推荐系统

我将介绍两部分内容，先是推荐系统的理解，这里主要从推荐系统的结构上面切入；然后是对论文中遇到的问题进行归类整理。

在推荐系统结构上面，我参考了一本2020年出版的《深度学习推荐系统》一书，书中推荐系统的逻辑结构大致如图所示。从图中可以看出，推荐系统的功能是将候选物品集中的物品进行筛选推荐给用户，而这个过程在如今的推荐系统中一般是分为两部分的，一部分是召回层，一部分是精细推荐层。由于在这个大数据时代数据集太大，即使是线性的推荐算法也会非常耗时，因此，设置召回层就是希望将这些耗时的操作进行简化，但又不能影响模型的性能，因此在召回层后面又会加入一层精细推荐层，以达到更好的效果。如何让召回层能够很快的从海量的数据中筛选出候选子集成为一个问题，因为要同时考虑效率和准确率。在目前，需要满足这种要求的一般会采用局部敏感哈希算法。局部敏感哈希算法跟哈希算法一样，属于一种映射思想，其将高维空间中的数据映射到低维空间，一般是一维空间，从而在一维空间中比较两个数据之间的相似程度。这里依据的一个原则是，在高维空间中相近的点肯定在低维空间中的投影相近，虽然在高维空间中距离较远的点也可能在低维空间中投影相近，但是这些点是很难通过精细筛选层的，因此使用局部敏感哈希算法对降低模型的准确性没有太大的影响，但由于一维空间距离的计算很高效，因此算法效率很高。当然，在召回层层想仅使用一种算法就达到很高的准确率是很难的，因此实际使用中往往是多种算法相结合，再取这些集合的可取之处作为子集。而精细筛选层往往是一些论文的讨论对象，因为其中含有很多可以调整的地方，比如模型结构等等。在推荐模型中，往往将用户信息、物品信息和一些场景信息作为输入，经过模型寻找出合适的推荐物品。从这个意义上来说，其实我们经常使用的搜索引擎其实就是一个推荐系统，虽然他只使用了用户输入的词条作为输入。当然，在目前的浏览器中已经有了一些如关键字自动补全的推荐之类的功能。

然后，更仔细的看这个模型，会发现，其中对输入信息需要采取一项成为特征工程的方式进行处理。这个特征工程其实主要是将输入信息转化为向量嵌入。因为，在目前的推荐系统中，可以使用的数据包括两类，一类是显式信息，比如用户的那些评论，还有一类是隐式信息，比如用户的点击或浏览记录。不管是哪种信息，最终都需要转化为数字才能被推荐模型处理，因此，特征工程的目的就是需要将这些原始的信息转化为数字信息，及高维向量。但由于实验发现，高维向量的转化工作其实也会影响到模型的性能，因此，如果能让这些向量之间蕴含数据之间的关系最后会提高模型的表现。对于显式数据而言，比如评论，作为文本，可以采用词向量转化模型来进行处理，而对于隐式数据而言，比如用户点击次数，则需要采用embedding层进行转化，因为这些隐式数据是很稀疏的。接着这个模型往下，会看到模型的训练分为在线训练和离线训练两种，在线训练目前可以采用的强化学习的方式，利用用户的反馈通过强化学习模型来优化参数，离线学习的研究较早，可采用的方式很多，比如阿里的GIN模型，使用深度学习的方式来寻找输入信息向量之间的相关关系。说到离线学习，就不得不提经常会遇到的协同过滤算法了。协同过滤算法虽然几乎是最早的推荐算法，但在目前仍然有很多用途，而且在论文中也经常会见到。协同过滤算法是假设以前表现相似的事物在将来的表现也会相似。意思是说，如果两个人A和B在之前一直买了很多一样的物品，那么在将来如果A买了物品X，那么B很可能也会买物品X，因此，作为推荐系统，将物品X推荐给B就是情理之中的事情了。这种情况属于用户协同过滤，还有一种是物品协同过滤，就是说，如果物品C和D往往是同时或者接连着被卖出去，那么当用户购买了物品C时，很有可能也会购买D，因此，作为推荐系统，将物品D推荐给用户就可能提高效益。虽然，用户协同过滤和物品协同过滤差不多，但是，由于用户集的变化比物品集的变化要更快，因此，在工程实际上考虑，物品协同过滤其实要比用户协同过滤用到广泛一些，而且，在可解释性上，根据后面要讲到的一篇论文所提到的，物品协同过滤比用户协同过滤更具有可解释性，因为买了相似物品的用户之间并不认识，没有必要因为一个不认识的人也买了什么物品而去买那件物品，然而，因为在物品C被购买后D往往会被购买，因此物品C和D之间应该存在某种关系，比如互补关系（像牙刷和牙膏那种），这样用户才更可能去考虑这件物品。在协同过滤算法的排序层过后，可以看到还有一个补充层，其中有个经常被讨论到的问题就是冷启动问题。冷启动问题属于在用户或物品刚加入推荐系统时，什么参考信息都没有，这时候推荐模型可能会把这些物品认为是冷门物品而不给予推荐，或者是推荐系统对新来的用户一无所知，不知道该推荐什么才好，这样会大大降低推荐系统的表现。因此，如何处理冷启动问题也是被很多论文讨论过的问题。从这个角度上来说，推荐系统其实也是一种预测系统，根据已有的信息来进行预测。

接下来回到论文方面，我把看过的论文大致归为五类。其中，有一部分是介绍论文提出的新思想和新概念之类的，这种文章目前我还不知道对我有什么用，可能以后需要用到再去研究吧。然后是对模型进行讨论的，这里就包括一些模型结构或超参数的讨论什么的，感觉有些可取之处。第三部分是设计在线学习方面的，这里我看到的那些论文没有涉及强化学习算法，只是在一些问题上提出了改进方法。第四部分是关于数据集方面的，这一部分的内容特别广，而且在我看来也有很多有营养的想法，所以我打算着重介绍这一部分的论文。最后一部分是关于如何让模型理解用户的个人爱好的问题，之所以放在最后是因为好多的论文在设计模型时都没有明确的考虑用户的爱好问题，因此作为区别于之前论文的一个补充我把这部分放在了最后。

（自由发挥，念完前面三节）

第四部分主要有五个方面的问题，分别是文本语义理解问题，产品的词向量建模问题，数据集的偏差消除问题和模糊数据的处理问题以及冷启动问题。（参考word文档进行讲解）

最后是关于用户喜好的两篇论文，一篇直接对用户喜好进行建模，这里作为用户喜好的依据是用户的评论，另一篇将用户喜好加入到搜索结果中，使最后的搜索结果对特定用户表现不同的效果，实现个性化。（按照ppt进行讲解）。

我的分享就是这么多，老师有什么指教吗？