一种上下文敏感的基于因子分解机的乡村旅游推荐算法设计

Based on Factorization Machine：A Context-aware Design of Countryside Tourism Recommendation Algorithm

**1 Introduction**

名词翻译

乡村旅游：Countryside Tourism

城镇旅游：Town tourism

特点：feature

季节性：sesonal

地理空间分布：geospatial distribution

乡村旅游[1,2]是指以乡村地区为活动场所，利用乡村独特的自然环境、田园景观、生产经营形态、民俗文化风情、农耕文化、农舍村落等资源，为游客提供观光、休闲、体验、健身、娱乐、购物、度假的一种新的旅游经营活动。近年来我国乡村旅游得到了快速发展，据中国产业调研网发布的《2017-2020年中国休闲农业与乡村旅游市场现状研究分析与发展前景预测报告》[3]显示，截至2015年4月底，全国有9.5万个村开展休闲农业与乡村旅游活动，休闲农业与乡村旅游经营单位达193万家，其中农家乐达220万家，规模以上园区超过4.1万家，年接待游客接近8.4亿人次，年营业收入超过3200亿元。同时黄蓉[4]博士通过问卷调查并运用Logistic模型分析，说明了中国国内旅游的一大趋势是从城镇流向农村。显然，若能通过旅游推荐技术，为用户提供个性化的旅游推荐，将会极大提高乡村旅游系统的可用性。

推荐系统被定义成一种信息过滤系统，用于预测用户对物品的“评分”或“偏好”[5,6]。协同过滤技术被广泛应用于推荐系统中，是提出最早在推荐系统中应用最广的推荐技术。在旅游推荐中也被广泛研究和应用[14-16]。协同过滤算法又可以分为两类方法[17-19]基于近邻方法（neighborhood-based）和基于模型方法（model-based）。基于近邻方法的思想是运用系统中的用户-物品评分直接来预测用户对新物品的评分，是使用已有的数据直接进行预测。而基于模型的方法是用这些评分来学习预测模型。主要思想是使用属性构建用户和物品之间的关联。基于模型的协同过滤算法有很多，包括贝叶斯聚类（Bayesian Clustering）[]、潜在语义分析(Latent Semantic Analysis)、潜在狄利克雷分布(Latent Dirichlet Allocation)、支持向量机（Support Vector Machines）、奇异值分解（Singular Value Decomposition）等。2010年Steffen Rendle提出的FM算法给推荐系统的研究提供了新的思路，该方法在推荐系统研究的贡献主要有两点：1是解决了数据高稀疏度问题，2是可以加入任意实值特征向量，使其能够更好的适应上下文感知的推荐。这使得我们的工作重心转移到如何提取特征向量问题上来。

虽然乡村旅游已经成为一种趋势，但是基于乡村旅游的推荐技术还没有人进行深入的研究，尽管推荐技术在旅游中已经有了广发的应用。目前有关乡村旅游的研究都是基于社会调研，或统计学上的乡村旅游影响因素，及乡村旅游特征的研究。本文根据乡村旅游季节性强和地理空间位置特征，提出了一种适合乡村旅游景点推荐的推荐方案。该方案基于因子分解机算法，结合了季节和地理空间位置特征上下文，在实际采集的数据集上进行了验证。**本文的主要贡献如下**：（in total，our ）

1、基于因子分解机技术FM，提出了一种在乡村旅游这个新领域推荐问题的有效解决方案。

2、基于乡村旅游季节性强特征，提出了一种季节提取，转化及应用于乡村旅游推荐算法的方法。

3、基于乡村旅游地理空间位置特征，提出了一种旅游景点位置提取，转化及应用于乡村旅游推荐算法的方法。

The remainder of this paper is organized as follows.

2 Material and methods

In this section, We first present an overview of Factorization Machines algorithm. Then we introduce Countryside Tourism’s features and the way extracting and applying thems.

2.1 Factorization machines

The Factorization machines was proposed by Steffen Rendle[ ] , that is a general predictor working with any real valued feature vector. It provide high accuracy in recommender systems for it is able to estimate reliable parameters under very high sparsity. The model equation for a factorization machine is defined as follows:

In which is the global bias，models the strength of the feature variable , models the interaction between the l features variable. In this paper we use 2-way, the model equation is defined as follows:

And is the dot product of two vectors .

