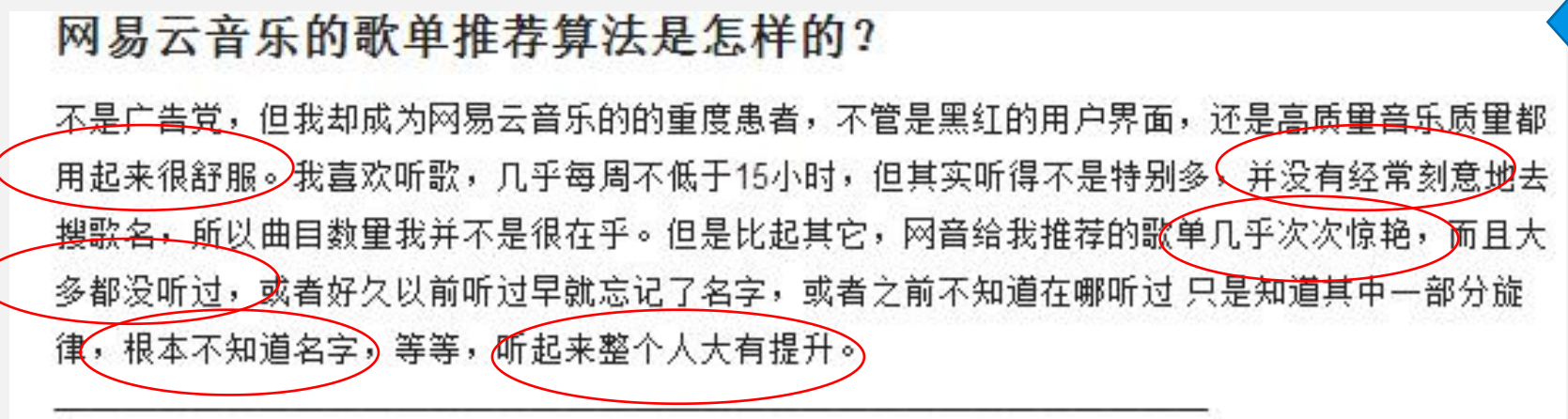
大数据分析与应用

第4章 推荐系统初步：推荐系统的产生

- 分类目录——Yahoo
- 搜索引擎——Google
- 推荐系统——电子商务、电影和视频网站、音乐与网络电台、社交网络、个性化阅读、基于位置的服务、个性化邮件、个性化广告

- 被搜索信息量巨大：希望能低成本的找到最合适的（物美价廉、最喜欢）
- 用户搜索需求不明确：“过度自由的选择是一件痛苦的事情”


网易云音乐的歌单推荐算法是怎样的？

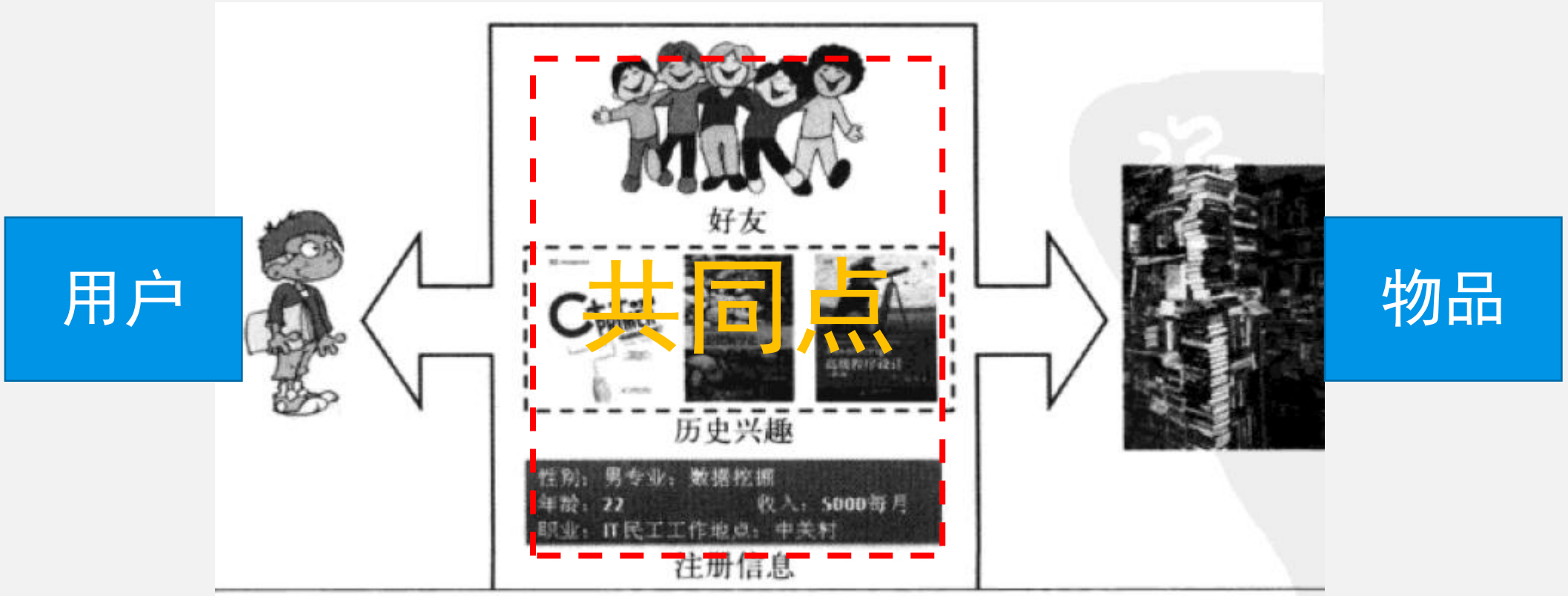
不是广告党，但我却成为网易云音乐的的重度患者，不管是黑红的用户界面，还是高质量音乐质量都用起来很舒服。我喜欢听歌，几乎每周不低于15小时，但其实听得不是特别多，并没有经常刻意地去搜歌名，所以曲目数量我并不是很在乎。但是比起其它，网音给我推荐的歌单几乎次次惊艳，而且大多都没听过，或者好久以前听过早就忘记了名字，或者之前不知道在哪听过 只是知道其中一部分旋律，根本不知道名字，等等，听起来整个人大有提升。

