based hybrid algorithm can make a significant improvement, i.e. increasing the accuracy by 61.3% over the tag-based collaborative filtering algorithm and 54.7% over the tag-content-based recommendation algorithm. The improvement of recommendation accuracy will enhance the user experience and make online travel websites more competitive.

Key words: recommendation system; personalized travel; data mining; tag-based; collaborative filtering; content-based; hybrid recommendation

0 引言

世界旅游组织的行业预测,未来20年世界旅游行业将持续保持较高的增长趋势.国际旅游人数预计平均增长率为每年3.3%.与此同时,大量的旅游用户也面临着一些旅游规划的问题.面对大量的景点选择,用户需要考虑诸多因素,如爱好、时间、价格、位置、交通等,这将耗费用户相当多的时间和精力去选择满意的景点.近年来,个性化推荐系统在各领域得到广泛应用[1-3].在线旅游网站通过有效利用推荐系统提高了用户服务质量,进而提高自身的竞争力.在线旅游推荐系统基于用户的历史数据分析用户的兴趣特点,可以从海量的旅游景点中为用户提供个性化的景点推荐服务,从而提高服务质量,提升在线旅游网站的竞争力.

当前,个性化旅游推荐一直面临着数据稀疏的问题.因为旅游景点繁多且在不断增加,但是用户旅游的频率相较于其他商品如音乐、电影等都低,用户旅游历史数据非常有限,造成数据稀疏.协同过滤算法是个性化推荐系统应用中非常经典和成熟的算法[4-5],它利用与目标用户兴趣相投的用户群体的喜好向目标用户作出推荐.协同过滤算法有一个普遍的的问题,即协同过滤在稀疏数据的情况下表现不佳.而由于旅游用户历史数据非常稀疏,因此用户相似度计算结果精确度不足,进而影响推荐效果.

为了克服数据稀疏带来的推荐效果不佳的问题,现有的研究中主要给出了以下的解决方法.一种方法是通过挖掘出更多的数据来提高用户相似度精度[6-7].文献[6]利用用户资料信息如用户的性别、年龄、工作信息等来计算用户相似度;也可以通过降维技术,如奇异值分解[7],来降低稀疏矩阵的维度,达到数据约减的目的,进而求出最好的低维近似.另一种方法是通过挖掘更多的相关信息来改善数据稀疏带来的问题[8-12].文献[9-12]通过分析用户在社交平台的数据信息推测用户的潜在兴趣.此外主题模型学习的方法也被用来解决数据稀疏问题[8-9].主题模型学习方法通过从用户的历史数据中提取出用户的兴趣,进而通过相似用户向目标用户作出感兴趣的推荐.如通过对用户在社交网站上上传的照片进行筛选和整理,用主题模型的方法对旅游主题进行分类,从而获得用户喜好.实际上,这种通过照片的相关信息进行主题提取和分类的方法是困难的.

为了更好地解决数据稀疏问题,我们利用更多景点自身的特征信息丰富数据.旅游景点本身具有一些特征,当我们在选择旅游目的地的时候,会考虑到景点特征相关的诸多因素,最普遍的考虑因素是景点位置和旅游时间[13].如十·一长假时,住在广州的客户可能会考虑这个时间适合旅游的景点,如内蒙古、新疆、九寨沟、三亚等,同时考虑到景点位置和假期长短,可能会选择去三亚.考虑景点位置其实本质上是考虑交通因素,我们更倾向于选择交通更便利花费更低的.由于交通因素复杂多变,而位置因素相对固定且实际交通成本一般与位置的距离呈正相关关系,因此我们用位置因素代替交通因素.除了景点位置和旅游时间外,对于景点的选择我们还会考虑到景点的类型和主题.例如,在十·一长假,一家三口出游,还会考虑景点的类型是否符合家庭成员的兴趣,同时也会考虑景点是否符合主题——家庭亲子游,因此我们选择利用景点自身的特征信息时,除了位置和时间信息外,还考虑景点的类型和主题信息.从而,我们得到关于景点的地域、时间、主题、类型四种因素的景点特征标签.

考虑多方面因素,可向用户推荐更合适的旅游景点.这些因素,是旅游景点本身的特征,如景点龙门石窟,它的地理位置是华中地区,适宜旅游时间是4—5月和9—10月,景点类型是人文,主题是历史.

