课程项目2

PostgreSQL上的Similarity Join实现

相似性查询

addressphone	
Address	**Phone**
11 North Michigan Avenue Chicago	(312) 521-7275

restaurantaddress	
Name	**Address**
Par; Grill restaurant	11 N Michigan Avenue Chicago

- addressphone.Address ≠ restaurantaddress.Address
- Similarity(addressphone.Address, restaurantaddress.Address) ≥ 0.7

两种相似性度量方式

- Levenshtein Distance
- Jaccard Index

Levenshtein Distance

- 两个字串之间，由一个转成另一个所需的最少编辑操作次数
- 允许的编辑操作：
  - 将一个字符替换成另一个字符
  - 插入一个字符
  - 删除一个字符
- 例：kitten转成sitting
  1. k -> s = sitten
  2. e -> i = sittin
  3. + g = sitting
  - $LD(kitten, sitting) = 3$

Jaccard Index

- 两个集合之间，交集的大小除以并集的大小
    - $J(A,B)=\frac{|A\cap B|}{|A\cup B|}$
- 本次实验以Bigram为基本单元
    - Bigram：字符串中连续的两个字符组成的基本单元
    - $Bigram(Apple)=\{\$A,Ap,pp,pl,le,e\$\}$
        - 开始字符前和结束字符后分别添加$符号
- 例：$J(Apple,Apply)=\frac{|\{\$A,Ap,pp,pl\}|}{|\{\$A,Ap,pp,pl,le,e\$,ly,y\$\}|}=\frac{4}{8}=0.5$
    - $Bigram(Apply)=\{\$A,Ap,pp,pl,ly,y\$\}$

测试查询

- Levenshtein Distance

```
select count(*)
from restaurantaddress ra, addressphone ap
where levenshtein_distance(ra.address, ap.address) < 4;
```

- Jaccard Index

```
select count(*)
from restaurantphone rp, addressphone ap
where jaccard_index(rp.phone, ap.phone) > 0.6;
```

结果验证（依次增大）

```
select count(*) from restaurantphone rp, addressphone ap where
levenshtein_distance(rp.phone, ap.phone) < 4;
select count(*) from restaurantaddress ra, restaurantphone rp where
levenshtein_distance(ra.name, rp.name) < 3;
select count(*) from restaurantaddress ra, addressphone ap where
levenshtein_distance(ra.address, ap.address) < 4;
select count(*) from restaurantphone rp, addressphone ap where
jaccard_index(rp.phone, ap.phone) > 0.6;
select count(*) from restaurantaddress ra, restaurantphone rp where
jaccard_index(ra.name, rp.name) > 0.65;
select count(*) from restaurantaddress ra, addressphone ap where
jaccard_index(ra.address, ap.address) > 0.8;
```

实验流程

1. 准备开发环境，推荐使用Linux Ubuntu 发行版
   - https://www.ubuntu.com/download/desktop
   - Windows下可使用VirtualBox配置虚拟环境
     - https://www.virtualbox.org/wiki/Downloads
2. 下载PostgreSQL源码
   - https://ftp.postgresql.org/pub/source/v10.4/postgresql-10.4.tar.gz
3. 阅读PostgreSQL文档，进行编译、安装
   - https://www.postgresql.org/docs/10/static/installation.html
4. 阅读文档、源码，修改源码，并调试
   - https://www.postgresql.org/docs/10/static/index.html

实验流程

5. 导入数据，运行实验测试样例，记录返回结果、运行时间等
6. 撰写实验报告
   - 系统与源码理解
   - 设计思路与实现方案
   - 关键代码说明
   - 实验与结果

   ○ 性能优化

验收方式

- 2人一组
  - 分组截止6月11日，邮件发送至qtwang16@fudan.edu.cn
- 7月7日23:55前通过elearning提交
- 压缩包命名方式：学号_姓名
- 内容
  - 实验报告
  - 源代码
- 评分依据
  - 实验报告
  - 代码质量
  - 实现与优化情况

参考文献

- Levenshtein Distance
  - https://en.wikipedia.org/wiki/Levenshtein_distance
- Jaccard Index
  - https://en.wikipedia.org/wiki/Jaccard_index
- Bigram
  - https://en.wikipedia.org/wiki/Bigram
- PostgreSQL Documents
  - https://www.postgresql.org/files/documentation/pdf/10/postgresql-10-A4.pdf