另一种重要方式是通过一些特征联系用户和物品，给用户推荐那些具有用户喜欢的特征的物品。这里的特征有不同的表现方式，比如可以表现为物品的属性集合（比如对于图书，属性集合包括作者、出版社、主题和关键词等），也可以表现为隐语义向量（latent factor vector），这可以通过前面提出的隐语义模型习得到。本章将讨论一种重要的特征表现方式——标签。

这里主要讨论UGC的标签应用，研究用户给物品打标签的行为，探讨如何通过分析这种行为给用户进行个性化推荐。

**UGC标签的代表性应用**

Delicous可算是标签系统里的开山鼻祖，它允许用户给互联网上的每个网页打标签，从而通过标签重新组织整个互联网。

CiteULike是一个著名的论文书签网站，它允许研究人员提交或者收藏自己感兴趣的论文并且给论文打标签，从而帮助用户更好地发现和自己研究领域相关的优秀论文。

Last.fm是一家著名的音乐网站，它让用户用标签标记音乐和歌手。通过分析用户的听歌行为预测用户对音乐的兴趣，从而给用户推荐个性化的音乐。

豆瓣是中国著名的评论和社交网站之一。它允许用户对图书和电影打标签，借此获得图书和电影的内容信息和语义，并用这种信息改善推荐效果。

标签系统的最大优势在于可以发挥群体的智能，获得对物品内容信息比较准确的关键词描述。

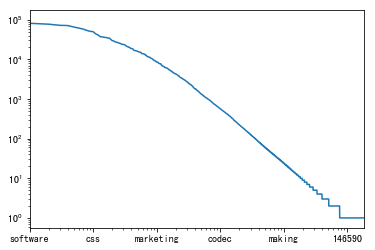
**标签系统中的推荐问题**

**用户为什么要打标签？**

社交需要/功能要求

**用户怎么打标签？**

统计结果表明，标签的分布也遵循幂律分布。



能看出来线性很强，还是比较符合幂律分布的。

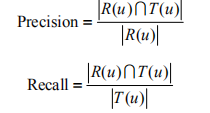
**用户打什么样的标签？**

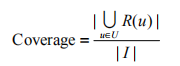
由于标签是人写的，就不可避免地会出现各种各样的不确定因素。可能出现的标签会有不同的种类和含义，需要具体问题具体分析，给标签赋予恰当的含义。

**基于标签的推荐系统**

以delicious数据集为例，通过学习训练集中的用户标签数据预测测试集上用户会给什么物品打标签。

评价指标：准确率，召回率，覆盖率和新颖度





**第一个推荐算法**

在最简单的情况下，推荐算法可以是这样的：

 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。

这时，用户u对物品i的兴趣可以被写为：



这里， n\_u,b是用户u打过标签b的次数，n\_b,i是物品i被打过标签b的次数。这里实际上就是计算了通过每一个标签b，可以从用户u“通往”标签i的路径权重之和。

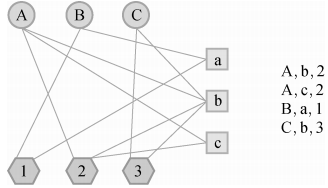
**可能的几种改进方向：**

惩罚热门物品：TF-IDF算法

克服数据的稀疏性：增加相似标签

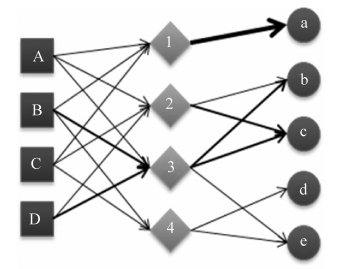
标签清理：去除无意义或负面意义的标签，提高标签的质量

**基于图的推荐算法**



这是一个用户-物品-标签图的例子，每条三折线表示一条记录，比如A-b-3就表示用户A对于物品b打了标签3。

回顾之前的算法。在简单算法中，我们把用户和物品之间的连接看成是由标签相连的“路径”。那么，既然我们使用了基于图的推荐算法，就可以尝试把它画成有向图的形式，具象地表达出路径的概念。



**基于标签的推荐解释**

基于标签的推荐其最大好处是可以利用标签做推荐解释。比方说我们使用用户相似度做协同过滤，那么解释就是“与你相似的用户也喜欢blablabla”，但是如果我们用标签推荐，就可以解释为“猜你喜欢悬疑/科幻/数据挖掘”等等，比如豆瓣的推荐界面。



同时，用户还可以自主选择标签，增加了推荐的实用程度。

关于这种方式的推荐，在推荐列表中，如何选择推荐物品的排序、如何向用户解释这些推荐结果，GroupLens的研究人员Jesse Vig有一些实验结果。

❑ 用户对标签的兴趣对帮助用户理解为什么给他推荐某个物品更有帮助；

❑ 用户对标签的兴趣和物品标签相关度对于帮助用户判定自己是否喜欢被推荐物品具有同

样的作用；

❑ 物品标签相关度对于帮助用户判定被推荐物品是否符合他当前的兴趣更有帮助；

❑ 客观事实类标签相比主观感受类标签对用户更有作用。

确实跟绕口令一样……

**给用户推荐标签**

正如上面所说过的一样，用户自由发挥的时候，谁也不知道这个标签是个什么意思，这显然并不是我们想看到的。

那怎么办呢？我们可以在用户生成标签的时候，就给他推荐一些候选答案，限制他的思维，让标签变得高质量。



**为什么要给用户推荐标签？**

方便用户输入标签/提高标签质量

**如何给用户推荐标签？**

我们有几种思路可以参考：

（0）推荐整个系统里最热门的标签【PopularTags算法】

（1）推荐这个物品最热门的标签【ItemPopularTags算法】

（2）推荐这个用户最常用的标签【UserPopularTags算法】

（3）将（1）（2）的结果加权

注意这一节和上一节的区别，这一节是预测“用户要打什么标签”，标签是预测目标。上一节是预测“用户要对什么物品打标签”，标签是预测路径。

根据书里给出的实验结果，第三种算法是最好的。这符合我们的直觉。不过，即使是最好的这种算法，依然存在着冷启动的问题。

解决这一问题有两个思路。

第一个思路是从物品的内容数据中抽取关键词作为标签。这方面的研究很多，特别是在上下文广告领域。

第二个思路是针对有结果，但结果不太多的情况。比如《MongoDB权威指南》一书只有一个用户曾经给它打过一个标签nosql，这个时刻可以做一些关键词扩展，加入一些和nosql相关的标签，比如数据库、编程等。