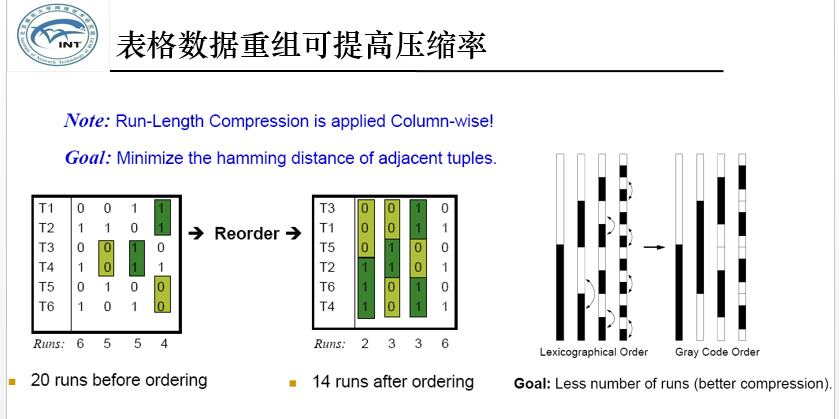
实验描述

我大概要做的事情是数据库表排序，之后便于压缩处理：

1. 背景

数据库表是一个0-1矩阵，就是只有0和1的矩阵

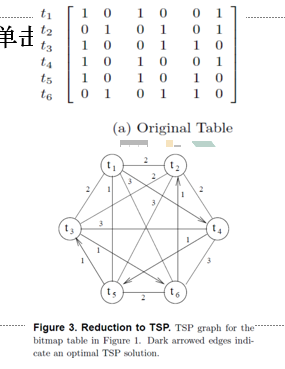


* 把0-1矩阵表格进行行排序，使得尽量多的相同值可以排列的靠近一些，方便压缩
* 比如左侧图 行排序为 T1 T2 T3 T4 T5 T6， 使用排序算法，排成右侧的行顺序 T3 T1 T5 T2 T6 T4；
* 压缩时是按照列来压缩，每个相同值的段，称作一个Run； Run数量越少，压缩效果越好
* 所以左侧第一列到第四列Run数分别是 6 5 5 4， 排序后，第一列到第四列的Run数分别是2 3 3 6， Run数减少了； 更直观的图可以看右侧的黑白图，分别是排序前排序后

怎样找到合适的排序顺序？

这个问题可以建模成hamming（海明）距离下TSP（旅行商问题），就是有n个城市，选择一条路径（排序），使得每个城市访问且只访问一次，路径总和最小。

Hamming（海明）距离就是两个码字的对应比特取值不同的比特数称为这两个码字的海明距离 如：10101和00110从第一位开始依次有第一位、第四、第五位不同，则海明距离为3



把数据库表的行作为图的顶点，顶点间海明距离作为边上的权值，找到最小路径（或者较小路径）。

以上就是背景知识，但不用理解的特别仔细

1. 我的实验目的是提出一种新的排序算法，带入到程序里，压缩效果跟现有的一些算法比较，如果占优，那就可以了。
2. 占优指标可以是Run数，运行速度，或者压缩率（压缩前压缩后的文件大小比值）
3. 这一套ReorderingRows 文件夹下Java程序主要代码介绍：

* sorting.java；

这个是主程序，它会调用各种排序算法来运行；通过命令行参数来看调用哪种算法；

* zipfian.java；

生成待排序文件，是zipfian分布的，zipfian是一种数学分布

* CountRunsBinary.java

可以计算文件的Run数

* Inmemory文件夹里面大多是算法，下面这些都是排序算法：

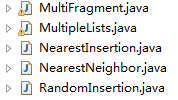
Nearestinsertion， nearestneighbor，TWO-Opt,THREE-OPT都是旅行商问题的算法













1. 我想做的是：

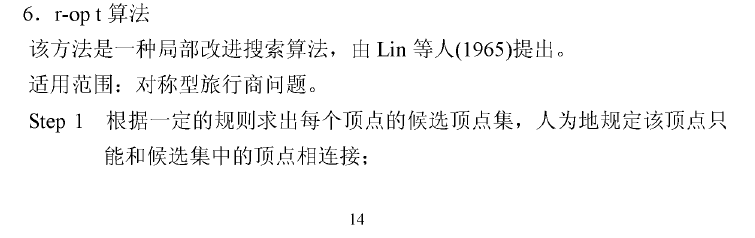
* 把文件包中TwoOpt.java排序算法调入到sorting.java里试试压缩效果，指标是Run数多少，或者压缩率多少
  + 现在的sorting.java程序里面没有带入TwoOpt算法，需要写点代码带入进去
  + 这一套java程序目前是不带压缩算法的，只是排序而已，所以还需要压缩处理；
    - 我找到了一个压缩算法jar包，但是不会用，叫EWAH算法；如果测压缩率的话，需要把压缩算法包带入进去使用；
* 把网上一套ThreeOpt算法C程序写成java， 或者按照这套java程序的风格自己写一个也行，带入到sorting.java里试试效果
* 如果ThreeOpt效果比较好的话，在ThreeOpt算法上做点小小修改，就是创新了
  + ThreeOpt可以跟TwoOPT、LEX等已有的几种算法来比较； Run数、运行时间或者压缩率上有点占优就行
* 如果在zipfian数据集上效果不错，可以在实际的数据集上测测，实际数据集的数据处理我还没弄好

