协同过滤推荐算法总结

推荐算法具有非常多的应用场景和商业价值，因此对推荐算法值得好好研究。推荐算法种类很多，但是目前应用最广泛的应该是协同过滤类别的推荐算法，本文就对协同过滤类别的推荐算法做一个概括总结，后续也会对一些典型的协同过滤推荐算法做原理总结。

推荐算法是非常古老的，在机器学习还没有兴起的时候就有需求和应用了。概括来说，可以分为以下5种：

1）**基于内容的推荐**：这一类一般依赖于自然语言处理NLP的一些知识，通过挖掘文本的TF-IDF特征向量，来得到用户的偏好，进而做推荐。这类推荐算法可以找到用户独特的小众喜好，而且还有较好的解释性。这一类由于需要NLP的基础，本文就不多讲，在后面专门讲NLP的时候再讨论。

2）**协调过滤推荐**：本文后面要专门讲的内容。协调过滤是推荐算法中目前最主流的种类，花样繁多，在工业界已经有了很多广泛的应用。它的优点是不需要太多特定领域的知识，可以通过基于统计的机器学习算法来得到较好的推荐效果。最大的优点是工程上容易实现，可以方便应用到产品中。目前绝大多数实际应用的推荐算法都是协同过滤推荐算法。

3）**混合推荐**：这个类似我们机器学习中的集成学习，博才众长，通过多个推荐算法的结合，得到一个更好的推荐算法，起到三个臭皮匠顶一个诸葛亮的作用。比如通过建立多个推荐算法的模型，最后用投票法决定最终的推荐结果。混合推荐理论上不会比单一任何一种推荐算法差，但是使用混合推荐，算法复杂度就提高了，在实际应用中有使用，但是并没有单一的协调过滤推荐算法，比如逻辑回归之类的二分类推荐算法广泛。

4）**基于规则的推荐**：这类算法常见的比如基于最多用户点击，最多用户浏览等，属于大众型的推荐方法，在目前的大数据时代并不主流。

5）**基于人口统计信息的推荐**：这一类是最简单的推荐算法了，它只是简单的根据系统用户的基本信息发现用户的相关程度，然后进行推荐，目前在大型系统中已经较少使用。

2. 协同过滤推荐概述

协同过滤(Collaborative Filtering)作为推荐算法中最经典的类型，包括在线的协同和离线的过滤两部分。所谓在线协同，就是通过在线数据找到用户可能喜欢的物品，而离线过滤，则是过滤掉一些不值得推荐的数据，比比如推荐值评分低的数据，或者虽然推荐值高但是用户已经购买的数据。

协同过滤的模型一般为m个物品，m个用户的数据，只有部分用户和部分数据之间是有评分数据的，其它部分评分是空白，此时我们要用已有的部分稀疏数据来预测那些空白的物品和数据之间的评分关系，找到最高评分的物品推荐给用户。

一般来说，协同过滤推荐分为三种类型。第一种是**基于用户(user-based)的协同过滤**，第二种是**基于项目(item-based)的协同过滤**，第三种是**基于模型(model based)的协同过滤**。

基于用户(user-based)的协同过滤主要考虑的是用户和用户之间的相似度，只要找出相似用户喜欢的物品，并预测目标用户对对应物品的评分，就可以找到评分最高的若干个物品推荐给用户。而基于项目(item-based)的协同过滤和基于用户的协同过滤类似，只不过这时我们转向找到物品和物品之间的相似度，只有找到了目标用户对某些物品的评分，那么我们就可以对相似度高的类似物品进行预测，将评分最高的若干个相似物品推荐