这些特征我们称之为标签，当然一个因素不仅仅只有一个标签，如龙门石窟的主题可以同时是宗教.景点的类型指的是该景点自身的特性，如景点类型是沙漠、雪地、人文等.景点的主题是指旅游者在该景点旅游时更符合的旅游主题，如徒步、亲子、宗教、艺术等.类型和主题并不是绝对独立的.我们发现在各种在线旅游网站上，如蚂蜂窝(www.mafengwo.cn)、百度旅游(lvyou.baidu.com)等，有海量的旅游游记.为了得到这四种因素相关的景点特征标签，我们先从这些海量游记中提取景点信息构建景点特征标签；然后根据与景点四种因素相关的标签，结合用户历史旅游景点数据，我们可以用用户历史景点的特征标签描述用户的兴趣特征，而不是直接用历史景点描述用户兴趣.因为景点的数量多且会不断增加，而用户历史旅游景点有限，景点的标签数量很少.目前 4 种景点因素中每个因素的的标签数量在 20 到 40 之间，用标签描述用户兴趣特征将会有效改善原始数据稀疏的问题.

除了在稀疏数据上的推荐效果不佳外，协同过滤算法本身还有一些其他的不足[5,14].首先，协同过滤算法考虑了相似用户，在向目标用户推荐时需要依赖其他用户数据.其次，协同过滤算法因为没有相似用户去过一个新的景点，无法将该景点推荐给用户；但是协同过滤算法由于考虑的是相似用户，所以在一定程度上可以为目标用户推荐新颖的景点.另外一种推荐算法——基于内容的推荐算法同样也有优劣.首先，基于内容的推荐算法具有用户独立性，可以仅依赖目标用户的个人历史数据构建用户个人信息.其次，基于内容的推荐算法可以较好地为用户推荐新的景点.基于内容的推荐算法由于仅基于目标用户个人的信息作出推荐，具有过度个性化的缺点，在景点推荐中较难推荐新颖的景点.

基于以上原因，我们在运用标签的同时，线性加权混合协同过滤和基于内容的推荐算法，以达到更好的推荐效果.

本文的主要贡献如下：

(Ⅰ)对于个性化旅游推荐，我们提出了基于标签的协同过滤算法.我们将用户感兴趣的标签而不仅仅是用户历史旅游景点应用到了协同过滤算法中，改善在了因数据稀疏所造成的推荐效果不佳的问题.

(Ⅱ)对于景点标签，除了常见的位置(本文统称为区域)、时间标签外，我们也增加了与景点类型、主题因素相关的标签，用以提高推荐效果.

(Ⅲ)我们将协同过滤算法和基于内容的推荐算法结合.单纯的协同过滤算法和基于内容的推荐算法本身各有优劣，将两者结合，取长补短，进一步提高了推荐的准确度.

1　相关工作

近年来，关于旅游推荐算法的相关技术一直在发展，最经典的协同过滤算法得到广泛应用[5].协同过滤算法基于用户的历史旅游景点数据，计算出不同用户之间的相似度，进而根据目标用户的相似用户群的旅游数据向目标用户作出推荐.尽管协同过滤算法在个性化推荐系统中表现良好，但是该算法面临数据稀疏的问题，尤其是用户旅游数据信息相比于一般商品数据信息更是稀疏.为了解决这一问题，学者提出了基于主题模型的方法[9,13].文献[13]构建了“用户-区域-季节”的主题模型以解决数据稀疏性问题，但是除区域和季节因素外，景点的类型和主题也是参考因素. 本文增加了这两个相关因素以进一步丰富数据.除此之外，有学者也整理采集用户在社交平台的签到、照片等数据[9,11-12]，但是从这些数据中提取景点信息其分类是困难的.

为了改善数据稀疏问题造成的影响，本文利用从旅游游记中提取的信息构建景点本身的特征信息，为景点建立标签；同时结合用户的旅游数据信息，用景点标签来描述用户兴趣.此外，我们为景点设置的标签与四个因素相关：区域、时间、主题、类型.现有的旅游推荐系统的研究主要关注景点的区域和时间因素，但我们的研究发现景点的主题和类型也是景点的一个重要特征因素.

2　模型

2.1　景点和用户数据的挖掘与处理

2.1.1　景点分类因素和标签

旅游计划中，用户对旅游景点的选取，需要考虑很多因素.如景点的所属区域、当季是否为最佳旅游时间等.为了方便起见，我们称这两个因素分别为区域和时间.

在目前的旅游网站上，存在着海量的旅游游记.通过对游记的分析，我们发现用户在计划旅游选取景点时，还会考虑旅游的主题和景点的类型因素.关于主题，在计划旅游时，不同的用户会偏好不同的主题，如一些用户会偏好符合“徒步”主题的景点，而另一些用户可能会偏好符合“亲子”主题的景点.关于类型，是景点本身的特性.有的景点的类型是“古

镇”,如凤凰古镇、束河古镇;而香港和上海则可以归到“城市”类型中.关于主题中的“亲子”、“情侣”这样的词汇,我们称之为景点主题标签.同样地,“城市”、“古镇”则称为景点类型标签.

