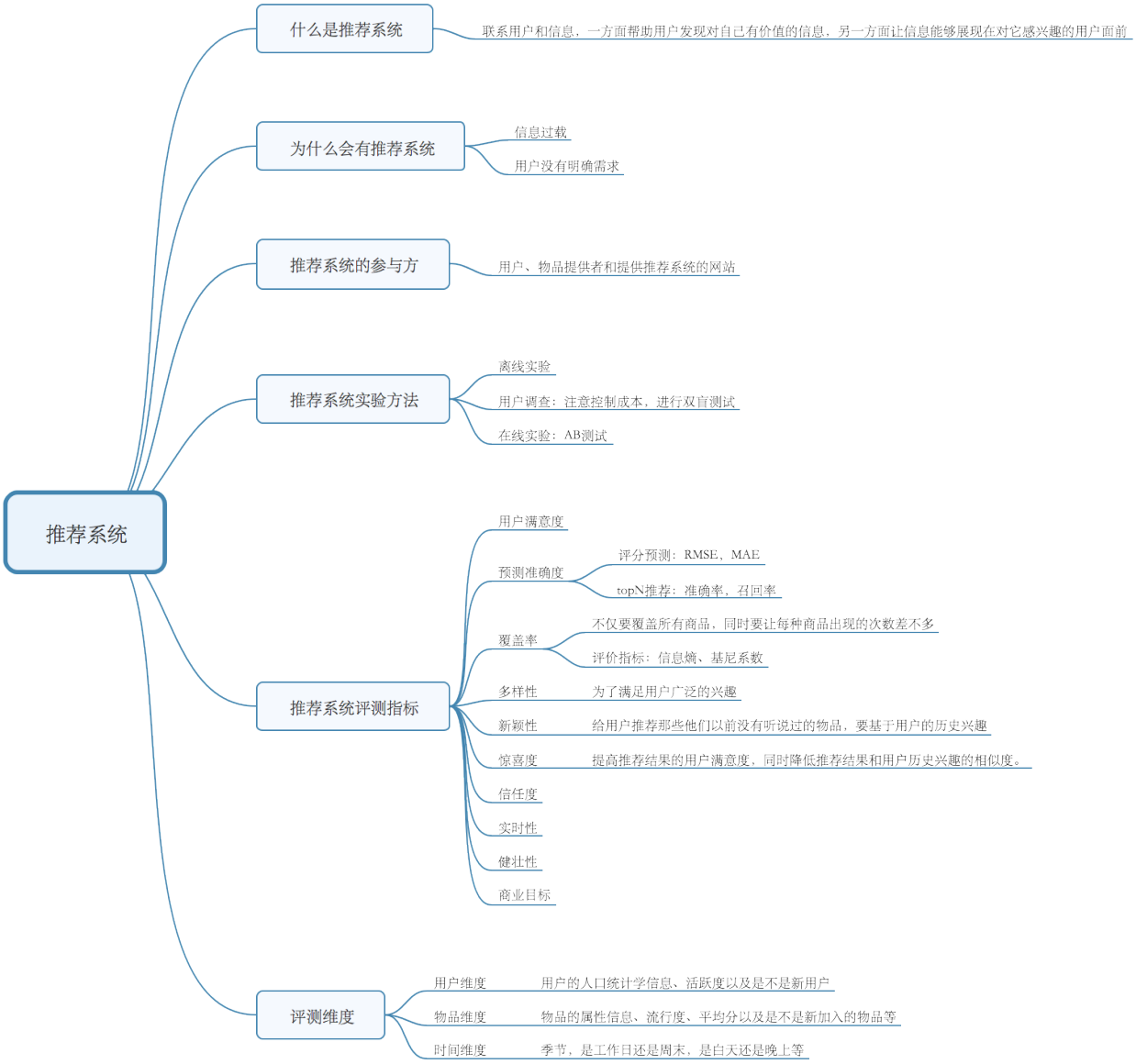
1 思维导图

- 《推荐系统实践--项亮》第一章的学习笔记



2. 推荐系统理解

- 推荐系统的任务就是联系用户和信息，一方面帮助用户发现对自己有价值的信息，另一方面让信息能够展现在对它感兴趣的用户面前，从而实现信息消费者和信息生产者的双赢。和搜索引擎不同的是，推荐系统不需要用户提供明确的需求，而是通过分析用户的历史行为给用 户的兴趣建模，从而主动给用户推荐能够满足他们兴趣和需求的信息。
- 个性化推荐的成功应用需要两个条件。第一是存在**信息过载**，因为如果用户可以很容易地从所有物品中找到喜欢的物品，就不需要个性化推荐了。第二是**用户大部分时候没有特别明确的需求**，因为用户如果有明确的需求，可以直接通过搜索引擎找到感兴趣的物品。
- 一个完整的推荐系统一般存在3个参与方:**用户、物品提供者和提供推荐系统的网站**。以图书推荐为例， 首先，推荐系统需要满足用户的需求，给用户推荐那些令他们感兴趣的图书。其次，推荐系统要让各出版社的书都能够被推荐给对其感兴趣的用户，而不是只推荐几个大型出版社的书。最后，

好的推荐系统设计，能够让推荐系统本身收集到高质量的用户反馈，不断完善推荐的质量，增加 用户和网站的交互，提高网站的收入。因此在评测一个推荐算法时，需要同时考虑三方的利益， 一个好的推荐系统是能够令三方共赢的系统。

3. 推荐系统实验方法

- 主要3中评测推荐效果的实验方法：离线实验、用户调查、在线实验

3.1 离线实验

- 离线实验的方法一般由如下几个步骤构成:
  - 通过日志系统获得用户行为数据，并按照一定格式生成一个标准的数据集;
  - 将数据集按照一定的规则分成训练集和测试集;
  - 在训练集上训练用户兴趣模型，在测试集上进行预测;
  - 通过事先定义的离线指标评测算法在测试集上的预测结果。

