近似链接实验报告

2015011313 徐鉴劲 计54

最终提交ID: 7449

实验目标

近似链接是要找出两个数据集之间所有满足相似性条件的数据对。给定相似函数sim(r, s)和阈值T，数据库R和S分别是字符串的集合，近似链接就是要找出所有满足条件的`{<r, s, sim(r, s)>, sim(r, s) > T}`。

本实验中，进行了Jaccard相似函数和编辑距离相似函数的近似链接。

算法流程

朴素方法-移植近似查询

首先我考虑移植近似查询中的成果。即使对于每一个字符串r属于数据库R，使用近似查询来不断地找出相似对，加入到最终的结果中。所要做出的改动主要是在数据格式方面，由原来的`pair<unsigned, int>`这种变成了`pair<JaccardJoinResult>`这种，更改对应的接口定义和数据增删代码。

1. 倒排列表建立

Jaccard 倒排列表

Jaccard是以单词为级别进行统计，而且重复不算，使用`map<string, vector<int>>`类型。

实现方法：按照空格分割字符，按照单词进行计数。

Q-Gram 倒排列表

每一个q-gram(q长度的字符串)映射到一个链表，链表中升序排列着q-gram所在字符串的序号。由于可能会使用二分查找，所以使用`map<string, vector<int>>`类型。

实现方法：顺序扫描数据集中的字符串，对于每个字符串，依次插入其序号至对应的q-gram处。

2. ED的过滤算法

使用一种改进版的DivideSkip算法，在短序列中使用扫描法统计，在长序列中使用二分法统计。

算法原理

1. 长度为qlen的询问字符串则总共有 qlen - q + 1 个 q-gram。
2. 每一次编辑最多可能改变q个q-gram，所以在 编辑距离 = threshold 的情况下，最多能改变 threshold * q 个q-gram，即改变了多于threshold * q 个q-gram的子串时，编辑距离也一定大于threshold。
3. 计算阈值T = qlen - q + 1 - threshold * q，当重合的q-gram个数小于T时，即改变多了，应该滤除；重合数大于等于T时，满足条件。
4. 考虑分开长短进行，在短序列中统计出重合了t_1个q-gram，在长序列中重合数量为t_2，长序列中没有扫描到的部分长度为l。如果满足$t_1 + t_2 \geq T$时，即满足条件。
   $t_1 + t_2 + l < T$时，即剩下所有都重合也不满足条件，失败退出。

实现流程

1. 首先建立query的q-gram，然后按照q-gram的频率（即其对应倒开列表的长度）升序排列。排在后边的就是短的序列。
2. 在qlen - q + 1 - L 个短序列中进行扫描，对于每一个q-gram，找到其倒开列表，对应的字符串计数加一，如果等于T - L，那么说明在L个长序列中（如果全部都有重合）可能产生符合条件的字符串，退出。
3. 对于每一个上一步中滤出的备选字符串，设当前总q-gram重合次数为k（k的初值就是短序列中q-gram的重合次数），对L个长倒开列表中的每一个进行二分查找，每当找到重合的q-gram，k自加一，同时判断是否成功；每当二分查找失败，判断失败条件。成功条件是$k \geq T$，失败条件是$k + L - i - 1 < T$。
4. 对剩下的备选字符串进行验证。

3. ED 的验证算法

直接进行动态规划，同时有一点小优化：动态规划中统计当前行的最小值，因为下一行的最小值必然大于当前行的最小值（结果也必然大于这个值），可以用这个最小值与编辑距离阈值进行比较，进行早退出。

4. Jaccard 的统计算法

对query按照空格进行分词，每个分词查询倒开列表，统计每一个字符串的相交个数。然后遍历整个相交个数列表，通过容斥原理求出Jaccard。

实验效果

实验结果是正确的，在exp2中运行了60s，在exp2-final中运行了360s。我的近似查询实现方法未经过有效优化，在exp1中运行10s，在exp1-final中运行300s。