# 用户管理

用户数据导入：从YFCC 100M数据集中提取10000多条用户数据，作为系统预用数据。

新用户注册：提供注册功能，实现新用户的注册。

用户登录：实现用户登录，基于用户进行个性化推荐。

# 照片管理

照片上传：用户上传其旅游照片。

信息提取：分析照片的EXIF信息，从中提取出照片拍摄时间，拍摄地等信息。

照片删除：用户可选择将其上传过的照片删除。

# 景点推荐

数据清洗：清洗不含GEO信息的照片数据，明显的噪音数据等。

## 景点聚类

使用DBSCAN算法，对照片信息聚类，获取景点信息。

### 密度聚类

DBSCAN是一种基于密度的聚类算法，这类密度聚类算法一般假定类别可以通过样本分布的紧密程度决定。同一类别的样本，他们之间的紧密相连的，也就是说，在该类别任意样本周围不远处一定有同类别的样本存在。

通过将紧密相连的样本划为一类，这样就得到了一个聚类类别。通过将所有各组紧密相连的样本划为各个不同的类别，则我们就得到了最终的所有聚类类别结果。

### DBSCAN密度定义

DBSCAN(Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法)是一种很典型的密度聚类算法，和K-Means，BIRCH这些一般只适用于凸样本集的聚类相比，DBSCAN既可以适用于凸样本集，也可以适用于非凸样本集。

DBSCAN是基于一组邻域来描述样本集的紧密程度的，参数(ϵ， MinPts)用来描述邻域的样本分布紧密程度。其中，ϵ描述了某一样本的邻域距离阈值，MinPts描述了某一样本的距离为ϵ的邻域中样本个数的阈值。

假设样本集是，则DBSCAN具体的密度描述定义如下：

1） ϵ-邻域：对于，其ϵ-邻域包含样本集D中与的距离不大于ϵ的子样本集，即， 这个子样本集的个数记为

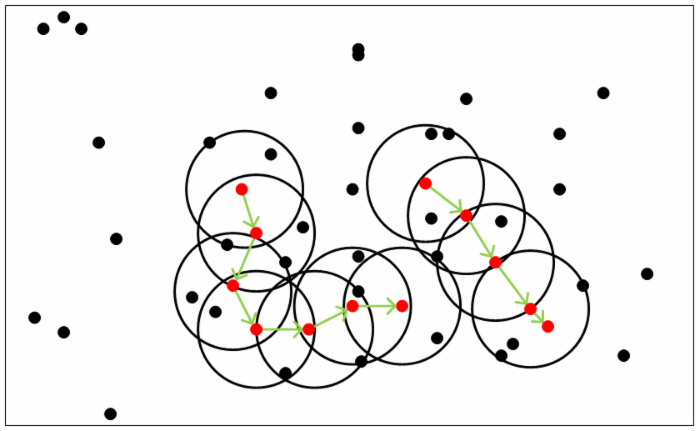
2) 核心对象：对于任一样本 ，如果其ϵ-邻域对应的 至少包含MinPts个样本，即如果，则是核心对象。

3）密度直达：如果位于的ϵ-邻域中，且是核心对象，则称由密度直达。注意反之不一定成立，即此时不能说由密度直达， 除非且也是核心对象。

4）密度可达：对于和，如果存在样本样本序列，满足 ，且由密度直达，则称由密度可达。也就是说，密度可达满足传递性。此时序列中的传递样本均为核心对象，因为只有核心对象才能使其他样本密度直达。注意密度可达也不满足对称性，这个可以由密度直达的不对称性得出。

5）密度相连：对于和，如果存在核心对象样本，使和均由密度可达，则称和密度相连。注意密度相连关系是满足对称性的。

从下图可以很容易看出理解上述定义，图中MinPts=5，红色的点都是核心对象，因为其ϵ-邻域至少有5个样本。黑色的样本是非核心对象。所有核心对象密度直达的样本在以红色核心对象为中心的超球体内，如果不在超球体内，则不能密度直达。图中用绿色箭头连起来的核心对象组成了密度可达的样本序列。在这些密度可达的样本序列的ϵ-邻域内所有的样本相互都是密度相连的。



### DBSCAN聚类算法

|  |
| --- |
| 算法1：DBSCAN聚类算法 |
| 输入：样本集，邻域参数(ϵ,MinPts), 样本距离度量方式 |
| 输出：簇划分C |
| 1. 初始化核心对象集合, 初始化聚类簇数k=0，初始化未访问样本集合, 簇划分 |
| 1. 对于j=1,2,...m, 按下面的步骤找出所有的核心对象：    1. 通过距离度量方式，找到样本的ϵ-邻域子样本集    2. 如果子样本集样本个数满足， 将样本加入核心对象样本集合： |
| 1. 如果核心对象集合，则算法结束，否则转入步骤4. |
| 1. 在核心对象集合Ω中，随机选择一个核心对象o，初始化当前簇核心对象队列, 初始化类别序号k=k+1，初始化当前簇样本集合, 更新未访问样本集合 |
| 1. 如果当前簇核心对象队列，则当前聚类簇生成完毕, 更新簇划分, 更新核心对象集合， 转入步骤3。 |
| 1. 在当前簇核心对象队列中取出一个核心对象o′,通过邻域距离阈值ϵ找出所有的ϵ-邻域子样本集，令 更新当前簇样本集合, 更新未访问样本集合 更新，转入步骤5. |
| 1. 输出结果： 簇划分 |

## 推荐模型

构建推荐模型，使用基于物品的协同过滤算法，ALS 矩阵分解算法等，使用投票法则整合算法。

### 协同过滤

### ALS矩阵分解

# 地图LBS

根据图片的经纬度信息，将景点信息显示在实际地图中。