优　　点	缺　　点
不需要有对实际系统的控制权	无法计算商业上关心的指标
不需要用户参与实验	离线实验的指标和商业指标存在差距
速度快，可以测试大量算法	

3.2 用户调查

- 用户调查是推荐系统评测的一个重要工具，很多离线时没有办法评测的与用户主观感受有关的指标都可以通过用户调查获得。在用户调查中，有一些需要注意的事项：
  - **成本控制**：用户调查成本很高，需要用户花大量时间完成一个个任务，并回答相关的问题。有些时候，还需要花钱雇用测试用户。因此，大多数情况下很难进行大规模的用户调查，而对于参加人数较少的用户调查，得出的很多结论往往没有统计意义。因此，我们在做用户调查时，一方面要控制成本，另一方面又要保证结果的统计意义。
  - **双盲实验**：即不要让实验人员和用户事先知道测试的目标，以免用户的回答和实验人员的测试受主观成分的影响。
  - **相同分布**：测试用户需要尽量保证测试用户的分布和真实用户的分布相同，比如男女各半，以及年龄、活跃度的分布都和真实用户分布尽量相同。
- 用户调查的优缺点也很明显。它的**优点**是可以获得很多体现用户主观感受的指标，相对在线实验风险很低，出现错误后很容易弥补。**缺点**是招募测试用户代价较大，很难组织大规模的测试 用户，因此会使测试结果的统计意义不足。此外，在很多时候设计双盲实验非常困难，而且用户在测试环境下的行为和真实环境下的行为可能有所不同，因而在测试环境下收集的测试指标可能在真实环境下无法重现。

3.3 在线实验

- 在完成离线实验和必要的用户调查后，可以将推荐系统上线做**AB测试**，将它和旧的算法进行比较。
- AB测试是一种很**常用的在线评测算法的实验方法。它通过一定的规则将用户随机分成几组，并对不同组用户采取不同的算法，然后通过统计不同组用户的各种不同的评测指标比较不同算法的好坏。**