知乎的评论

- 被搜索信息量巨大
- 用户搜索需求不明确

第4章 推荐系统初步：定义

- 推荐算法是通过一定的方式将用户与物品联系起来。(目标)



4.1 什么是好的推荐系统？

□推荐系统的三个参与方：

用户（消费者）

物品（生产者）

推荐系统平台提供者

□ 事后的评价关注用户满意度

- 预测准确度（有没有去消费）：最原始的评价指标
- 用户满意度（消费的体验如何）：需要用户主动的评价

4.1什么是好的推荐系统？

■ 其他指标

- ✓ 覆盖率：描述一个推荐系统对物品长尾的发掘能力；
- ✓ 多样性：覆盖用户不同兴趣领域；
- ✓ 新颖性：给用户推荐他们没有听说过的物品；
- ✓ 惊喜度：与用户历史兴趣无关，却让用户觉得满意；
- ✓ 信任度：用户对结果的信任程度；
- ✓ 实时性：给用户推荐具有时效性的商品；
- ✓ 健壮性：推荐系统的作弊问题；
- ✓ 商业目标：商家与平台的盈利

常见方法

●协同过滤方法：信息来源于用户行为数据

4.2 基于邻域的方法：基于用户/物品的协同过滤

C8 基于模型的方法：矩阵分解/排序

●基于内容的推荐方法：信息来源于物品和消费者的特征

4.3 基于标签的方法：人工标签、自动标签嵌入（C8）

4.4 基于约束的推荐：预算、时效、地理信息（C8）

4.2.利用用户行为数据方法：数据特点

用户行为数据：显性反馈数据与隐性反馈数据

	显性反馈	隐性反馈
视频网站	用户对视频的评分	用户观看视频的日志、浏览视频页面的日志
电子商务网站	用户对商品的评分	购买日志、浏览日志
门户网站	用户对新闻的评分	阅读新闻的日志
音乐网站	用户对音乐/歌手/专辑的评分	听歌的日志

两类用户行为数据的性质

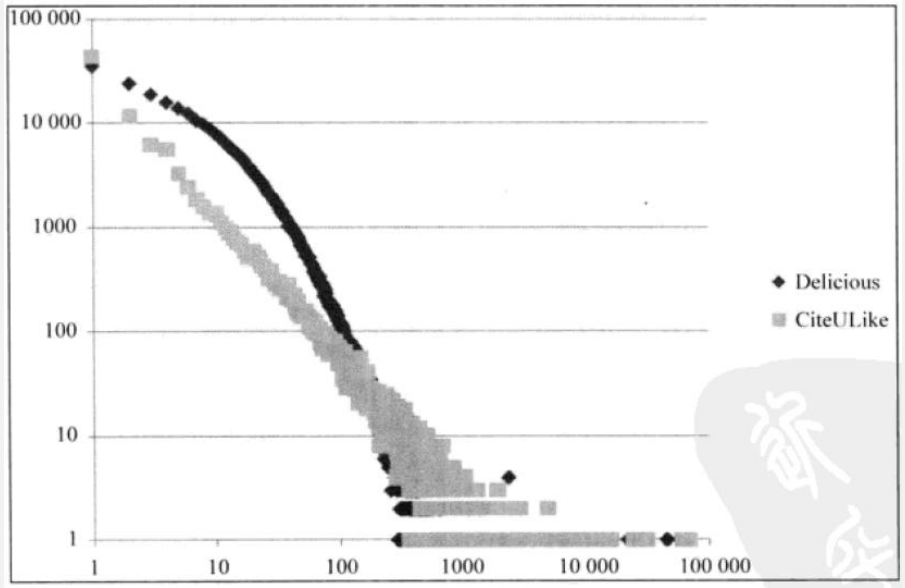
	显性反馈数据	隐性反馈数据
用户兴趣	明确	不明确
数量	较少	庞大
存储	数据库	分布式文件系统
实时读取	实时	有延迟
正负反馈	都有	只有正反馈