因子分解机模型可以无限增加因子的数量，所以应用因子分解机我们的重点工作可以放到因子的选择和转化上，接下来本文将介绍如何提取乡村乡村旅游的特征因子。

2.2 特征描述

1. Sesonal[ 21]，Although town tourism is also seasonal, countryside tourism is particularly seasonal. For example: For those who love the natural scenery, different seasons scenery will be very different. So, we assume that Seasonal features will have an impact on user ratings.

2. In the Geographical distribution, most tourists choose to travel by car, that make it express characteristics of geography. According to Lu Xiaoli et al. [ ] for nearly 20 years of domestic and foreign countryside tourism research hot sum up: tourists countryside tourism travel threshold is 20km from the city. That space within the city range from 20 ~ 100km of countryside tourism is a frequent occurrence area. So, we can know the countryside tourists have a preference area. Thereby we assume that geospatial distribution features will have an impact on user ratings.

2.3 Seasonal features

Due to the varying lengths of the seasons in different parts of China, we have devise two options. First according to the user comment time to extract the month information. Then, one solution is four seasons, in (3,4,5) is spring, in (6,7,8) is summer, in (9,10,11) is autumn, in (12,1,2) is winter. Mark the four seasons with 1, 2, 3 and 4 respectively.

Another solution is twelve seasons, that just use user comment month to mark the twelve seasons. In the experimental part we will conduct two sets of experiments to verify which scheme is better.

2.4 Geographical distribution features

Accroding to the countryside tourists have a preference area assume, the Geographical factors must reflect tourists’s preference area. There are two things we should do: get the location,?????

Based on the name of the collected attractions, get the location of the attraction through the Baidu API, the result is a set of longitude and latitude (x,y). The next step is to reduce the dimension.

本文采取降维处理的依据是“游客乡村旅游出游阈值为距离城市 20km，即距离城市20~100km之内的空间范围是乡村旅游的频繁发生地带”，得到以下三点：1、可以推出对于每个用户都有自己的旅游偏好区域D；2、在用户偏好区域D中的景点将会获得用户相对较高的评分；3、基于以上两点，将景点进行区域划分，找出最优的划分方法则可以模拟用户的偏好区域，从而提高预测评分的准确度。根据以上三点，本文采取给景点划分区域进行编号，则落入同一区域的景点可以获得相同的编号，体现出了在地理空间位置上的联系。区域编号作为因子分解机算法的因子。

具体操作：

假设前提：地图是个椭球体

那么：根据中华人民共和国国家统计局(http://www.stats.gov.cn/ztjc/ztsj/gjsj/2005/200607/t20060721\_59675.html)数据，地球赤道半径为R1≈ 6378km，地球极半径长度。

地球上同一纬度上两点A(lonA,α),B(lonB,α)的距离计算公式如下：

r=R1\*cosα

L1=

其中r为所在纬度平面的半径，α为弧度。L1为两点之间的距离。

地球上同一经度上两点C(β,latC),D(β,latD)的距离计算公式如下：

L2=

其中L2为两点之间的距离。

具体实现：取所选区域的边界minLng, maxLng, minLat, maxLat。以区域的左下角为坐标原点，沿横向纬度和纵向经度为坐标轴建立坐标系。以步长step=s将横纵坐标等分，则分割出N块区域，依次编号（1，2，…，N-1,N）如图所示：

步骤：

1、将步长转化为度数，计算同一纬度上经度步长，同一经度上纬度步长（定值）；

2、计算经度方向上坐标最大值k；

3、计算景点经度方向上坐标x，纬度方向上坐标y，容易得该景点的模块标签号为：blockNumber=y\*k+x;

4、根据景点的经纬度信息，依照以上公式映射入块

具体计算流程如下所示：

|  |
| --- |
| Get Block Number |
| Input：minLng, maxLng, minLat, maxLat, stepSize(km), dot(lng,lat)  Output：blockNumber |
| 计算每经度、每纬度的距离  perLng  perLat  计算经纬度步长（单位：度）  stepSizeLng  stepSizeLat  计算坐标轴横向最大块数k  计算坐标点坐标  计算块标号  return blockNumber |

步长s的取值，直接影响到总块值以及落入每块的景点个数。在实验部分，我们将将步长分别取值30，40，50，60……100，110，120，130，140，150km 不同值来测试预测的准确度，从而找出最优步长。

2.5基于季节地理位置上下文的推荐算法设计

In this paper, we choose the least square error as the loss function.

To prevent overfitting，use L2 regularization,so the equation turn into follws:

In which  represents the regularization coefficient of the parameter .