- AB测试的**优点**是可以公平获得不同算法实际在线时的性能指标，包括商业上关注的指标。 AB测试的**缺点**主要是周期比较长，必须进行长期的实验才能得到可靠的结果。因此一般不会用 AB测试测试所有的算法，而只是用它测试那些在离线实验和用户调查中表现很好的算法。其次， 一个大型网站的AB测试系统的设计也是一项复杂的工程。
- 一般来说，一个新的推荐算法最终上线，需要完成上面所说的3个实验。
    - 首先，需要通过离线实验证明它在很多离线指标上优于现有的算法。
    - 然后，需要通过用户调查确定它的用户满意度不低于现有的算法。
    - 最后，通过在线的AB测试确定它在我们关心的指标上。

4. 评价指标

4.1 用户满意度

- 用户作为推荐系统的重要参与者，其满意度是评测推荐系统的最重要指标。但是，用户满意度没有办法离线计算，只能通过用户调查或者在线实验获得。
- **在线系统中，用户满意度主要通过一些对用户行为的统计得到**。比如在电子商务网站中，用户如果购买了推荐的商品，就表示他们在一定程度上满意。因此，我们可以利用购买率度量用户的满意度。此外，有些网站会通过设计一些用户反馈界面收集用户满意度。比如在视频网站中，都有对推荐结果满意或者不满意的反馈按钮，通过统计两种按钮的单击情况就可以度量系统的用户满意度。更一般的情况下，我们可以用点击率、用户停留时间和转化率等指标度量用户的满意度。

4.2 预测准确度

- 预测准确度度量一个推荐系统或者推荐算法预测用户行为的能力。这个指标是最重要的推荐系统离线评测指标。**在计算该指标时需要有一个离线的数据集，该数据集包含用户的历史行为记录。然后，将该数据集通过时间分成训练集和测试集。最后，通过在训练集上建立用户的行为和兴趣模型预测用户在测试集上的行为，并计算预测行为和测试集上实际行为的重合度作为预测准确度**。预测准确度指标有分为以下几种：
- **评分预测**
- 预测用户对物品评分的行为称为**评分预测**，在评分预测中，预测准确度一般通过均方根误差RMSE和 平均绝对误差MAE计算，对于测试集中的一个用户u和物品i，令rui是用户u对物品i的实际评分，而rˆui 是推荐算法给出的预测评分，RMSE和MAE的定义分别为:

$$\mathrm{RMSE}=\frac{\sqrt{\sum_{u,i\in T}\left(r_{ui}-\hat{r}_{ui}\right)^2}}{|T|}$$

$$\mathrm{MAE}=\frac{\sum_{u,j\in T}\left|r_{ui}-\hat{r}_{ui}\right|}{|T|}$$

- **TopN推荐**
- 网站在提供推荐服务时，一般是给用户一个个性化的推荐列表，这种推荐叫做TopN推荐。TopN推荐的预测准确率一般通过准确率(precision)/召回率(recall)度量。
- 令R(u)是根据用户在训练集上的行为给用户作出的推荐列表，而T(u)是用户在测试集上的行为列表。那么，推荐结果的召回率Recall 和准确率Precision定义为:

$$\mathrm{Recall}\ =\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|}$$

$$\text{Precision} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|}$$

- 有时为了全面评测TopN推荐的准确率和召回率，一般会选取不同的N，计算一组准确率和召回率，然后画出相应的曲线。
- **覆盖率**
- 覆盖率(coverage)描述**一个推荐系统对物品长尾的发掘能力**。覆盖率有不同的定义方法，**最简单的定义为推荐系统能够推荐出来的物品占总物品集合的比例**。假设系统的用户集合为U，推荐系统给每个用户推荐一个长度为N的物品列表R(u)。那么推荐系统的覆盖率可以通过下面的公式计算:

$$\text{Coverage} = \frac{|\bigcup_{u \in \text{U}} R(u)|}{|I|}$$

- 从上面的定义也可以看到，热门排行榜的推荐覆盖率是很低的，它只会推荐那些热门的物品，这些物品在总物品中占的比例很小。**一个好的推荐系统不仅需要有比较高的用户满意度，也要有较高的覆盖率**。
- 但是上面的定义过于粗略。覆盖率为100%的系统可以有无数的物品流行度分布。为了更细致地描述推荐系统发掘长尾的能力，需要统计推荐列表中不同物品出现次数的分布。**如果所有的物品都出现在推荐列表中，且出现的次数差不多，那么推荐系统发掘长尾的能力就很好**。因此， 可以通过研究物品在推荐列表中出现次数的分布描述推荐系统挖掘长尾的能力。如果这个分布比较平，那么说明推荐系统的覆盖率较高，而如果这个分布较陡峭，说明推荐系统的覆盖率较低。
- 在信息论和经济学中有两个著名的指标可以用来定义覆盖率。信息熵H和基尼系数G：p(i) 表示物品i的流行度除以所有物品流行度之和；i_j 表示按照物品流行度p()从小到大排序的物品列表中第j个物品；

