

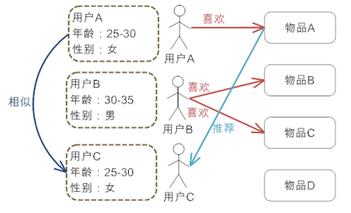
[图 1. 推荐引擎工作原理图](http://www.ibm.com/developerworks/cn/web/1103_zhaoct_recommstudy1/index.html)

* **我们需要的数据源包括：**

1. 用户对ad的偏好，主要是用户查看ad的记录，包括查看地点、查看时间、查看设备等；**（必需）**
2. 用户的基本信息，如住址；（可选）
3. 要推荐ad的元数据，如关键字、类目、价格区间、新旧程度等。（可选）

* **现阶段选择基于人口统计学的推荐机制，理由有三：**

1. 最易于实现；
2. 不需要当前用户查看ad的历史数据，对新用户同样适用；
3. 不管是推荐类目还是推荐ad都可以使用这一方法。



[图 2. 基于人口统计学的推荐机制的工作原理](http://www.ibm.com/developerworks/cn/web/1103_zhaoct_recommstudy1/index.html)

* **具体实现：**

1. 对用户查看ad的记录进行聚类。用于相似度度量的属性可以使用但不限于地理位置等；
   1. 输入：用户查看ad的记录；
   2. 输出：相似度高的记录将被聚成一类；
   3. 选项：记录的时间跨度；聚类类别数；
2. 根据聚类结果得到同一类用户（如地理位置相近的）曾经查看过的ad列表，对每一类的列表排序；
   1. 输入：每一类的ad列表
   2. 输出：每一类的ad有序列表
   3. 选项：根据查看频率/更新时间排序
3. 对于给定位置的用户，把该用户划归为某一类用户，返回该类的ad有序列表。
   1. 输入：用户的位置信息
   2. 输出：该用户所属类的ad有序列表

***\* 类目的推荐类似***

* **聚类方法：**

1. 根据指定的属性（如地址、查看时间等）的近似度进行聚类，需要定义近似度度量。如所有发生在晚上的交通大学的ad查看记录是高度近似的。
2. 或通过已有的数据集进行训练，变成分类问题，需要已经分好类的训练数据集。难点在于在对数据特征不清楚的情况下，怎样的ad查看记录是属于一类的较难确定。