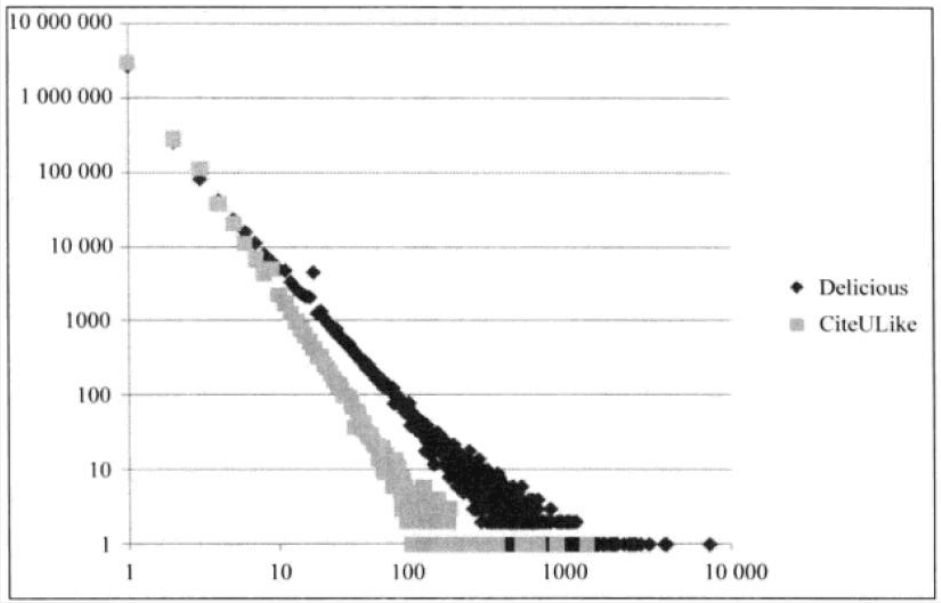
4.2. 利用用户行为数据

数据格式举例

user id	产生行为的用户的唯一标识
item id	产生行为的对象的唯一标识
behavior type	行为的种类（比如是购买还是浏览）
context	产生行为的上下文，包括时间和地点等
behavior weight	行为的权重（如果是观看视频的行为，那么这个权重可以是观看时长；如果是打分行为，这个权重可以是分数）
behavior content	行为的内容（如果是评论行为，那么就是评论的文本；如果是打标签的行为，就是标签）