$$H = -\sum_{i=1}^{n} p(i) \log p(i)$$

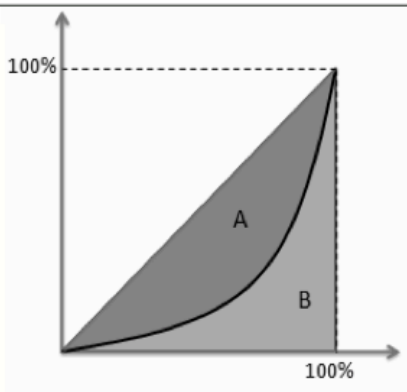
$$G = \frac{1}{n-1} \sum_{j=1}^{n} (2j - n - 1) p\left(i_j\right)$$

基尼系数的计算原理

首先，我们将物品按照热门程度从低到高排列，那么右图中的黑色曲线表示最不热门的x%物品的总流行度占系统的比例y%。这条曲线肯定是在$y=x$曲线之下的，而且和$y=x$曲线相交在(0,0)和(1,1)。

令SA是A的面积，SB是B的面积，那么基尼系数的形象定义就是SA / (SA + SB)，从定义可知，基尼系数属于区间[0,1]。

如果系统的流行度很平均，那么SA就会很小，从而基尼系数很小。如果系统物品流行度分配很不均匀，那么SA就会很大，从而基尼系数也会很大。



社会学领域有一个著名的马太效应，即所谓强者更强，弱者更弱的效应。如果一个系统会增大热门物品和非热门物品的流行度差距，让热门的物品更加热门，不热门的物品更加不热门，那么这个系统就有马太效应。比如，首页的热门排行榜就有马太效应。进入排行榜的都是热门的物品，但它们因为被放在首页的排行榜展示有了更多的曝光机会，所以会更加热门。相反，没有进入排行榜的物品得不到展示，就会更不热门。搜索引擎的PageRank算法也具有一定的马太效应，如果一个网页的某个热门关键词排名很高，并因此被展示在搜索结果的第一条，那么它就会获得更多的关注，从而获得更多的外链，PageRank排名也越高。

那么，推荐系统是否有马太效应呢？推荐系统的初衷是希望消除马太效应，使得各种物品都能被展示给对它们感兴趣的某一类人群。但是，很多研究表明现在主流的推荐算法（比如协同过滤算法）是具有马太效应的。评测推荐系统是否具有马太效应的简单办法就是使用基尼系数。如果$G1$是从初始用户行为中计算出的物品流行度的基尼系数，$G2$是从推荐列表中计算出的物品流行度的基尼系数，那么如果$G2 > G1$，就说明推荐算法具有马太效应。

- **多样性**

- 为了满足用户广泛的兴趣，推荐列表需要能够覆盖用户不同的兴趣领域，即推荐结果需要具有多样性。多样性推荐列表的好处用一句俗话表述就是“不在一棵树上吊死”。尽管用户的兴趣在较长的时间跨度中是一样的，但具体到用户访问推荐系统的某一刻， 其兴趣往往是单一的，那么**如果推荐列表只能覆盖用户的一个兴趣点，而这个兴趣点不是用户这个时刻的兴趣点，推荐列表就不会让用户满意。反之，如果推荐列表比较多样，覆盖了用户绝大多数的兴趣点，那么就会增加用户找到感兴趣物品的概率。因此给用户的推荐列表也需要满足用户广泛的兴趣，即具有多样性。**

