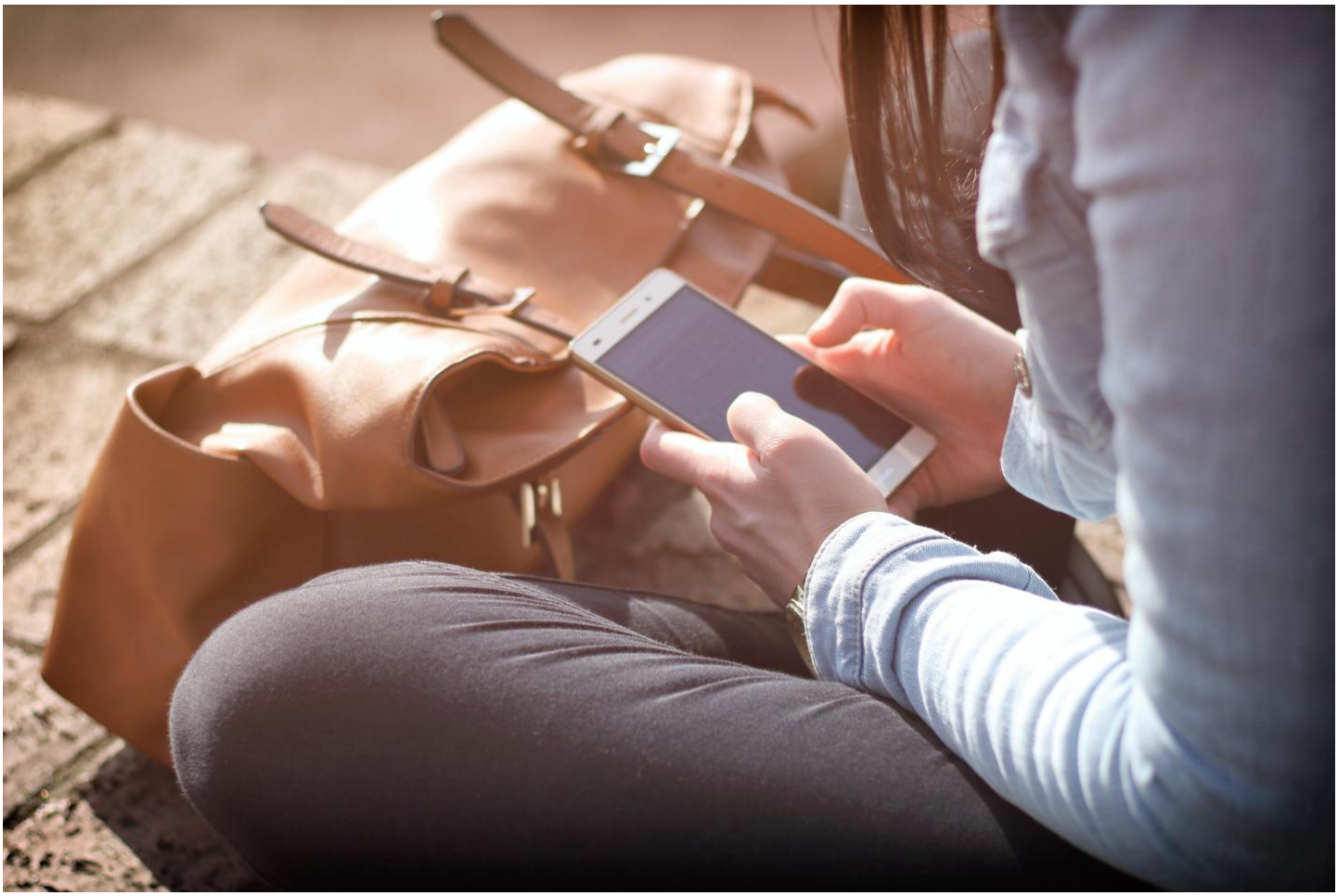
知乎



推荐系统：协同过滤及其利弊

第四范式...
已认证的官方帐号

关注他

11 人赞同了该文章

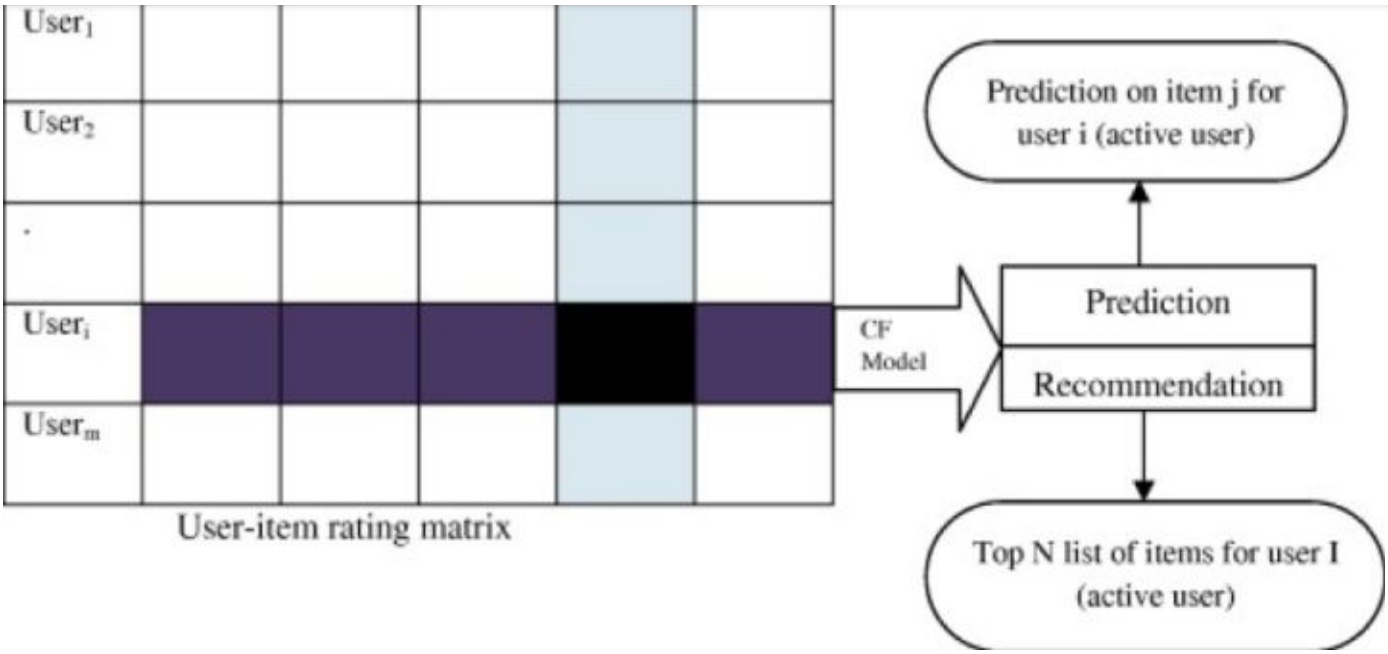
在上一篇文章中，我们谈到了**推荐系统**中基于内容的过滤及其利弊，今天我们来看看协同过滤。

与基于内容的过滤（CBF）不同，协同过滤（Collaborative Filtering）技术独立于域，适用于无法利用元数据充分描述的项目，如电影、音乐等。

协同过滤技术（CF）首先会构建用户项目偏好的数据库，即user-item矩阵，然后，计算用户画像之间的相似性，匹配具有相似的兴趣爱好的用户，完成整个推荐。这些用户获得的推荐项目，是他之前未评级但已被其它相似用户评价过的项目。

由CF生成的结果可能是预测，也可能是推荐。预测表示用户i的项目j的预测得分的数值Rij，而推荐是用户最喜欢的前N个项目的列表，如图下所示。

协同过滤可以分为两类：1）基于记忆；2）基于模型



协同过滤过程

基于记忆

在用户的搜索过程中，与他兴趣爱好相似的用户之前评价过的项目扮演着重要角色。一旦匹配到与该用户兴趣爱好相似的其他用户，就可以使用不同的算法，结合该用户和其他用户的兴趣爱好，生成推荐结果。

基于记忆的CF可以通过基于用户（user-based）和基于项目（item-based）两种技术实现。

基于用户的CF通过比较用户对同一项目的评级来计算用户之间的相似性，然后计算活跃用户对项目的预测评级，并将该预测作为类似的其他用户对项目评级的加权平均值。

基于项目的CF则利用项目之间的相似性预测结果：从用户-项目矩阵中检索活跃用户评价的所有项目，建立项目相似性的模型，计算项目之间的相似度，然后选择前K个最相似的项目，计算前K个项目的加权平均值，生成预测。

计算物品/用户之间的相似性有很多方法：

计算欧几里得距离:

$$d(x,y) = \sqrt{(\sum (x_i - y_i)^2}$$

$$\frac{1}{1+d(x,y)}$$

皮尔逊相关系数：

$$p(x,y)=\frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}=\frac{n\sum x_i y_i-\sum x_i \sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}$$

其中sx,sy表示x和y的标准差。

Cosine相似度：

$$T(x,y)=\frac{x\bullet y}{\|x\|^2\times\|y\|^2}=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

Tanimoto系数，也称作Jaccard系数：

$$T(x,y)=\frac{x\bullet y}{\|x\|^2+\|y\|^2-x\bullet y}=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}+\sqrt{\sum y_i^2}-\sum x_i y_i}$$

基于模型

基于模型的CF会使用先前的用户评级来建模，提高协同过滤的性能。建模过程可以通过机器学习或数据挖掘来完成。这些技术包括奇异值分解（SVD）、潜在语义分析、回归分析、聚类分析等。