通过以上分析,我们将景点划分为区域、时间、主题、类型4种因素.此外,从海量的旅游游记中挖掘信息,我们提取出关于4种因素的标签.

2.1.2 用户历史数据和标签

得到用户历史数据之后,将会得到非常稀疏的用户-景点矩阵,这样稀疏的数据不适合用在协同过滤算法中.为了改善这种状况,我们提出设置用户标签.

在旅游推荐中,用户标签和景点标签相关.每个用户去过多个景点,这一系列景点有自己的特征标签,同样的标签可能会出现多次.我们用出现频率最高的多个景点特征标签来描述该用户的兴趣标签.

2.2 模型

我们的个性化旅游推荐系统模型分为两部分.一部分是离线模块,用于景点-标签集的构建.另一部分是在线模块,用于为用户做旅游景点的推荐.如图1、图2所示.

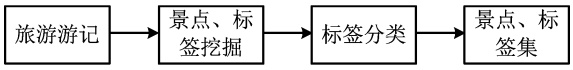

图1 离线模块

Fig.1 Off-line module

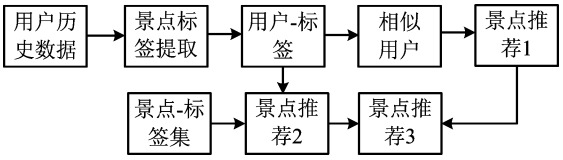

图2 在线模块

Fig.2 Online module

离线模块的目标是构建景点-标签数据库.首先从海量的游记中挖掘出景点和描述景点的特征标签,将景点和对应的标签联系起来;然后对标签进行分类;最后再将景点和对应的4种因素下的标签联系起来,构建景点-标签数据库.

在线模块为用户作出个性化的景点推荐.首先根据用户历史旅游景点数据,结合离线模块构建的景点-标签数据,将用户旅游景点的相应标签提取出来,用这些景点标签描述用户的兴趣特征,构建用户-标签矩阵;然后我们用两种方法推荐景点:一种方法是基于标签的协同过滤算法,这需要计算相似用户,得到景点推荐1;另一种方法是结合景点-标签数据的基于标签内容的推荐算法,得到景点推荐2;最后,我们用这两种算法加权后的混合推荐算法,进行更进一步的推荐,得到景点推荐3.

3 算法

3.1 基于标签的协同过滤算法

协同过滤算法是推荐系统非常经典成熟的算法,因此我们基于协同过滤算法作推荐.

首先,协同过滤算法收集用户的历史数据.普遍地,用户数据包含了用户对商品的显式打分,如对电影的评分.对于旅游推荐来说,由于旅游项目不同于一般产品的特殊性,如开销更大、时间安排难度更大等,因此用户旅游频率较于看电影、听歌、购物等是非常低的.数据本身少而稀疏,且具有有效评分的数据更少.为了充分利用现有的数据,我们认为用户去过某个景点即表示用户对该景点感兴趣,频次越高表明用户兴趣越高.这通常也是符合实际情况的.

其次,以用户为基础的协同过滤需要找到与目标用户兴趣特征相似的用户群,因此需要计算用户之间的相似度.用户旅游数据稀疏,在计算用户相似度的时候会使得计算结果精度不够,从而导致最终推荐结果的准确度不足.为了改善这一问题,我们通过从用户游记中挖掘到的景点信息即关于区域、时间、主题、类型四个因素的特征标签,来描述景点的自身特征.对于用户来说,用户对某景点感兴趣也暗示用户对该景点的特征(或者称为标签)是感兴趣的,因此我们可以利用这些标签与用户的历史旅游数据结合,将用户的历史景点信息转换为与四种因素相关的多个标签信息,并用这些标签信息来描述用户的兴趣,构建用户-标签矩阵进而计算用户间的相似度.

关于相似度的计算,目前使用较多的相似度算法有余弦相似性、调整余弦相似性、皮尔逊相关系数、欧氏距离等.经过多次实验,我们选择欧氏距离来计算用户相似度,公式表示如下:

$$\mathrm{Sim}(u_i,u_j)=\alpha-\sum_{\mathrm{type}}\left[w_{\mathrm{type}}*\sqrt{\sum_{p_n}(u_{i,\mathrm{type}}-u_{j,\mathrm{type}})^2}\right] \quad (1)$$

式中,u_i 和 u_j 分别表示用户 i 和用户 j;type表示4种景点因素的标号,取值1到4;w_{type} 表示每种因素(地域、时间、主题、类型)的权重;P_n 表示景点标号;$u_{i,\mathrm{type}}$ 表示用户 i 关于第type种因素的标签向量;$u_{j,\mathrm{type}}$ 表示第 j 个用户的关于第type种因素的标签向量;α 是常量,调整相似度为正值.