用户行为数据中的长尾现象





4.2 利用用户行为数据：协同过滤

■ 1．基于用户的协同过滤算法：给用户推荐和他兴趣相似的其他用户喜欢的物品。

- （标志推荐系统的诞生（1992））
- 基于用户的相似度计算

■ 2．基于物品的协同过滤算法：给用户推荐和他之前喜欢的物品相似的物品

- （目前业界使用最多的算法，Amazon，YouTube：过时）
- 基于物品的相似度计算（物品的相似度中包含着用户的相似度）

4.2 利用用户行为数据：基于用户的协同过滤算法

算法描述：

1. 找到和目标用户兴趣相似的用户集合；
2. 找到这个集合中的用户喜欢的，并且目标用户没有听说过的物品；
3. 将找到物品推荐给目标用户

兴趣相似如何度量？

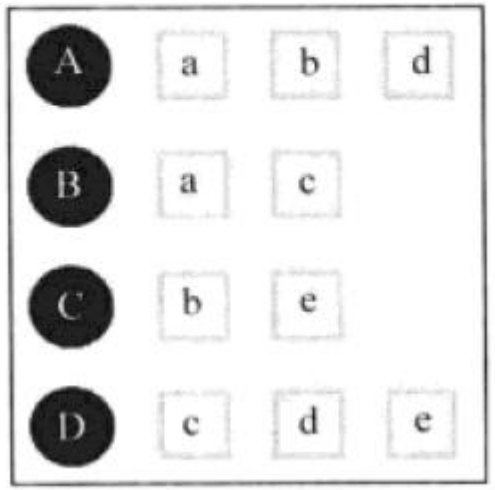
余弦相似度

$$w_{uv}=\frac{|N(u)\cap N(v)|}{\sqrt{|N(u)||N(v)|}}$$

兴趣相似度

$$w_{uv}=\frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|}$$

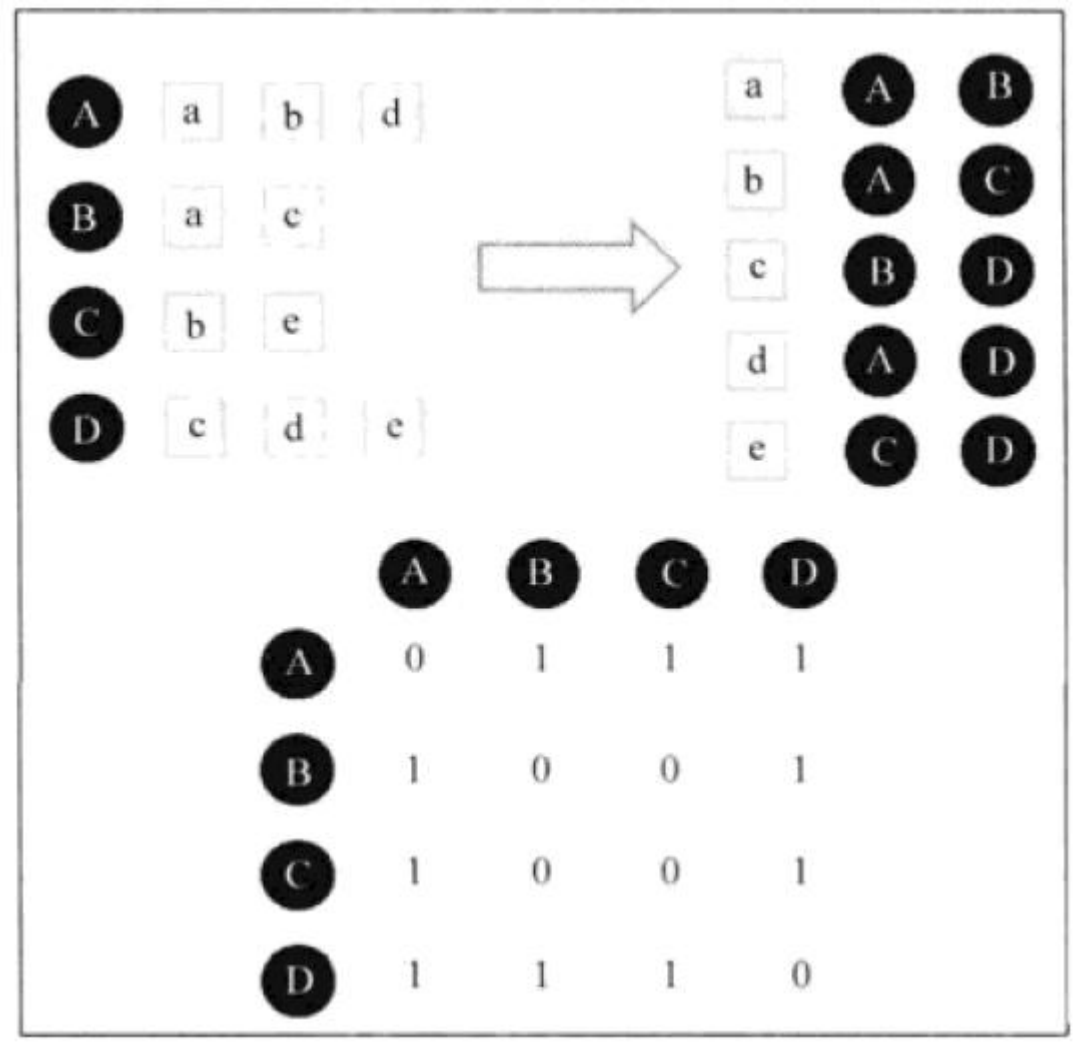
$$w_{AB}=\frac{|\{a,b,d\}\cap\{a,c\}|}{\sqrt{|\{a,b,d\}|\,|\{a,c\}|}}=\frac{1}{\sqrt{6}}$$



4.2 利用用户行为数据：基于用户的协同过滤算法

- 所有用户之间两两计算相似度使得计算复杂度极大
- 而相似度矩阵可能是稀疏的

- 方案：
  1. 得到关于物品的倒排表
  2. 构建相似物品数量矩阵（相似度分子部分）
  3. 在非零元素上计算相似度



4.2 利用用户行为数据：基于用户的协同过滤算法

最后在相似度的基础上构建用户与物品的关联：

1. 选取与目标用户u最接近的N个用户（下标为v）；
2. 在其中找出对物品i有过行为（购买、浏览、评论）的用户集合不同的行为有不同的关联度r；
3. 计算目标用户u与物品i的推荐度；