Learning algorithm use SGD(Stochastic Gradient Descent).This algorithms are very popular for optimizing factorization models as they are simple, work well with different loss functions. The process of SGD training FM algorithm is as follows:

|  |
| --- |
| SGD(Stochastic Gradient Descent) |
| **Input:**Training set D, regularization parameters, learning rate , initialization  **Output:** Model parameters  Initialization:  Repeat  FOR DO  ;  FOR DO  ;  FOR DO  ;  end  end  end  **Until** stopping criterion is met. |

**需要添加一些内容**

**在算法中，提取出的季节信息和 景点地理位置信息，作为算法中的因子对预测结果产生影响。**

**3 Experiment**

In this section, we first introduce the dataset. Then, the metrics and experimental method are presented. Next, three benchmark methods are presented. Finally, we show the detail experimental results.

3.1 数据集介绍

Our dataset collect travel information from four provinces including Hebei, Henan, Shangdong and Shanxi. The data set includes 45254 users, 3848 items (the attractions), and 260694 ratings (1 to 5 stars). The data set information is shown in Table 1：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Users | Items | Ratings | Sparsity  % |
| MF-travel | 45254 | 3848 | 260694 | 0.15 |

3.2 Metrics评价方法

我们使用了两种在推荐领域经常使用的评价指标 RMSE与ASE

We use two authoritative evaluation metrics in the recommended areas: Mean Absolute Error (MAE) and Root Mean Square Error(RMSE). The metrics are defined as:

Where N denotes the number of elements of the data set.

3.3 Experimental result comparison

In this section we design two experiments about season and geospatial distribution. Finally come to the best season label and geospatial distribution label.

3.3.1 Experiment for season

**数据集处理后的格式为**

uirt(user – item – rating – time )

uirtSeaon(user – item – rating – time – season )

uirtSeaon12(user – item – rating – time – season )

uirtLo(user – item – rating – time - location)

uirtSeason12Lo(user – item – rating – time -season- location)//用来证明两个因素的正相关性

uirtSe12Comm(user – item –rating – time - season –comCount - greatCo – goodCo - badCo)可能不用

uirtSe12CommLo(user – item – rating – time – season – comCount - greatCo-

goodCo – badCo - loction) 可能不用

前期默认试验参数设置: learnRate=0.001 , factor=6 , iterations =1000

第一组验证季节选择4还是12

第一组实验是验证季节因子对推荐预测评分的影响，本文从用户评论时间中提取出月份，季节有两种方式来表示，一是4季，考虑到不同地区季节间的差异，本文的第二种表示方法为直接以月份表示，共12季。以uirt作为对比基本对比试验，最终试验结果如下图所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | Uirt | UirtSe4 | UirtSe12 |
| MAE | 0.5627437481404048 | 0.5603457314107018 | 0.5556676766284818 |
| RMSE | 0.7517287833666975 | 0.7508853044321885 | 0.7433529494967073 |

第二组实验找出最优的步长S，实验参数选择：

第二组试验是验证用户偏好区域对推荐预测评分的影响，本文通过设置不同步长S进行试验，以此来找出最优步长。以是数据集uirt作为基本对比，以数据集uirtLo做试验。

实验结果如下图/表：

从图中可以看出，在步长80km时，平均误差（MAE）与均方根误差（RMSE）达到最小，即效果最好。

下面就以上两个试验为基础，验证两个因子的正相关性，季节设定为12季，步长取80km。实验结果如下图：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Uirt | UirtSe12 | UirtLo | UitrtSe12Lo |
| MAE | 0.5627437481404048 | 0.5556676766284818 | 0.547972243042463 | 0.5462491722962254 |
| RMSE | 0.7517287833666975 | 0.7433529494967073 | 0.731266452982731 | 0.7303562418646597 |

从图中可以看出，季节和地理特征两个因子成正相关。

为了调试到最优效果，本文了在不同维度f和不同迭代次数对试验效果的影响。实验结果如图所示：

参数调试试验:

在维度f=6的条件下,学习速率learnRate=0.01 , 0.005 , 0.001 的效果比较

Iterations = 100 , 500 , 1000

在学习速率learnRate= 的条件下,维度factor = 5 , 10 , 15 ,20

Iterations = 100,500,1000

3.4 Baseline approaches 与基准算法的对比实验

基准算法介绍

与基准算法的对比实验

3.5 Results comparison

实验结果比较

4 Discussion

5 Conclusion and future work