CS2045M: 高级数据结构与算法分析 **2025-2026 学年秋冬学期**

Lecture 3: 倒排索引

编写人：吴一航 yhwu_is@zju.edu.cn

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。所以改进为链表存储，这就是倒排索引。注意在倒排索引里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 InvertedFileIndex.zip，其中有相关论文和参考资料等。因为本讲内容简单且宽泛，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。事实上搜索引擎设计中还有非常多具有影响力的算法，如 HITS，PageRank 等，感兴趣的读者可以自行了解。