$$p(u,i)=\sum_{v} w_{uv} r_{vi}$$

4. 将推荐度高的物品推荐给客户；

4.2 利用用户行为数据：基于用户的协同过滤算法

- 基于用户算法的挑战：
  - 用户数的不断增长，相似度矩阵不断增长
  - 难以解释推荐的结果

4.2 利用用户行为数据：基于物品的协同过滤算法

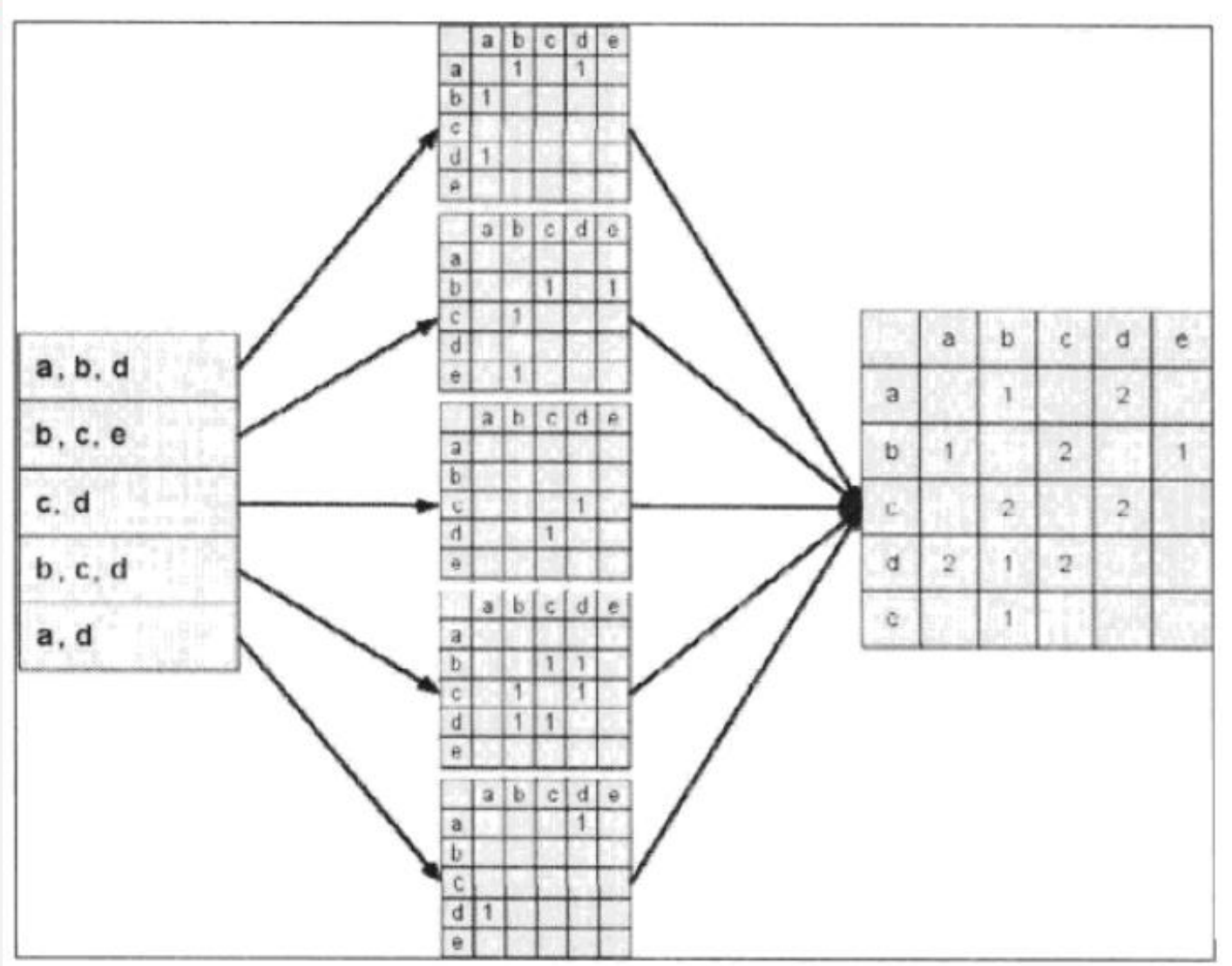
基于物品的协同过滤算法主要分为两步。
(1) 计算物品之间的相似度。
(2) 根据物品的相似度和用户的历史行为给用户生成推荐列表。

