SimSearcher设计文档

1.建立索引CreateIndex

把所有输入的字符串分割成Q-gram子串集，对于每个子串，通过数据结构unordered\_map<string, vector<unsigned>> 暂存，key为子串，vector是包涵子串的单词id,对上述结构以vector的size进行排序，最终用unordered\_map<string, unsigned>保存子串对应的vector序列id, vector<vector<unsigned>> sortGramList表示排序后的倒排索引。

2.求出候选集

采用了mergeSkip方法。

SearchED:依照公式T= |Q|-qGram+1-2\*thre求出T,再由公式L=T/(ulogM+1)求出长索引链的数目。对于T>0的情况，之后执行MakeGram,Mergeskip,MergeOpt函数，否则直接遍历所有单词进行比较。

SearchJaccard: 依照公式T= max(|Q|\*thre,(|Q|+|S|min)\*thre/(1+thre))求出T,再由公式L=T/(ulogM+1)求出长索引链的数目。之后执行MakeGram,Mergeskip,MergeOpt函数。

MakeGram:找出所有包涵query单词的qGram集的倒排索引链，把他们存入possibleList中。

Mergeskip:在短索引链中使用Mergeskip算法找出可能符合要求得的单词的ID,存入ShortList中。其中的heap结构使用C++库中的优先队列实现。

MergeOpt:对于每个shortList中的每个元素，在所有长索引表中进行二分查找，把符合条件的存入longList中。

3.筛选候选结果

对候选集中的所有字符串和查询字符串进行对比。

getED使用动态规划的方法求出两个单词的ED;

getJaccard把两个字符串都通过空格分割，把子串存入set中消除相同的串，再通过set进行比较。