最后,在计算出用户相似度得到相似用户群后,

对目标用户的兴趣景点进行预测，做出 top-N 推荐，为目标用户推荐其最感兴趣的 N 个景点.

具体地，对于每一个目标用户，我们先找到和目标用户相似度最高的 k 个用户，然后将 k 个相似用户去过的景点中最受相似用户欢迎的 n 个景点推荐给目标用户.我们用景点热度表示景点的受欢迎程度，如下公式(2)所示.一个景点被参观的频次越多，则该景点的热度越高.特别地，在为目标用户推荐时，我们提到的某个景点热度是在相似用户群中该景点的热度，而不是在所有用户中该景点的热度.

$$\mathrm{Pop}(p)=\sum_{u_i} V_{u_i,p} \tag{2}$$

式中，p 表示景点数量；u_i 表示用户 i ；$V_{u_i,p}$ 表示用户 i 参观景点 p 的频次，如果没有用户去过景点 p ，则景点 p 的热度为 0.算法 3.1 描述了基于标签的协同过滤算法.

算法 3.1　基于标签的协同过滤算法.

输入：用户-标签矩阵，4 种因素相关的景点-标签矩阵
输出：为用户推荐的 n 个景点
step 1：
for i ←1 to userNumber
　for j ← 1 to userNumber
　　do
计算用户 i 与用户 j 的相似度；
　end
end
step 2：
for i ←1 to userNumber
　do
获取与用户 i 相似度最高的前 k 个用户；
end
step 3：
for i ←1 to userNumber
　do
将相似用户中热度最高的 n 个景点推荐给用户 i；
end

3.2　基于标签内容的推荐算法

该算法的思路是，对于目标用户而言，当景点在内容特征上与该用户以往感兴趣的景点相似时，用户将来很可能对这些景点同样感兴趣.不同于协同过滤算法，基于内容的推荐算法不需要通过相似用户来推荐，即向目标用户推荐时不需要其他用户的数据，没有冷启动问题和数据稀疏问题，因此我们尝试用该方法向目标用户推荐.

基于内容的推荐算法首先需要对项目的内容进行特征提取，本文的项目即景点.在 3.1 节中，我们已经构建了与 4 种因素(区域、时间、主题、类型)相关的景点-标签数据，本文的标签即景点的特征数据，且是 4 个维度的特征.

我们通过目标用户去过的景点的特征来学习目标用户的兴趣特征；通过用户的兴趣特征与候选景点的特征数据比较，为用户推荐相关性最大即用户最可能感兴趣的景点.

如下，算法 3.2 描述了基于标签内容的算法.

算法 3.2　基于标签内容的推荐算法.

输入：用户-标签矩阵，4 个景点-标签矩阵，权重 w
输出：为用户推荐的 n 个景点
step 1：
for i ←1 to userNumber
　for j ← 1 to 4
　　do
　获取用户 i 在第 j 个因素下最感兴趣的 q 个标签
　　end
end
step 2：
for j ← 1 to 4
　for i ←1 to userNumber
do
为用户 i 推荐在第 j 种因素下感兴趣的的 n 个景点
　end
end
step 3：
for i ←1 to userNumber
　do
权重 w 作用于四种因素下的推荐结果上，得到最佳的 n 个推荐景点
end

3.3　加权的混合推荐算法

基于标签的协同过滤算法和基于标签内容的推荐算法各有优劣.

首先，向目标用户推荐景点时，基于标签的协同过滤算法考虑了相似用户，推荐依赖于其他用户；而基于标签内容的推荐算法具有用户独立性，以目标用户个人的历史数据构建用户特征信息.其次，基于标签的协同过滤算法无法将一个新的景点推荐给用户，因为没有相似用户去过该景点；而基于标签内容的推荐算法却可以较好地为用户推荐新的景点.此外，基于标签的协同过滤算法由于考虑了相似用户的兴趣，所以在一定程度上可以向目标用户推荐新颖的景点；而基于标签内容的推荐算法仅基于目标用户个人的信息作出推荐，具有过度个性化的缺点，在景点推荐时较难推荐新颖的景点.基于以上考虑，

为达到更好的推荐效果，我们采用线性加权的方式结合这两种算法做出景点推荐，公式表示如下：

$$\mathrm{recom3}_U = w_1 * \mathrm{recom1}_U + w_2 * \mathrm{recom2}_U \quad (3)$$

式中，U 表示用户集；$\mathrm{recom3}_U$ 表示在加权混合推荐算法下对用户的景点推荐结果；$\mathrm{recom1}_U$ 表示在基于标签的协同过滤算法下对用户的景点推荐结果；$\mathrm{recom2}_U$ 表示在基于标签内容的推荐算法下对用户的景点推荐结果；w_1，w_2 分别表示两种推荐结果线性加权的权重.