4.2 利用用户行为数据：基于物品的协同过滤算法

物品相似度的计算：

1. 购买记录向量化
2. 向量相加得到相似矩阵
3. 挑选非零元素计算相似度（余弦相似度）

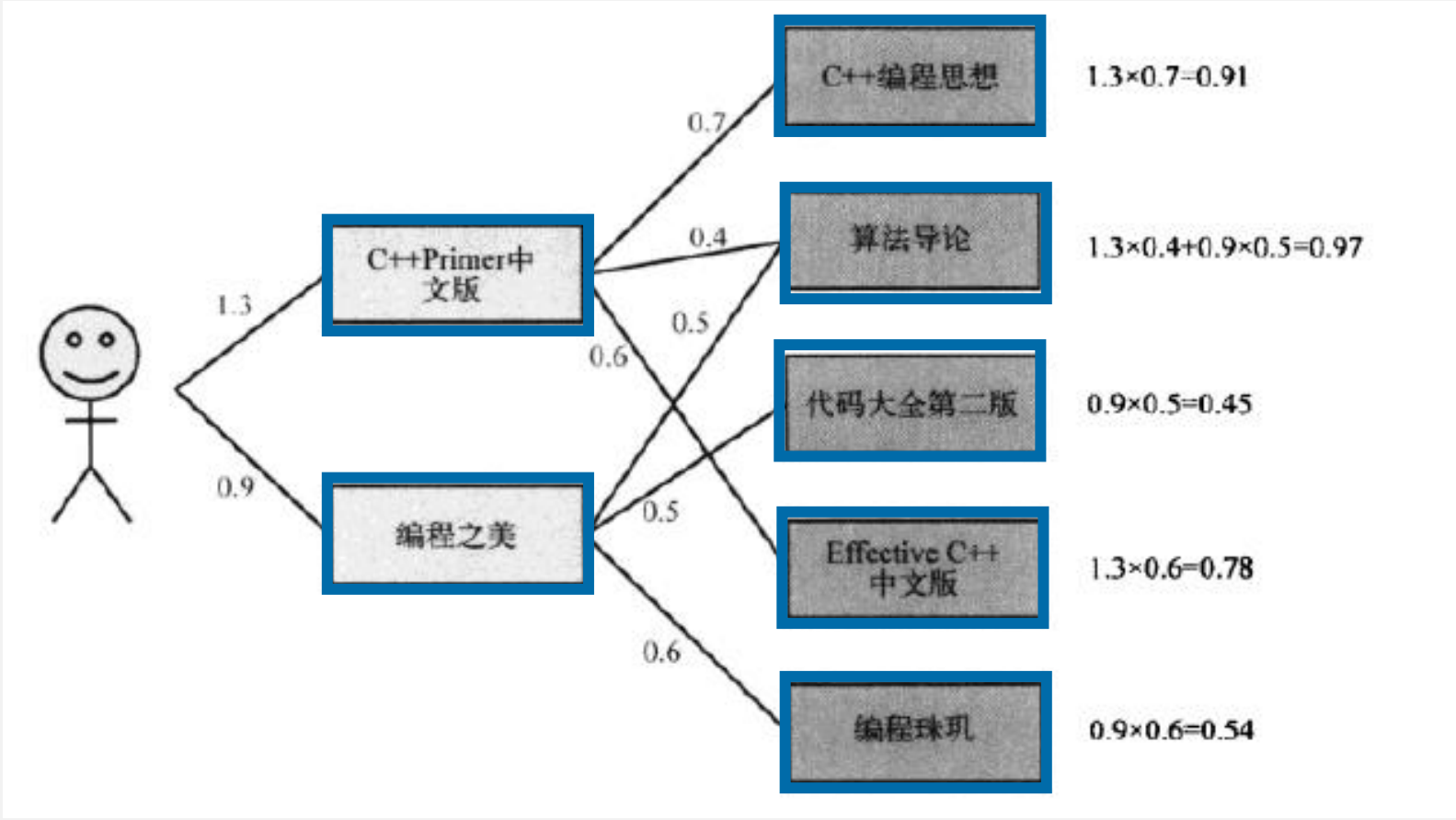
$$w_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i)||N(j)|}}$$



