# 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

[[1]](#footnote-1) 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。

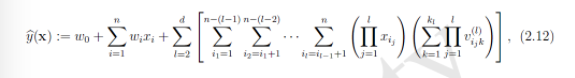
The remainder of this paper is organized as follows.

# 2、Material and methods（介绍相关算法与特征提取应用）

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 特征提取

### 2.2.1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.

### 2.2.2Seasonal 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.

### 2.2.3Geographical 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 latitude and longitude(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：

取所选区域的上下左右最大边界EXmax、WXmax、SYmax、NYmax，以步长S将横纵坐标与纵坐标进行等分，则分出N个区块。

假设将地图看做是个完美的球体，地图半径为R ≈ 6371km，则同一纬度上的两点A(lonA,α),B(lonB,α)的距离，同一经度上两点C(β,latC),D(β,latD)计算公式如下：

所在纬度平面半径：r=R\*cosα

同纬度两点之间距离：L1=

同经度两点之间距离：L2=

可以推算出，所选区域展开后可视为梯形

以左下角为原点，沿经纬度方向建立坐标系。

以步长step=s沿横轴纵轴进行分割，从左下角以1开始按顺序依次编码，N块标签为1-N。则景区根据经纬度计算落入哪块区域。

**如何映射入块：**

**添加伪代码执行步骤/公式**

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20，30，40，50，60……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。

2.3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  { IF  {;  FOR DO  {;}  }}}  **Until** stopping criterion is met. |

# 3、Experiment

## 3.1数据集介绍

数据集介绍

## 3.2. Metrics评价方法

## 3.3Experimental result comparison实验结果比较

设计两组实验：第一组验证季节选择4还是12

第二组实验找出最优的步长s

4.3.1. Baseline approaches 基准方法-----对比实验的基准方法

基准算法介绍，基准算法的对比实验

4.3.2. Results comparison结果比较

实验结果比较

# 4 Discussion

讨论

# 5 Conclusion and future work

The countryside tourism has some obviously features compared to town tourism, especially in the seasonal and geospatial distribution has obvious features. Since the Factorization machine algorithm has the excellent characteristics of an infinite increase factor, so what we should do is to extract valid factors and effectively express factors. In this paper we turn the two features into factors.

乡村旅游区别于城镇旅游，有着自身特有的特征，尤其是在季节性和地理空间分布上有着明显的特征。由于因子分解机具有无限增加因子的好的优良特性我们要做的只是如何提取有效的因子，以及如何表达因子。

1. 基金项目：中央高校基本科研业务费专项（2017XS002）;

   作者简介：高万林（1965—），男，教授，博士生导师，主要从事计算机网络与信息安全研究，E-mail：gaowlin@cau.edu.cn [↑](#footnote-ref-1)