

Web 信息处理与应用：课后作业 1

网页信息处理 + 网页索引部分

请于 2020 年 10 月 25 日前将作业电子版发送至课程邮箱：ustcweb2019@163.com

## 1 计算题

1.1 请推荐如下查询的处理次序。

(tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)

其中，每个词项对应的倒排记录表的长度分别如下：

词项	倒排记录表长度
----	---------

eyes	213 312
------	---------

kaleidoscope	87 009
--------------	--------

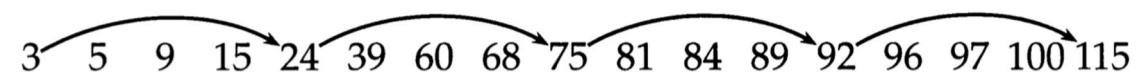
marmalade	107 913
-----------	---------

skies	271 658
-------	---------

tangerine	46 653
-----------	--------

trees	316 812
-------	---------

1.2 考虑利用如下带有跳表指针的倒排记录表



和一个中间结果表（如下所示，不存在跳表指针）进行合并操作。

3 5 89 95 97 99 100 101

采用基于跳表指针的倒排记录表合并算法，请问：

- 1) 跳表指针实际发生跳转的次数是多少？
- 2) 当两个表进行合并时，倒排记录之间的比较次数是多少？
- 3) 如果不使用跳表指针，那么倒排记录之间的比较次数是多少？

1.3 写出倒排记录表 (777, 17743, 294068, 31251336) 的可变字节编码。在可能的情况下对间距而不是文档 ID 编码。写出 8 位块的二进制码。

## 2 问答题

2.1 基于机械分词的常见方法中对于“最大匹配”的依赖，可能导致什么隐患？如何利用 N-最短路径缓解这一隐患？如何选择一个恰当的 N 值？

2.2 如何结合查询词项的分布细节，设计相对合理的跳表指针步长？

2.3 在信息检索系统中，如何同时使用位置索引和停用词表？潜在问题有哪些，如何解决？