4.2 利用用户行为数据：基于物品的协同过滤算法

用户与物品之间的关联计算：

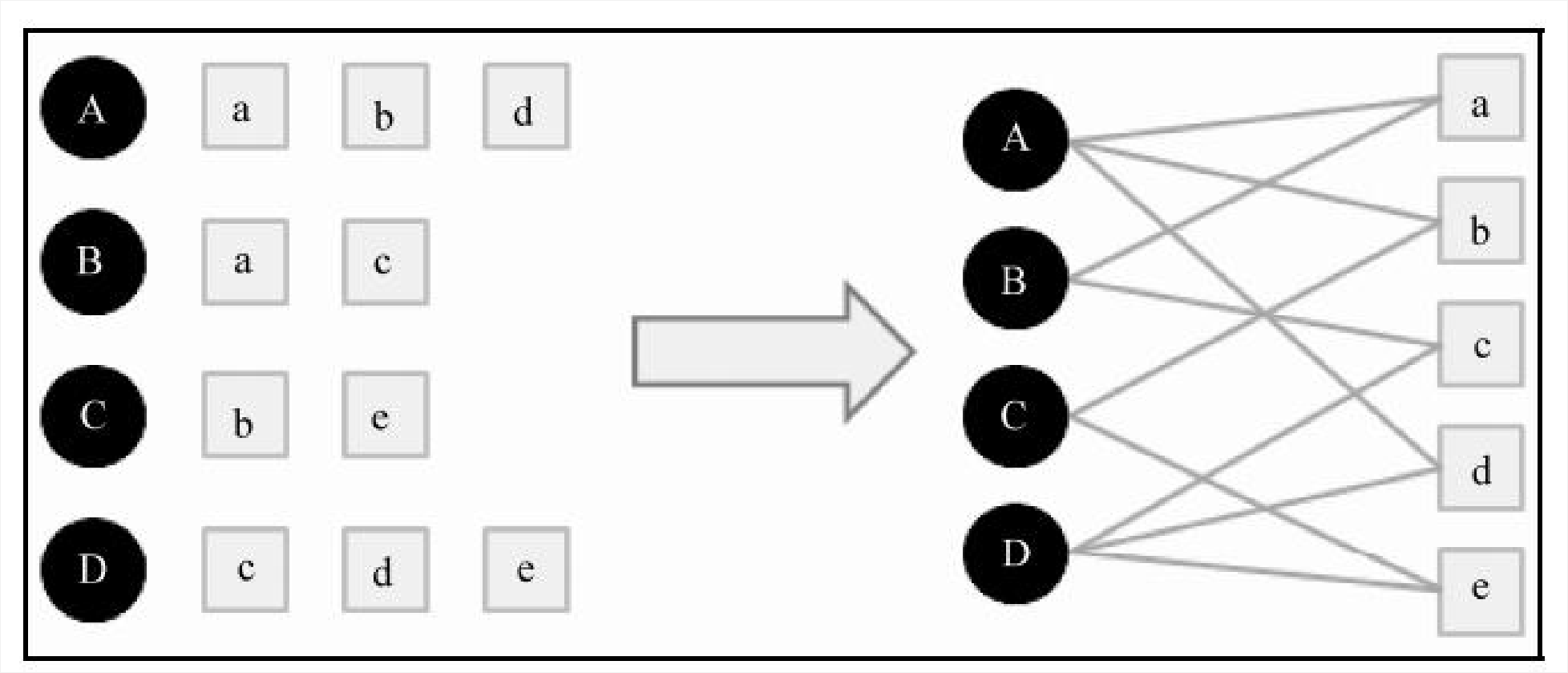
$$p(u,i)=\sum_{j} r_{uj} w_{ji}$$



4.2 利用用户行为数据：比较

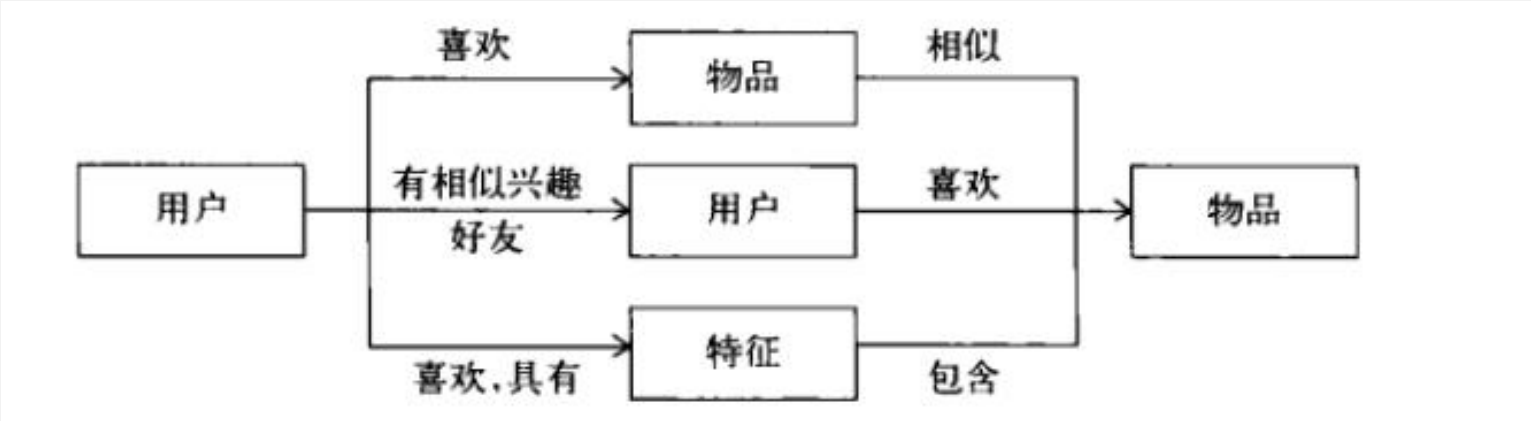
	基于用户	基于物品
性能	适用于用户较少的场合，如果用户很多，计算用户相似度矩阵代价很大	适用于物品数明显小于用户数的场合，如果物品很多（网页），计算物品相似度矩阵代价很大
领域	时效性较强，用户个性化兴趣不太明显的领域	长尾物品丰富，用户个性化需求强烈的领域
实时性	用户有新行为，不一定造成推荐结果的立即变化	用户有新行为，一定会导致推荐结果的实时变化
冷启动	在新用户对很少的物品产生行为后，不能立即对他进行个性化推荐，因为用户相似度表是每隔一段时间离线计算的	新用户只要对一个物品产生行为，就可以给他推荐和该物品相关的其他物品
推荐理由	很难提供令用户信服的推荐解释	利用用户的历史行为给用户做推荐解释，可以令用户比较信服

4.2 利用用户行为数据：衍生方法



4.3 利用用户标签的数据

- 标签是一种无层次化结构的、用来描述信息的关键词，它可以用来描述物品的语义。
  根据给物品打标签的人的不同，标签应用一般分为两种：
- 一种是让作者或者专家给物品打标签；
- 另一种是让普通用户给物品打标签，也就是UGC（User Generated Content，用户生成的内容）的标签应用。



用户为什么要打标签？
用户怎么打标签？
用户打什么样的标签？

4.3 利用用户标签的数据

打标签的动机：

- 便于内容上传者组织自己的信息（内容生成者的动机）
- 便于帮助其他用户找到信息（内容生成者的动机）
- 有些标注用于更好地组织内容，方便用户将来的查找（内容消费者的动机）
- 另一些标注用于传达某种信息，比如照片的拍摄时间和地点等。