给用户。比如你在网上买了一本机器学习相关的书，网站马上会推荐一堆机器学习，大数据相关的书给你，这里就明显用到了基于项目的协同过滤思想。

我们可以简单比较下基于用户的协同过滤和基于项目的协同过滤：基于用户的协同过滤需要在线找用户和用户之间的相似度关系，计算复杂度肯定会比基于基于项目的协同过滤高。但是可以帮助用户找到新类别的有惊喜的物品。而基于项目的协同过滤，由于考虑的物品的相似性一段时间不会改变，因此可以很容易的离线计算，准确度一般也可以接受，但是推荐的多样性来说，就很难带给用户惊喜了。一般对于小型的推荐系统来说，基于项目的协同过滤肯定是主流。但是如果是大型的推荐系统来说，则可以考虑基于用户的协同过滤，当然更加可以考虑我们的第三种类型，基于模型的协同过滤。

基于模型(model based)的协同过滤是目前最主流的协同过滤类型了，我们的一大堆机器学习算法也可以在这里找到用武之地。下面我们就重点介绍基于模型的协同过滤。

3. 基于模型的协同过滤

基于模型的协同过滤作为目前最主流的协同过滤类型，其相关算法可以写一本书了，当然我们这里主要是对其思想做有一个归类概括。我们的问题是这样的m个物品，m个用户的数据，只有部分用户和部分数据之间是有评分数据的，其它部分评分是空白，此时我们要用已有的部分稀疏数据来预测那些空白的物品和数据之间的评分关系，找到最高评分的物品推荐给用户。

对于这个问题，用机器学习的思想来建模解决，主流的方法可以分为：用关联算法，聚类算法，分类算法，回归算法，矩阵分解，神经网络,图模型以及隐语义模型来解决。下面我们分别加以介绍。

3.1 用关联算法做协同过滤

一般我们可以找出用户购买的所有物品数据里频繁出现的项集活序列，来做频繁集挖掘，找到满足支持度阈值的关联物品的频繁N项集或者序列。如果用户购买了频繁N项集或者序列里的部分物品，那么我们可以将频繁项集或序列里的其他物品按一定的评分准则推荐给用户，这个评分准则可以包括支持度，置信度和提升度等。

常用的关联推荐算法有Apriori，FP Tree和PrefixSpan。如果大家不熟悉这些算法，可以参考我的另外几篇文章：

Apriori算法原理总结

FP Tree算法原理总结

PrefixSpan算法原理总结

3.2 用聚类算法做协同过滤

用聚类算法做协同过滤就和前面的基于用户或者项目的协同过滤有些类似了。我们可以按照用户或者按照物品基于一定的距离度量来进行聚类。如果基于用户聚类，则可以将用户按照一定距离度量方式分成不同的目标人群，将同样目标人群评分高的物品推荐给目标用户。基于物品聚类的话，则是将用户评分高物品的相似同类物品推荐给用户。

常用的聚类推荐算法有K-Means, BIRCH, DBSCAN和谱聚类，如果大家不熟悉这些算法，可以参考我的另外几篇文章：

K-Means聚类算法原理

BIRCH聚类算法原理

DBSCAN密度聚类算法

谱聚类（spectral clustering）原理总结

3.3 用分类算法做协同过滤

如果我们根据用户评分的高低，将分数分成几段的话，则这个问题变成分类问题。比如最直接的，设置一份评分阈值，评分高于阈值的就是推荐，评分低于阈值就是不推荐，我们将问题变成了一个二分类问题。虽然分类问题的算法多如牛毛，但是目前使用最广泛的是逻辑回归。为啥是逻辑回归而不是看起来更加高大上的比如支持向量机呢？因为逻辑回归的解释性比较强，每个物品是否推荐我们都有一个明确的概率放在这，同时可以对数据的特征做工程化，得到调优的目的。目前逻辑回归做协同过滤在BAT等大厂已经非常成熟了。

常见的分类推荐算法有逻辑回归和朴素贝叶斯，两者的特点是解释性很强。如果大家不熟悉这些算法，可以参考我的另外几篇文章：

逻辑回归原理小结

朴素贝叶斯算法原理小结

3.4 用回归算法做协同过滤

用回归算法做协同过滤比分类算法看起来更加的自然。我们的评分可以是一个连续的值而不是离散的值，通过回归模型我们可以得到目标用户对某商品的预测打分。

常用的回归推荐算法有Ridge回归，回归树和支持向量回归。如果大家不熟悉这些算法，可以参考我的另外几篇文章：

线性回归原理小结

决策树算法原理(下)