合。缺点是难以进行模型评估，一般通过行业经验判断结果是否合理。

关联规则最经典的是购物篮分析，啤酒和尿布就是一个经典案例。运用在早期亚马逊、京东、淘宝等购物推荐场景中，往往表现为“买过这本书的人还买了XXX”，“看了这部电影的人还想看XXX”，其推荐结果包含的个性化信息较低，相对简单粗暴。

聚类分析（Clustering Analysis）

聚类技术已应用于不同的领域，如模式识别、图像处理、统计数据分析等。聚类就是将数据对象分组成为多个类或者簇 (Cluster)，从而让同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。所以，在很多应用中，一个簇中的数据对象可以被作为一个整体来对待，从而减少计算量或者提高计算质量。人们日常生活的“物以类聚，人以群分”，核心的思想就是聚类。通过聚类，人们能意识到密集和稀疏的区域，发现全局的分布模式，以及数据属性之间的有趣的相互关系。在CF中，聚类分析可以作为其他算法的预处理步骤，简化计算量，提高分析效率。

决策树（Decision Tree）

机器学习中，决策树是一个预测模型；代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若想要有复数输出，可以建立独立的决策树以处理不同的输出。从数据产生决策树的机器学习技术叫做决策树学习, 通俗说就是决策树。决策树的简单策略好比公司招聘面试，面试官筛选一个人的简历，如果候选人的各项条件都符合，那么进入初面，初面合格再进入下一轮面试。

人工神经网络（Artificial Neural Network）

人工神经网络从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也经常被简称为“神经网络”或“类神经网络”。神经网络是一种运算模型，由大量的节点（或称神经元）之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数（activation function）。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重，这相当于人工神经网络的记忆。网络的输出则依网络的连接方式、权重值和激励函数的不同而不同。网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

法，就叫回归分析。回归分析又可以分为两大类，根据因变量和自变量的个数来分类的话，可分为一元回归分析与多元回归分析；根据因变量和自变量的函数表达式来分类的话，可分为线性回归分析与非线性回归分析。

贝叶斯分类器（Bayesian Classifiers）

由于推荐问题可以看成分类问题，因此可以使用机器学习领域中的分类算法加以解决。朴素贝叶斯分类算法是贝叶斯分类算法中比较简单的一种，它的基本思想是：对于给出的待分类物品和既定的类别，计算该物品在各个类别中出现的频率，哪个类别计算出的概率大就将物品归于那个类。在推荐系统中，朴素贝叶斯分类能够在已知某些评分的情况下，通过计算概率预测未知评分。

朴素贝叶斯分类器的主要优点是对孤立的噪声点和不相关的属性具有鲁棒性，并且通过在概率估算中忽略实例来处理缺失值。素贝叶斯分类实现起来比较简单，准确率高，但是分类的时候需要学习全部样本的信息。因此，朴素贝叶斯分类适用于数据量不大，类别较少的分类问题。

协同过滤技术的优缺点

协同过滤与CBF相比，优势就是可以在根据各个用户的历史信息推荐项目，跟项目本身的内容属性无关。尽管CF技术取得了一定成功，但仍然存在一些问题：

1.冷启动问题

在产品刚刚上线、新用户到来的时候，如果没有用户在应用上的行为数据，也无法预测其兴趣爱好。另外，当新商品上架也会遇到冷启动的问题，没有收集到任何一个用户对其浏览，点击或者购买的行为，也无从对商品进行推荐。

2.数据稀疏性问题

当用户仅对数据库中可用的项目中的一小部分进行评分时，就会导致这种问题。**数据规模越大，一般而言越稀疏。**

3.可扩展性问题

这是与推荐算法相关的另一个问题，因为计算通常随着用户和项目的数量线性增长。当数据集的量

知乎

4．同义问题

同义词是指名称不同但非常相似的项目。大多数推荐系统很难区分这些项目之间的不同，如婴儿服装和婴儿布料。协同过滤通常无法在两个术语之间建立匹配，也无法计算二者之间的相似性。自动术语扩展、词库构建、奇异值分解（SVD），尤其是潜在语义索引，能够解决同义问题，但缺点是某些添加的术语可能与预期的含义不同，从而导致推荐性能的快速下降。

相关阅读：

推荐系统过滤技术：基于内容的过滤及其利弊

推荐系统的工作流程

用Python搭建推荐系统的最佳开源包

如何用Python搭建一个简单的推荐系统？

想要了解推荐系统？看这里！（2）——神经网络方法

想要了解推荐系统？看这里！（1）——协同过滤与奇异值分解

入门推荐系统，你不应该错过的知识清单

推荐系统相关术语知多少

如欲了解更多，欢迎搜索并关注先荐微信公众号（**ID：dsfsxj**）。

本账号为第四范式智能推荐产品先荐的官方账号。账号立足于计算机领域，特别是人工智能相关的前沿研究，旨在把更多与人工智能相关的知识分享给公众，从专业的角度促进公众对人工智能的理解；同时也希望为人工智能相关人员提供一个讨论、交流、学习的开放平台，从而早日让每个人都享受到人工智能创造的价值。

编辑于 2019-08-30