全局排序和排序过程

面临的问题：文档质量参差不齐（完整度？类型？可读性？）

全局排序

> 文档质量与相关性结合

$$Score(K,D) = w_1 \cdot Relevance(K,D) + w_2 \cdot Quality(D)$$

PageRank

Main idea :

1. 被引用越多的网页应该是越权威的网页
2. 被权威网页引用的网页很可能也是权威网页

> 基本流程：基于在网页间的游走
>
> 随机选取一个页面作为起始页面，随机进入该页面链接的一个页面，如此操作N次，最后停留在网页W_i的概率即为W_i的PageRank,如果遇到死角，则跳到一个随机选择的网页

> Markov Chain : $S_n = S_0 \cdot T^n$, 最终可到达状态S使得 $S = S \cdot T$
>
> - S为T的特征向量(可证明S是T的最大特征值所对应的特征向量)
>
> 实际中基于 $S_n = S_0 \cdot T^n$ 进行若干次计算，到达所需精度

排序过程

> 首先返回给用户排序最高的K个结果,当用户需要更多结果是，在增大K值

提前淘汰大部分文档

1. 找到 A个最可能包含 Top-K的文档(要求A中的文档包含一定数量的关键词),Then ,Top-K
2. Champion List $[*]$

在每个关键词的posting list 中选出r个得分最高（比如tf值）的文档，使用champion list 完成查询

3. 使用全局排序决定每个posting list的文档顺序。查询时，先使用postinglist的前端

按Impact顺序访问索引

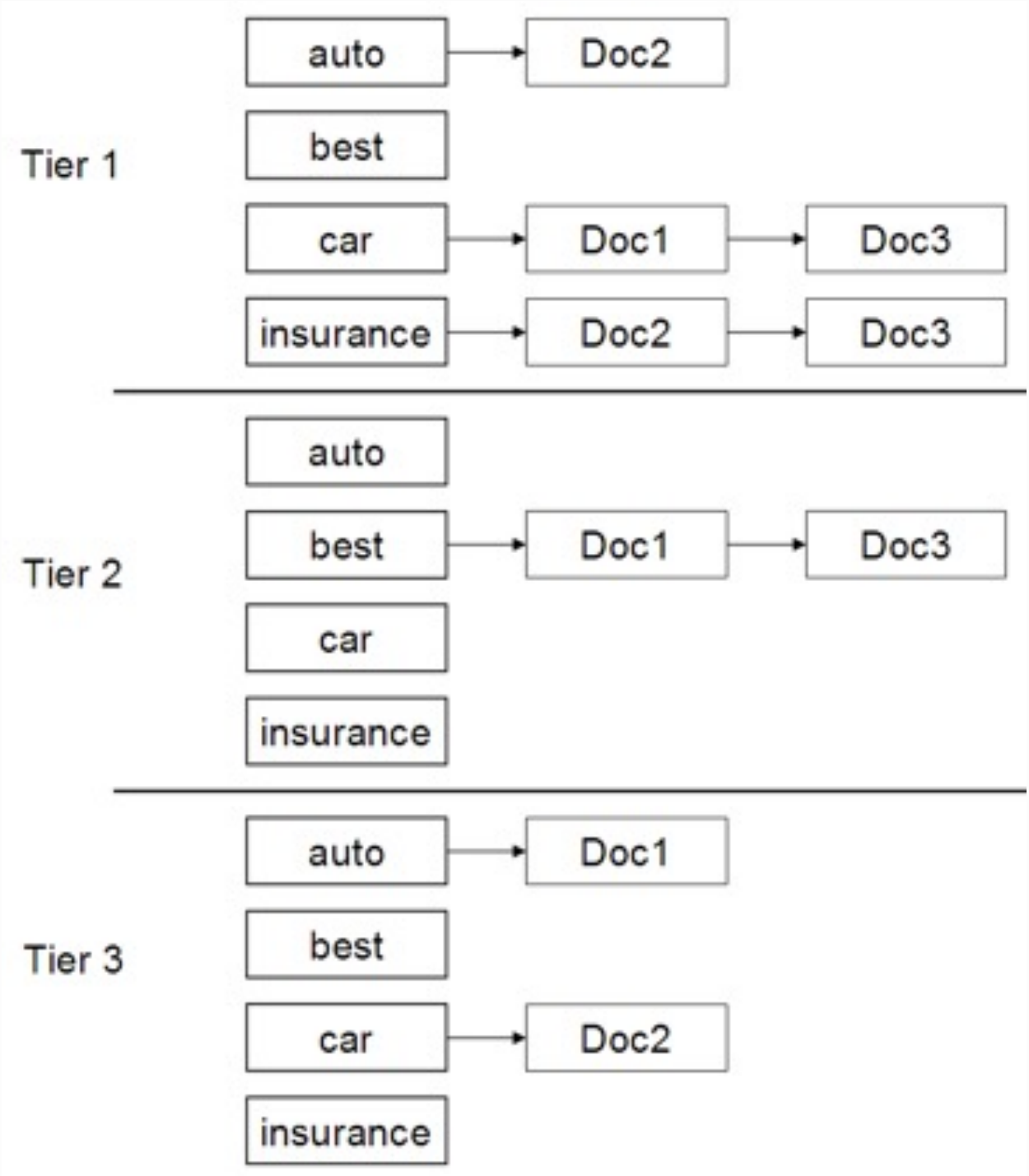
基本思路： 对posting list 按照tf排序 \ 按照idf的顺序逐个访问posting list \ 适时停止访问

聚类

在N个文档中选择$N^{1/2}$个文档作为类中心 \ 按照相关性，将文档聚为$N^{1/2}$个类 \ 用户提交一个查询q \ 找到离q最近的类中心o \ 在o所在的类中找到Top-K.

分层索引

基本思想：先将倒排索引分层次



Note

LSA (Lanten Semntic Analysis)/ LSI Model

每个人独特的词汇使用模式