4 实验结果与分析

下面将基于真实的数据评估算法性能.本文提出的个性化旅游推荐算法主要在3个方面做了改进：

(Ⅰ)基于标签的协同过滤算法，在计算用户相似度时利用用户-标签数据而不是用户-景点数据，用户-标签数据相较于用户-景点数据的稀疏度大大降低，且用景点本身的特征标签来描述用户兴趣特征，改善了因数据稀疏造成的推荐效果不佳的问题.

(Ⅱ)关于景点标签，当前的研究中普遍提取与景点相关的时间、位置信息.本文算法中，除了利用这两点因素外，还结合了景点类型、主题相关的标签，用更丰富的信息达到提升推荐效果的目的.

(Ⅲ)在构建了景点-标签数据后，直接利用基于标签内容的推荐算法作出推荐.同时，因为协同过滤算法和基于标签内容的推荐算法各有优劣，我们将两者通过线性加权的方式结合，取长补短，进一步提高推荐的准确度.

基于以上几个方面，我们设置了对比实验，主要涉及以下几种推荐算法.

基于景点热度的推荐算法[15]：该算法先利用所有用户的历史景点数据计算出景点热度；然后将热度最高的 N 个景点推荐给目标用户.在购物网站、视频网站等会有热门项目推荐，热门项目即多数用户喜欢的项目，如某件商品、某部电影.将热门推荐应用到旅游推荐中，即将热度高的景点推荐给用户.如十·一国庆长假，若此时全国最热门的景点有长城、故宫、泰山、华山等，那么基于景点热度的推荐算法将会给用户优先推荐这些热门景点.该算法考虑了所有人的兴趣喜好，在线旅游网站如百度旅游、蚂蜂窝等都有热门景点推荐，但是对每个用户的个性化信息利用不足.

传统的协同过滤算法[5]：该算法利用用户的景点数据，基于用户历史旅游景点计算用户间相似度，进而向目标用户推荐.如我们计算出与目标用户兴趣相似的用户群最喜欢去的景点有布达拉宫、泸沽湖，那么传统的协同过滤算法将会推荐目标用户十·一长假游玩以上景点.该算法没有利用更多的数据处理方法，没有解决数据稀疏问题.

基于标签的协同过滤算法：通过为景点构建的四种因素相关的景点-标签数据来表示景点特征，进而通过用户历史旅游景点的特征提取出用户兴趣特征标签，基于用户标签计算用户间相似度，进而向目标用户推荐.如我们提取出用户的兴趣特征标签是国庆、秋天、西南、古镇、徒步、文化等，那么基于标签的协同过滤算法将通过标签计算出目标用户的相似用户群；再将相似用户群最喜欢去的而目标用户未去过的景点如大理古城推荐给目标用户.

基于标签内容的推荐算法：通过计算得到用户兴趣特征标签，进而得到用户最感兴趣的标签；再将这些标签下相关的热门景点推荐给目标用户.如我们提取出用户的兴趣特征标签是秋天、佛教、徒步、文化等，那么基于标签内容的推荐算法将为目标用户推荐最具有这些标签特征的景点如嵩山等.

基于标签的混合算法：将基于标签的协同过滤算法和基于标签内容的推荐算法通过线性加权相结合，进而为用户做推荐.

基于时间、区域的推荐算法：该方法运用景点标签时，仅用时间、区域两种因素相关的标签，与本文提出的其他几种方法做对比.

我们先在20组不同小数据集上进行实验，每种算法在小数据集上的实验结果与我们预期的结果一致.

4.1 数据集

本文的数据主要由两部分组成.一部分是游记，另一部分是用户数据.我们抓取了来自在线旅游网站——百度旅游(lvyou.baidu.com)和蚂蜂窝(www.mafengwo.cn)上的9 857篇游记，用来提取景点特征标签.四种因素(区域、时间、主题、类型)相关的标签共计106个.用户数据来自于百度旅游，由1 335个用户的88 395条用户记录构成，时间跨度为2012年至2015年.根据时间顺序将数据集分为训练数据和测试数据.具体地，将这些数据分为5组，每次任选其中两组作为训练数据和测试数据做交叉实验.