支持向量机原理(五)线性支持回归

3.5 用矩阵分解做协同过滤

用矩阵分解做协同过滤是目前使用也很广泛的一种方法。由于传统的奇异值分解SVD要求矩阵不能有缺失数据，必须是稠密的，而我们的用户物品评分矩阵是一个很典型的稀疏矩阵，直接使用传统的SVD到协同过滤是比较复杂的。

目前主流的矩阵分解推荐算法主要是SVD的一些变种，比如FunkSVD，BiasSVD和SVD++。这些算法和传统SVD的最大区别是不再要求将矩阵分解为
$U\Sigma V^T$
的形式，而变是两个低秩矩阵
P^TQ
的乘积形式。对于矩阵分解的推荐算法，后续我会专门开篇来讲。

3.6 用神经网络做协同过滤

用神经网络乃至深度学习做协同过滤应该是以后的一个趋势。目前比较主流的用两层神经网络来做推荐算法的是限制玻尔兹曼机(RBM)。在目前的Netflix算法比赛中， RBM算法的表现很牛。当然如果用深层的神经网络来做协同过滤应该会更好，大厂商用深度学习的方法来做协同过滤应该是将来的一个趋势。后续我会专门开篇来讲讲RBM。

3.7 用图模型做协同过滤

用图模型做协同过滤，则将用户之间的相似度放到了一个图模型里面去考虑，常用的算法是SimRank系列算法和马尔科夫模型算法。对于SimRank系列算法，它的基本思想是被相似对象引用的两个对象也具有相似性。算法思想有点类似于大名鼎鼎的PageRank。而马尔科夫模型算法当然是基于马尔科夫链了，它的基本思想是基于传导性来找出普通距离度量算法

难以找出的相似性。后续我会专门开篇来讲讲SimRank系列算法。

3.8 用隐语义模型做协同过滤

隐语义模型主要是基于NLP的，涉及到对用户行为的语义分析来做评分推荐，主要方法有隐性语义分析LSA和隐含狄利克雷分布LDA，这些等讲NLP的再专门讲。

4. 协同过滤的一些新方向

当然推荐算法的变革也在进行中，就算是最火爆的基于逻辑回归推荐算法也在面临被取代。哪些算法可能取代逻辑回归之类的传统协同过滤呢？下面是我的理解：

a) **基于集成学习的方法和混合推荐**:这个和混合推荐也靠在一起了。由于集成学习的成熟，在推荐算法上也有较好的表现。一个可能取代逻辑回归的算法是GBDT。目前GBDT在很多算法比赛都有好的表现，而有工业级的并行化实现类库。

b)**基于矩阵分解的方法**：矩阵分解，由于方法简单，一直受到青睐。目前开始渐渐流行的矩阵分解方法有分解机(Factorization Machine)和张量分解(Tensor Factorization)。

c) **基于深度学习的方法**：目前两层的神经网络RBM都已经有非常好的推荐算法效果，而随着深度学习和多层神经网络的兴起，以后可能推荐算法就是深度学习的天下了？目前看最火爆的是基于CNN和RNN的推荐算法。

5. 协同过滤总结

协同过滤作为一种经典的推荐算法种类，在工业界应用广泛，它的优点很多，模型通用性强，不需要太多对应数据领域的专业知识，工程实现简单，效果也不错。这些都是它流行的原因。

当然，协同过滤也有些难以避免的难题，比如令人头疼的“冷启动”问题，我们没有新用户任何数据的时候，无法较好的为新用户推荐物品。同时也没有考虑情景的差异，比如根据用户所在的场景和用户当前的情绪。当然，也无法得到一些小众的独特喜好，这块是基于内容的推荐比较擅长的。

以上就是协同过滤推荐算法的一个总结，希望可以帮大家对推荐算法有一个更深的认识，并预祝大家新年快乐！

分类: 0081. 机器学习