8.1 大数据的广泛应用

--李晚月，安航正，章舒婕

一 数据产业的发展趋势

1.连接泛在化：

互联网+，移动互联网，广播电视网，通信网，物联网，车联网，船联网

2.计算多元化：

机器学习，数据挖掘，流式计算，批量计算，图数据计算，高性能计算，量子计算

3.数据产业化：

工业大数据，政府大数据，科学大数据，商业大数据，教育大数据，社会网络大数据

4.应用智能化：

智慧城市，智慧家居，智慧制造，智能医疗

5.创新生态化：

跨界融合创新，产业协同创新，技术集成创新

二 大数据的广泛应用

•**社会网络大数据**：网络社会化，人们大量的时间生活在网络中，人的网络行为生成的海量的数据，社会网络已经成为联机网络信息空间和物理世界不可或缺的桥梁。

•**城市大数据：**智慧城市通过大数据研究，解决城市扩张而来的诸多问题。智能交通，利用交通大数据，采集交通实时监控数据，进行智能疏导、公共车辆管理，车辆辅助控制等应用。

•**商业大数据**：关联购买行为的客户群体大数据分析，已经渗透到零售行业、餐饮行业和电信行业等。推荐系统进行商业个性化推荐。

•**教育大数据**：通过对教育中搜索的数据进行分析，用于教育发展、创造巨大潜在价值的数据集合。例如信息化校园、智能辅导系统和在线题库、认知诊断或习题资源分析和挖掘、MOOC平台活跃度预测。

三 通过大数据分析的预测：

• **体育赛事预测**

• **股票市场预测**

• **用户行为预测**：基于用户搜索行为、浏览行为、评论历史和个人资料等数据，互联网业务可以洞察消费者的整体需求，进而进行针对性的产品生产、改进和营销。

• **人体健康疫情预测**

• **疾病疫情预测**：基于人们的搜索情况、购物行为预测大面积疫情爆发的可能性。

• **交通行为预测**：基于用户和车辆的LBS定位数据，分析人车出行的个体和群体特征，进行交通行为的预测。

8.2.1推荐系统

--潘念然，徐雅婕，赵亚然

* **推荐系统简介**

**推荐系统**(Recommendation System)通过用户与信息产品之间二元关系，利用已有的选择过程或相似性关系挖掘每个用户潜在感兴趣的对象，进行个性化推荐，其本质就是信息过滤。

* **推荐系统的组成**

完整的推荐系统由收集、分析和推荐3个部分组成。

* 收集用户信息的行为的记录模块
* 分析用户喜好的模型的分析模块
* 推荐算法模块
* **推荐技术分类**
* 基于内容(Content-based Recommendations，CBR) 的推荐:关于推荐对象的特征提取。
* 基于关联规则的推荐:挖掘用户与物品的关联规则
* 基于知识的推荐:基于某- -领域一整套 规则和线路推荐
* 协同过滤(Collaborative Filtering ，CF)
  + 基于领域/记忆的模型

1. 基于用户的协同过滤(User CF) >基于用户的Top-N算法
2. 基于物品的协同过滤(Item CF) >基于物品的Top-N算法
   * 隐语义模型LFM(latent factor model)是通过隐含特征联系用户兴趣和物品。
   * 基于图的模型 随机游走算法

* 混合推荐
* **基于用户的协同过滤**

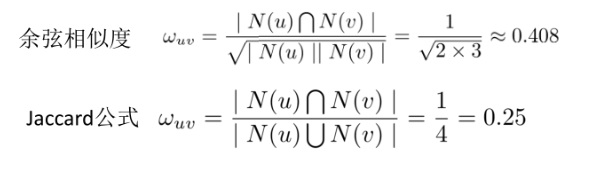
基于用户的协同过滤推荐算法(User CF) ,通过相似性算法来测评用户之间的相似性，然后基于用户之间的相似性进行产品推荐。