4.2　性能评估

我们用准确率[16]评估本文提出的 3 种算法以及其他两种算法的性能.

$$\text{Precision}=\frac{\sum_{u\subseteq U}|R(u)\cap T(u)|}{\sum_{u\subseteq U}|R(u)|} \quad (4)$$

式中,$R(u)$是根据用户在训练集上的行为向用户作出的推荐;$T(u)$是用户在测试集上的行为列表;Precision 即准确率,表示准确推荐的景点在所有推荐景点中的比率.推荐结果的准确率越高表明算法的推荐性能越好.

4.3　性能比较

在不同的景点推荐数目下,我们测试了五种推荐算法的准确率.

表 2 至表 6 中,N 表示为用户推荐景点的数量.算法 a 至算法 e 分别表示如下:

算法 a:基于景点热度的推荐算法;

算法 b:传统的协同过滤算法;

算法 c:基于标签的协同过滤算法;

算法 d:基于标签内容的推荐算法;

算法 e:基于标签的混合推荐算法.

4.3.1　在四种因素标签下各算法的性能比较

表 1 和图 3 数据表示五种算法的准确率.需要注意的是,对于算法 c,d,e 来说,这三种算法的标签均是关于 4 种因素(地域、时间、主题、类型)的.

表 1　4 种因素的标签下各算法的性能比较

Tab.1　Performance comparison of algorithms with four kinds of tags

N 准确率算法	a	b	c	d	e
5	0.048 2	0.039 8	0.066 3	0.078 3	0.120 5
10	0.047 0	0.037 3	0.062 0	0.058 6	0.096 4
15	0.043 8	0.033 3	0.051 8	0.049 4	0.084 3
20	0.042 8	0.028 3	0.046 7	0.047 3	0.077 7
25	0.040 7	0.028 0	0.045 8	0.050 2	0.061 0

由图表数据可知,相较于其他 4 种推荐算法,混合推荐算法的准确率是最高的,最高达 0.120 5,其次分别是基于标签内容的推荐算法、基于标签的协同过滤算法、基于景点热度的推荐算法、传统的协同过滤算法.与传统的协同过滤算法相比,本文提出的基于标签的协同过滤算法推荐结果的准确率有了明显的提高.

基于标签的协同过滤算法平均准确率为 0.054 5,比基于景点热度的推荐算法提高了22.5%,比传统的协同过滤算法提高了 63.7%.由此可见,个性化推荐中加入标签信息,推荐性能有显著提升.

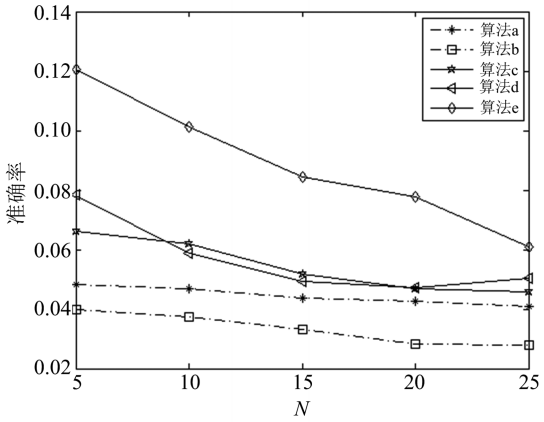

图 3　四种因素相关的标签下各算法的性能比较

Tab.3　Performance comparison of algorithms with four kinds of tags

基于标签内容的推荐算法平均准确率为 0.056 8,比基于景点热度的推荐算法提高了27.6%,比传统的协同过滤算法提高了 70.6%.

混合推荐算法的准确率比基于标签的协同过滤算法提高了 61.3%,比基于标签内容的推荐算法提高了 54.7%.由此可见,将两种算法混合的推荐结果性能比单个推荐算法的推荐性能显著提高.

4.3.2　在两种因素标签下各算法的性能比较

表 2 和图 4 数据中,算法 c,d,e 中使用的标签是仅关于地域和时间两种因素的,这两种因素是现有的多数研究中主要考虑的因素.

表 2　两种因素相关的标签下各算法的性能比较

Fig.2　Performance comparison of algorithms with two kinds of tags

N 准确率算法	a	b	c	d	e
5	0.048 2	0.039 8	0.057 8	0.062 7	0.092 8
10	0.047 0	0.037 3	0.049 4	0.053 6	0.082 5
15	0.043 8	0.033 3	0.042 6	0.048 4	0.073 9
20	0.042 8	0.028 3	0.039 8	0.046 7	0.070 8
25	0.040 7	0.028 0	0.037 6	0.041 8	0.065 5

由表 2 和图 4 可以看出,基于标签的混合推荐算法的性能依然远高于其他 4 种算法.

4.3.3　在两种和 4 种因素的标签下推荐性能比较

下面我们主要讨论分别应用两种和 4 种因素相关标签的推荐性能的比较.