标签的类别

- 表明物品是什么 比如是一只鸟，就会有“鸟”这个词的标签；是豆瓣的首页，就有一个标签叫“豆瓣”；是乔布斯的首页，就会有个标签叫“乔布斯”。
- 表明物品的种类 比如在**Delicious**的书签中，表示一个网页类别的标签包括 **article**（文章）、**blog**（博客）、 **book**（图书）等。
- 表明谁拥有物品 比如很多博客的标签中会包括博客的作者等信息。
- 表达用户的观点 比如用户认为网页很有趣，就会打上标签**funny**（有趣），认为很无聊，就会打上标签**boring**（无聊）。
- 用户相关的标签 比如 **my favorite**（我最喜欢的）、**my comment**（我的评论）等。
- 用户的任务 比如 **to read**（即将阅读）、**job search**（找工作）等。

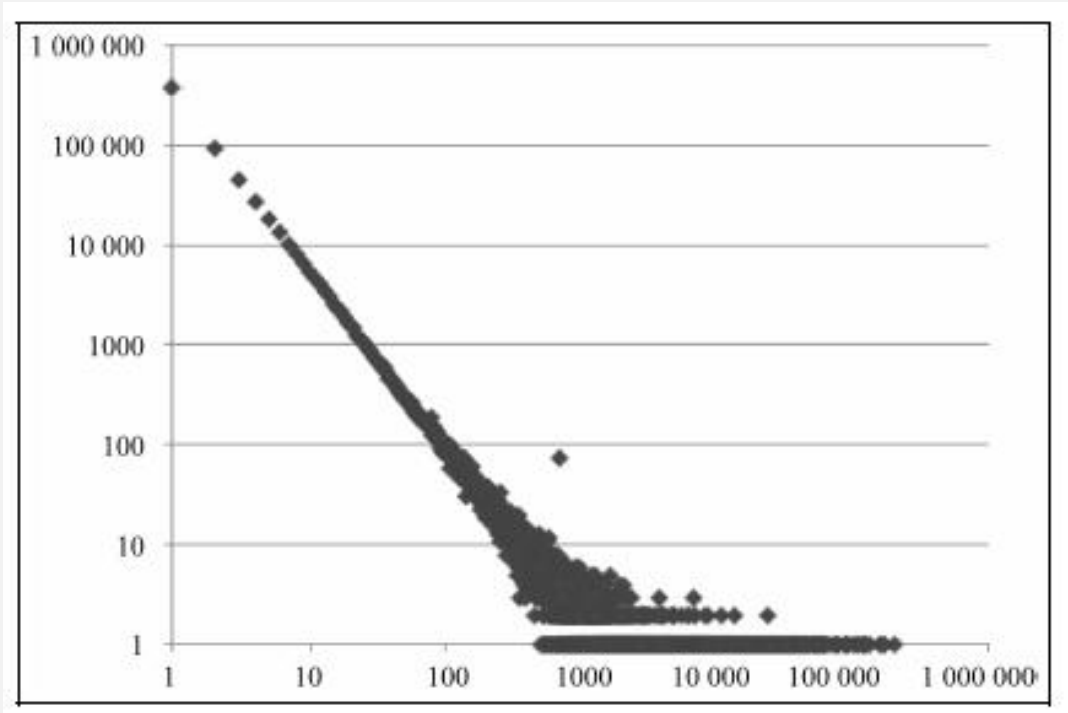
- 对于具体的领域标签可能更加专业化，例如电影网站：语言、奖项、主演等标签

4.3 利用用户标签的数据

用户、物品、标签、记录的数据量对比

	用户数	物品数	标签数	记录数
Delicious	11 200	8791	42 233	405 665
CiteULike	12 466	7318	23 068	409 220

标签也有长尾分布



4.3 利用用户标签的数据

算法：

统计每个用户最常用的标签。

对于每个标签，统计被打过这个标签次数最多的物品。

对于一个用户，首先找到他常用的标签，然后找到具有这些标签的最热门物品推荐给这个用户。

对于上面的算法，用户u对物品i的兴趣公式如下：

$$p(u,i)=\sum_{b} n_{n,b} n_{b,i}$$

4.3 利用用户标签的数据：为用户推荐标签

- 方便用户输入标签：让用户从键盘输入标签无疑会增加用户打标签的难度，这样很多用户不愿意给物品打标签，因此我们需要一个辅助工具来减小用户打标签的难度，从而提高用户打标签的参与度。
- 提高标签质量：同一个语义不同的用户可能用不同的词语来表示。这些同义词会使标签的词表变得很庞大，而且会使计算相似度不太准确。而使用推荐标签时，我们可以对词表进行选择，首先保证词表不出现太多的同义词，同时保证出现的词都是一些比较热门的、有代表性的词。

推荐策略：

1. 给用户u推荐整个系统里最热门的标签
2. 给用户u推荐物品i上最热门的标签
3. 以加权的同时推荐1、2

推荐系统实验

- 对照基于用户的协同过滤算法的伪代码，完善代码注释。
- 在样例代码的数据集上，实现基本的基于物品的协同过滤算法（IBCF）