SimJoin 实验报告

计64 钮泽平 2015010467

Edit Distance

根据论文 Pass-Join: A Partition-based Method for Similarity Joins 进行实现，效果非常好

算法描述

- 这个算法基于鸽巢原理，将一个串分割成$\tau+1$段，那么如果另一个串和它距离小于τ，则这些segements中一定至少有一个是它的子串。
- 将所有串分成$\tau+1$段，建立倒排索引
- 两个串串长相差一定不超过τ
- 对于$[l-\tau, l+\tau]$范围内的串，生成第一个串可能的substring，使用Multi-match-aware-Substring算法
- Multi-match-aware-Substring的主要思想是，对于一个确定位置的segment，将串分成`左，segment，右`三个部分，因为基于中间是匹配的假设，用长度和编辑距离的关系，限制与这个segment匹配的串的起始位置，生成子串集合
- 根据子串集合选取倒排列表，里面的元素就是candidates.

主要数据结构

- L_l^i，表示长度为l的串，第i个seg的倒排索引，使用unordered_map组织，自然地组织为二维数组

Jaccard Similarity

根据论文 Can we Beat the Prefix Filtering? An Adaptive Framework for Similarity Join and Search 进行实现

算法描述

- 框架是基于对每个串查找候选集合，再进行验证
- 算法主要特点是使用统计的方法高效地计算对于一个串，l-prefix最合适的l值是多少，达到"自适应"的效果。
- 写报告的时候`cost-estimation`部分还没有实现完，但是通过手动调节l-prefix的l，已经得到了比较好的结果。

主要数据结构：

- 增量索引，$\Delta I_l (1 \le l \le t)$，表示当后缀长度为$t-l$时比后缀长度为$t-(l-1)$时，前缀增加的term的倒排索引
- 和论文比，比较麻烦的是要将$Jaccard$转化为$Overlap$。初始化时，使用整个数据集中最大可能的t，t_{max}进行索引构建，对于$Overlap = t$的情况，ΔI_1是建立好的索引的一个前缀的增量索引的并集。
- 索引使用`unordered_map`.
- 在通过计数生成CandidateSet的时候，加入jaccard的集合大小过滤（还没实现position filter）