由图 5、6、7 可以看出,我们在基于标签的协同过滤算法(算法 c)、基于标签内容的推荐算法(算法

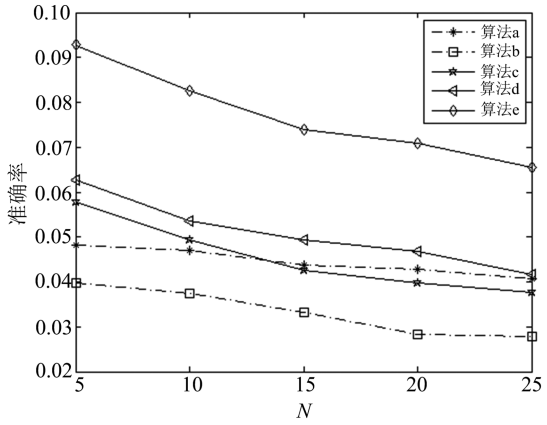

图 4 两种因素相关的标签下各算法的性能比较

Fig.4 Performance comparison of algorithms with two kinds of tags

d)、基于标签的混合推荐算法(算法 e)中分别运用与 4 种因素有关的标签和两种因素(地域和时间)相关的标签作对比,我们发现前者的推荐结果准确率明显高于后者.通过计算我们得到如表 3、4、5 的数据.

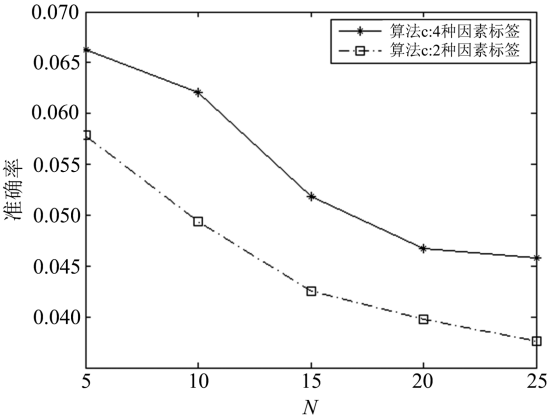

图 5 基于标签(2 种和 4 种)的协同过滤的算法

Fig.5 Tag-based cooperative filtering algorithm

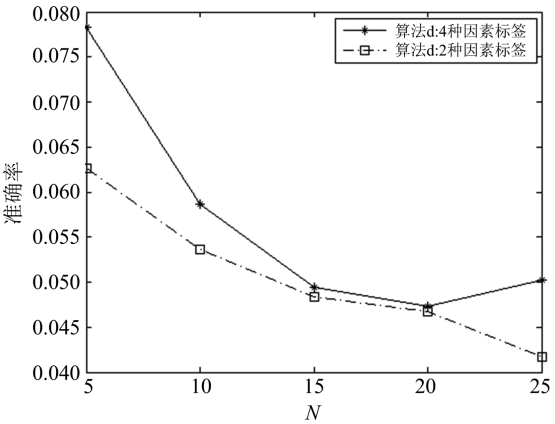

图 6 基于标签(2 种和 4 种)内容的推荐算法

Fig.6 Tag-content-based recommendation algorithm

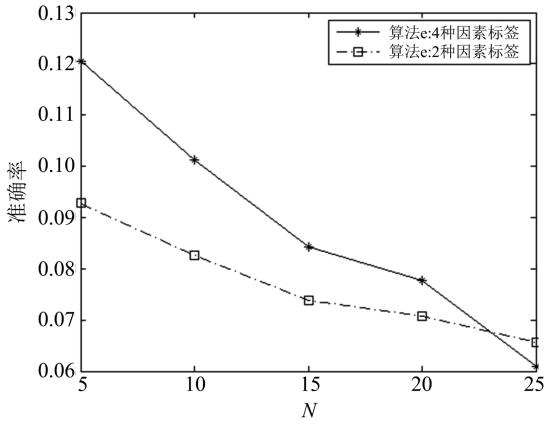

图 7 基于标签(2 种和 4 种)的混合推荐算法

Fig.7 Tag-based hybrid recommendation algorithm

表 3 基于标签的协同过滤算法的性能比较

Tab.3 Performance comparison of tag-based cooperative filtering algorithms

N 准确率算法	应用 2 种因素相关标签	应用 4 种因素相关标签	准确率提高
5	0.057 8	0.066 3	14.7%
10	0.049 4	0.062 0	25.5%
15	0.042 6	0.051 8	21.6%
20	0.039 8	0.046 7	17.3%
25	0.037 6	0.045 8	21.8%
均值	0.0454 4	0.0545 2	20.0%

