数据库专题训练

数据库新型检索技术

作业二 近似连接

助教 李浩达 lihaoda9@163.com

清華大學

实验框架

- 请参考框架代码，实现SimJoiner类的方法：
  - joinJaccard()函数
  - joinED()函数
- 请不要修改这两个方法的声明，可以根据需要自行添加其他方法。

JoinResult类

```
template <typename _IDType, typename _SimType>
struct JoinResult
{
        _IDType id1;   // file1中的记录id
        _IDType id2;   // file2中的记录id
        _SimType s;    // 相似度/编辑距离
};
typedef JoinResult<unsigned, double> JaccardJoinResult;
typedef JoinResult<unsigned, unsigned> EDJoinResult;
```

输入文件格式同实验1，每行一个记录（字符串），记录号为从0开始的行号。

joinJaccard函数

- 函数声明：int joinJaccard(const char *filename1, const char *filename2, double threshold, vector<JaccardJoinResult> &result);
  - filename1， filename2：输入文件名
  - threshold：Jaccard阈值
  - vector<JaccardJoinResult> &result，返回的结果，需按照id1、id2从小到大排序，且无重复结果
  - 返回值同实验1createIndex

joinED函数

- 函数声明：int joinED(const char *filename1, const char *filename2, unsigned threshold, vector<EDJoinResult> &result);
  - filename1， filename2：输入文件名
  - threshold：ED阈值
  - vector<EDJoinResult> &result，返回的结果，需按照id1、id2从小到大排序，且无重复结果
  - 返回值同实验1createIndex
  - 可以自行选定q值

实验要求

- 实验平台：Ubuntu, g++ 4.8
- 评测标准:
  - ✓ 正确性：
    - ✓ 返回的结果均满足查询要求;满足查询要求的结果全部被返回
  - ✓ 时间：跑的越快越好，最终以速度排名为依据给分
  - ✓ 空间：要求能够跑动最终评测数据集（一般不需考虑）
- 提交材料：
  - OJ上的submission id
  - 简要的文档，描述算法设计、关键优化等
  - 网络学堂提交
- 在线评测截止时间:以评测平台为准
- 材料提交截止时间:以网络学堂为准

评测说明

- 最终编译会采用给定的makefile，大家可以自行测试自己的代码是否能通过编译
- 可以使用c++11中的特性来简化代码，可以使用stl标准库
- 请不要使用多线程等手段来加速程序
- 最终提交文件中请不要包含main函数，以避免链接失败。最终评测流程为：
  - 将提交的代码压缩包解压缩
  - 将评测用的main.cpp，makefile复制到同一目录
  - 编译，运行得到的程序
- 请不要尝试攻击实验室服务器☺

Thanks, Questions?

清華大學