- 多样性描述了推荐列表中物品两两之间的不相似性。因此，多样性和相似性是对应的。假设s(i, j)∈[0,1] 定义了物品i和j之间的相似度，那么**用户u的推荐列表R(u)的多样性**定义如下：

$$\text{Diversity} = 1 - \frac{\sum_{i,j \in R(u), i \neq j} s(i,j)}{\frac{1}{2}|R(u)|(|R(u)|-1)}$$

- **推荐系统的整体多样性可以定义为所有用户推荐列表多样性的平均值：**

$$\text{Diversity} = \frac{1}{|U|}\sum_{u \in U} \text{Diversity}\left(R(u)\right)$$

- **新颖性**

- 新颖的推荐是指给用户推荐那些他们以前没有听说过的物品。在一个网站中**实现新颖性**的最简单办法是，把那些用户之前在网站中对其有过行为的物品从推荐列表中过滤掉。比如在一个视频网站中，新颖的推荐不应该给用户推荐那些他们已经看过、打过分或者浏览过的视频。
- 评测新颖度的最简单方法是利用推荐结果的平均流行度，因为越不热门的物品越可能让用户觉得新

颖。因此，如果推荐结果中物品的平均热门程度较低，那么推荐结果就可能有比较高的新颖性。

- **惊喜度**
- 惊喜度(serendipity)是最近这几年推荐系统领域最热门的话题。**如果推荐结果和用户的历史兴趣不相似，但却让用户觉得满意，那么就可以说推荐结果的惊喜度很高**，而推荐的新颖性仅仅取决于用户是否听说过这个推荐结果。提高推荐惊喜度需要提高推荐结果的用户满意度，同时降低推荐结果和用户历史兴趣的相似度。

- **信任度**
- 度量推荐系统的信任度只能通过问卷调查的方式，询问用户是否信任推荐系统的推荐结果。
- 提高推荐系统的信任度主要有两种方法。**首先需要增加推荐系统的透明度(transparency)， 而增加推荐系统透明度的主要办法是提供推荐解释**。只有让用户了解推荐系统的运行机制，让用户认同推荐系统的运行机制，才会提高用户对推荐系统的信任度。**其次是考虑用户的社交网络信息，利用用户的好友信息给用户做推荐，并且用好友进行推荐解释**。这是因为用户对他们的好友一般都比较信任，因此如果推荐的商品是好友购买过的，那么他们对推荐结果就会相对比较信任。
- **实时性**
- 在很多网站中，因为物品(新闻、微博等)具有很强的时效性，所以需要在物品还具有时效性时就将它们推荐给用户。推荐系统的实时性包括两个方面。首先，**推荐系统需要实时地更新推荐列表来满足用户新的行为变化。实时性的第二个方面是推荐系统需要能够将新加入系统的物品推荐给用户**。这主要考验了推荐系统处理物品冷启动的能力。

- **健壮性**
- 健壮性(即robust,鲁棒性)指标衡量了一个推荐系统抗击作弊的能力。算法健壮性的评测主要利用模拟攻击。首先，给定一个数据集和一个算法，可以用这个算法给这个数据集中的用户生成推荐列表。然后，用常用的攻击方法向数据集中注入噪声数据，然后利用算法在注入噪声后的数据集上再次给用户生成推荐列表。最后，通过比较攻击前后推荐列表 的相似度评测算法的健壮性。如果攻击后的推荐列表相对于攻击前没有发生大的变化，就说明算法比较健壮。
- **商业目标**
- 网站评测推荐系统更加注重网站的商业目标是否达成，而商业目标和网站的盈利模式是息息相关的。

5. 评价维度

- 一个推荐算法， 虽然整体性能不好，但可能在某种情况下性能比较好，而增加评测维度的目的就是知道一个算法在什么情况下性能最好。这样可以为融合不同推荐算法取得最好的整体性能带来参考：
    - **用户维度**：主要包括用户的人口统计学信息、活跃度以及是不是新用户等。
    - **物品维度**：包括物品的属性信息、流行度、平均分以及是不是新加入的物品等。
    - **时间维度**：包括季节，是工作日还是周末，是白天还是晚上等

- 如果能够在推荐系统评测报告中包含不同维度下的系统评测指标，就能帮我们全面地了解推荐系统性能，找到一个看上去比较弱的算法的优势，发现一个看上去比较强的算法的缺点。