表 4 基于标签内容的推荐算法的性能比较

Tab.4 Performance comparison of tag-content-based recommendation algorithms

N 准确率算法	应用 2 种因素相关标签	应用 4 种因素相关标签	准确率提高
5	0.062 7	0.078 3	24.9%
10	0.053 6	0.058 6	9.3%
15	0.048 4	0.049 4	2.1%
20	0.046 7	0.047 3	1.3%
25	0.041 8	0.050 2	20.1%
均值	0.0506 4	0.0567 6	12.1%

通过表 3、4、5 的数据,我们最终得到分别在基于标签的协同过滤算法、基于标签内容的推荐算法、基于标签的混合推荐算法下,运用 4 种因素相关标签时的准确率比用两种因素相关标签时明显提升,平均准确率分别提高 20.0%、12.1%、14.1%.由此可见,当我们运应用与时间、地域相关标签的时候,如

果再用上与主题、类型因素相关标签，推荐效果将显著提高.

表 5　基于标签的混合推荐算法的性能比较

Tab.5　Performance comparison of tag-based hybrid recommendation algorithms

N 准确率算法	应用 2 种因素相关标签	应用 4 种因素相关标签	准确率提高
5	0.092 8	0.120 5	29.8%
10	0.082 5	0.096 4	16.8%
15	0.073 9	0.084 3	14.1%
20	0.070 8	0.077 7	9.7%
25	0.065 5	0.061 0	—
均值	0.077 1	0.0879 8	14.1%

5　结论

本文提取景点标签并用景点标签描述用户兴趣特征，利用基于标签的协同过滤算法、基于标签内容的推荐算法、基于标签的混合推荐算法向用户作个性化景点推荐.景点标签和四种因素相关，分别是区域、时间、主题、类型，丰富了现有的通过主题模型提取的标签类别.基于 4 种标签的算法计算用户相似度进而作出推荐，改善了推荐效果.同时本文将基于标签的协同过滤算法与基于标签内容的推荐算法相结合，这种混合算法结合了两种算法的优势，进一步提高了推荐结果的准确度.

参考文献(References)

[1] G.Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[2] GE Y, XIONG H, TUZHILIN A, et al. An energy-efficient mobile recommender system[J]. Knowledge Discovery and Data Mining, 2010, 18(1): 899-908.

[3] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[4] LINDEN G, SMITH B, YORK J. Amazon. com recommendations: Item-to-item collaborative filtering [J]. IEEE Internet Computing, 2003, 7(1): 76-80.

[5] 冷亚军，陆青，梁昌勇. 协同过滤推荐技术综述[J]. 模式识别与人工智能，2014，27(8)：720-734.
LENGYajun, LU Qing, LIANG Changyong. Survey of recommendation based on collaborative filtering[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(8): 720-734.

[6] PAZZANI M J. A framework for collaborative, content-based and demographic filtering[J]. Artificial Intelligence Review, 1999, 13(5-6): 393-408.

[7] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[8] LIU Q, GE Y, LI Z M, et al. Personalized travel package recommendation[C]// Proceedings of 11th International Conference on Data Mining. Vancouver, Canada: IEEE Computer Society, 2011: 407-416.

[9] BAO J, ZHENG Y, MOKBEL M F. Location-based and preference-aware recommendation using Sparse Geo-Social Networking data[C]// Proceedings of the 20th International Conference on Advances in Geographic Information Systems. Redondo Beach, USA: ACM Press, 2012: 199-208.

[10] QIAN X M, FENG H, ZHAO G S, et al. Personalized recommendation combining user interest and social circle[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7): 1487-1502.

[11] FENG H, QIAN X M. Mining user-contributed photos for personalized product recommendation [J]. Neurocomputing, 2014, 129: 409-420.

[12] 刘树栋，孟祥武. 基于位置的社会化网络推荐系统[J]. 计算机学报，2015，38(2)：322-336.
LIUShudong, MENG Xiangwu. Rcommender systems in location-based social networks[J]. Chinese Journal of Computers, 2015, 38(2): 322-336.

[13] LIU Q, CHEN E H, XIONG H, et al. A cocktail approach for travel package recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 278-293.

[14] UMYAROV A, TUZHILIN A. Improving collaborative filtering recommendations using external data [C]// 8th International Conference on Data Mining. IEEE Press, 2008: 618-627.

[15] 聂恩伦，陈黎，王亚强，等. 基于 K 近邻的新话题热度预测算法[J]. 计算机科学，2012，39(6A)：257-260.
NIEEnlun, CHEN Li, WAGN Yaqiang, et al. Algorithm for prediction of new topic's hotness using the K-nearest neighbors[J]. Computer Science, 2012, 39(6A): 257-260.

[16] 项亮. 推荐系统实践[M]. 北京：人民邮电出版社，2012.