21120491：高级数据结构与算法分析 **2023-2024 学年春夏学期**

Lecture 3: 倒排索引

教师：刘金飞，助教：吴一航 日期: 2024 年 3 月 12 日

3.1 内容概述

本讲是这门课程中最轻松的一讲。本讲的目标是应用我们之前学习的数据结构，讨论关于搜索引擎的设计问题。本讲主要涉及了如下主题：

1. 倒排索引的引入和定义：从两个 naive 的想法出发，其一是遍历搜索，这样太耗时间；其二是稀疏矩阵，这样存储比较浪费空间。所以我们改进为链表存储，这就是倒排索引。注意在倒排索引里面我们保存单词的出现次数是因为当多个单词同时搜索时，从出现最少的词入手搜索会更快。
2. 如何构建倒排索引：逐个词语读入插入构建。其中会有很多问题，例如分词，stemming，stop words 等，还有通过搜索树或 hash 访问等；除此之外还有存储上的考量，因为内存不够需要存储到外存，外存可以分布式存储（两种方式），然后还有更新时可以用 cache 等改进存储效率。
3. 搜索引擎的评价：区分 Data Retrieval 和 Information Retrieval，了解准确率和召回率两个重要的衡量参数（与此对应还有假阳性和假阴性），其中阈值的设置是重要的影响因素。

希望进一步了解的同学可以参考学在浙大上传的压缩包，其中有相关论文和参考资料等。因为本讲内容太宽泛，因此细节不在此赘述，对 PPT 中写得不够完善的部分也可以参考陈越老师的 MOOC。

3.2 讨论

本节讨论题分为必做和选做两个部分，其中**必做题每一组同学都要全部完成**；关于选做题，因为本节实际上介绍的是更广泛的领域——信息检索——在搜索引擎方面很多年以前的研究成果，所以这个我们可以给出不同方向，或者更新的话题供同学们自行探索。选做题请每个小组从三个话题中**选取最感兴趣一个**进行自由探索（多做不加分）。

必做题：课堂上我们讨论了准确率（Precision）和召回率（Recall）这两个评价搜索引擎的指标，了解了它们之间的 tradeoff，那么你们认为有什么方法可以同时提高这两个比率吗？这里只需写出大致的想法，无需准确的解决方案。

选做题 1（更新的搜索算法）：本题希望同学们自行探索更新的搜索引擎算法，你们的任务是：

1. 了解并介绍除了准确率（Precision）和召回率（Recall）之外的其它常用搜索引擎功能评价标准；
2. 了解并介绍 Google PageRank 算法的基本原理和流程；
3. 了解并介绍更先进的搜索引擎算法，例如基于 AI 的搜索算法的大致工作原理。

选做题 2（网飞推荐系统）： 网飞（Netflix）是目前最大、最成功的娱乐和媒体公司之一，主要商业模式是订阅流媒体服务（如视频订阅观看）；按收入计算，它是全球第七大互联网公司，该公司拥有 1.5 亿订户，月费产生的年收入为 190 亿美元。网飞的成功离不开推荐算法，他们用推荐算法给用户提供最精确的推荐，从而留住订阅用户；他们甚至基于用户偏好写新的剧本，如《纸牌屋》等，影片受到很多人的追捧。网飞公司在 2006 年设置了 Netflix 百万大奖赛，公开了大约 1 亿个匿名影片评级，数据集仅包含了影片名称、评价星级和评级日期，没有任何文本评价的内容。比赛要求参赛者预测 Netflix 的客户分别喜欢什么影片，要把预测的效率提高 10% 以上。

你们的任务是：

1. 了解网飞使用的推荐算法，选取重要的部分介绍其中的关键流程与技术。事实上很多音乐软件（如网易云等）也有相似的推荐功能，那么如果你是网易云的新用户，网易云会首先给你推荐什么？当你经过一段时间每天深夜都要戴着耳机打开网易云之后，你认为你每天看到的推荐歌单是通过怎样的算法得到的呢？
2. （推荐信安同学完成）了解并介绍网飞推荐系统设计赛中的隐私问题，简单介绍攻击泄露隐私的手段，和一些可行的防御手段。

选做题 3（谷歌广告拍卖）： 事实上，很多时候搜索引擎提供的答案并非完全都是直接回答问题的网页，而是经常包括大量的广告内容，或者搜索引擎的侧边栏会有很多与搜索内容相关的广告出现。那么搜索引擎公司（例如谷歌）是如何决定哪家公司能将广告放在特定的位置的呢？这就是所谓 “关键字搜索拍卖” 的由来。2006 年，来自关键字搜索拍卖的利润占据了谷歌总利润的 98%，令人震撼。

你们的任务是：

1. 了解并介绍谷歌关键字搜索拍卖（广义二价拍卖）的基本模型，如谁赢得拍卖，赢得拍卖的应当支付什么价格；
2. 了解并介绍衡量拍卖的指标，例如真实性、利润最大化、社会福利最大化等，基于此比较广义二价拍卖与其它拍卖形式（如 VCG，一价拍卖等）的差别，大致说明为什么谷歌选择广义二价拍卖这一机制。