* 步骤：

1. 找到和目标用户信息相似的用户集合。
2. 找到这个集合中用户喜欢的物品推荐给目标用户。

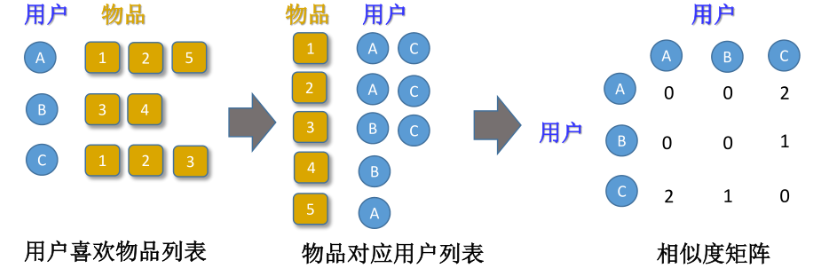
* 具体实现：
* 步骤1:找到和目标用户兴趣相似的用户集合( 喜欢物品的相似)。

首先计算两个用户u和v相似度。令N(u)表示u用户喜欢的物品集合，比如{物品1，物品3}，|N(u)|表示集合中元素个数为2，采用余弦相似度或Jaccard公式对两个用户度量Wuv。



有些情况是，可能很多用户之间没有对同样的物品感兴趣，那么相似度公式的分子为0，他们的相似度也是0。

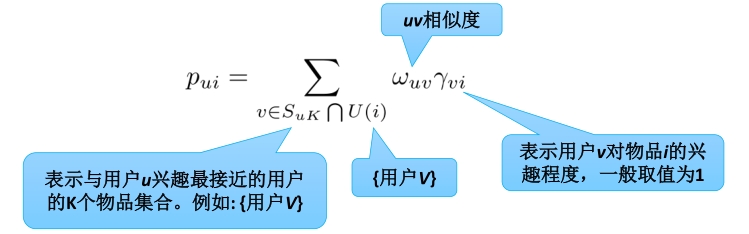
这里可以利用物品到用户的倒排表，仅对有对相同物品产生交互行为的用户进行计算其相似度矩阵。



* 步骤2:找到用户集合所喜欢的物品且目标用户未选择的，选取物品推荐给

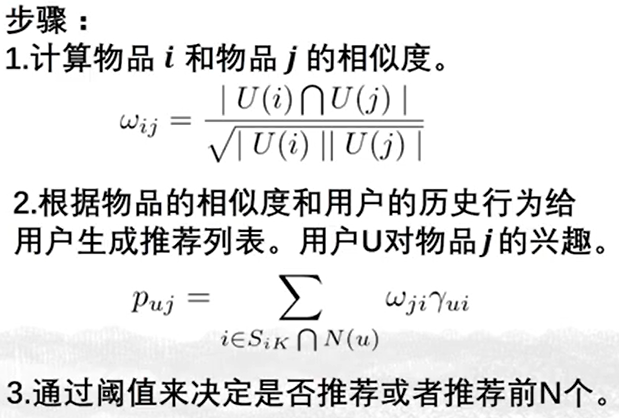
目标用户。

计算用户u对物品i的感兴趣程度Pui 然后通过设定阈值来决定是否推荐物品i。



基于用户的Top-N算法，则不设置阈值，而是通过对物品感兴趣程度的大小选取前N个物品的顺序清单。

基于用户的协同过滤推荐算法：通过用户对不同物品的评分来评测不同物品之间的相似性，给用户推荐和他之前喜欢的物品的相似的物品



UserCF和ItemCF的区别:

基于用户的协同过滤是推荐用户所在兴趣小组中用户点击的热点，更注重社会化。例如：新闻网给目标客户推荐其他有相同或相似兴趣爱好的人关注的新闻，给用户最新最及时最感兴趣的新闻信息。推荐结果有一定的新颖性。

基于项目的协同过滤是根据用户历史行为推荐相似物品，更注重个性化。例如：购物、书籍、电影，用户的兴趣爱好一般比较固定，网站一般是给目标用户推荐他所感兴趣领域的产品。

