1. 用户画像有助于了解如何定义标签，从而根据标签来对用户进行内容推荐。

用户画像是数据驱动(DT)的基本应用，而DT是互联网下半场的发展方向，对用户信息数据挖掘的深度决定了业务拓展的宽度和未来发展方向。

1. 给用户打标签有4个维度：用户、消费、行为、分析

用户标签（静态）：性别、年龄、地域、收入

消费标签：消费习惯、购买意向、是否对促销敏感

行为（显式行为和隐性行为）：收藏、喜欢、评分（显式）；频次、时长（隐式）。

内容分析：对平时其他浏览内容的分析：体育、游戏、八卦

1. 用户生命周期有三个阶段：

获客：关注拉新，获取客户

粘客：个性化推荐，搜索排序，场景运营等

留客：分析潜在流失客户，短信激活。寻找降低流失率的关键节点。

1. 标签来源：

原始数据：前端埋点记录或后端脚本日志

PGC：工作人员专家生产，提前设置好的默认标签

UGC：普通用户生产，包括用户评价中的高频词汇提取

标签是对高维事物的抽象降维

1. K-means工作原理：

有监督学习带label的学习方法。提前已经确定好要分几类(K)

S1.选取K个点作为初始的类中心点，这些点一般从数据中随机抽取；

S2.将每个点分配到最近的类中心点，这样就形成K个类，然后重新计算每个类的中心点；

重复S2直到类不发生变化，或者设置结束的最大迭代次数。

1. 距离的表示方式：欧式距离、曼哈顿距离、切比雪夫距离、余弦距离（表示向量间的距离）
2. 数据规范化的方法：

Min-Max法：将原始数据投射到指定空间[min,max]

新数值=（原数值-min）/（max-min）

当min=0，max=1时为[0，1]规范化

对应sklearn中的MinMaxscaler方法

Z-score规范化：即正态分布形式的转化

新数值=（原始值-均值）/标准差

对应sklearn中的preprocessing.scale

小数定标规范化：通过numpy移动小数点的方式快速规范化

1. 评价指标

准确率accuracy = （TP+TN）/（TP+FP+FN+TN）

召回率 recall= TP/(TP+FN)

精确率 precision= TP/(TP+FP)

1. TF-IDF

TF:Term Frequency词频=

一个单词的重要性和它在文档中出现的次数呈正比

IDF：Inverse Document Frequency 逆向文档频率

一个单词在文档中的区分度，这个单词出现的文档数越少，区分度越大，IDF越大

1. 基于内容的推荐系统步骤：
2. 物品表示，为每个item抽取features
3. 特征学习，利用一个用户过去喜欢（不喜欢）的item的特征数据，学习用户的喜好特征（profile）
4. 生成推荐列表Recommendation Generation

通过用户profile与候选item的特征，推荐相关性最高的item