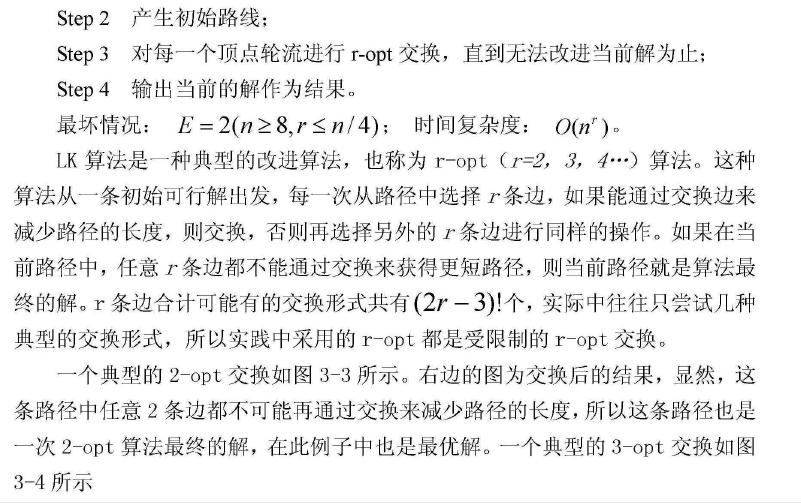
2-opt，3-opt算法原理的简单介绍（仅供理解参考，代码都有）：

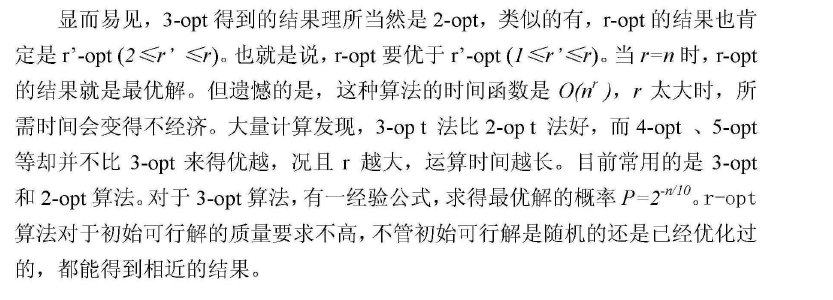
简单说：

-2opt算法就是去掉原来路径中两条边，换上另外两条，如果路径总长度减小，就交换；直到没有可交换的

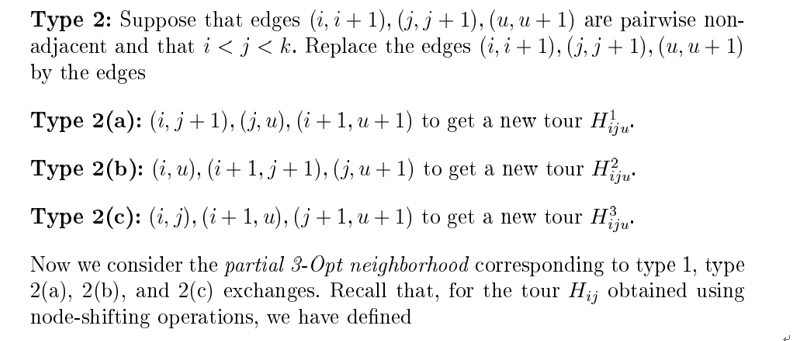
-3opt算法就是去掉原来路径中三条边，换上另外三条，如果路径总长度减小，就交换；直到没有可交换的； 但是换哪三条边是可以有各种变化的。







3-opt的算法的一种：



1. 我自己目前的实验过程，供参考：

实验过程：

1. 首先使用zipfian.java 生成三个文件；三个输入参数分别是 指定生成的文件名，行数和列数

运行时使用的参数：分别制定生成的文件，行数和列数

Zipfiantable1 8192 4

Zipfiantable2 131072 4

Zipfiantable3 524288 4

1. 使用多种算法进行排序，sort.java文件，四个输入参数分别是是否要排序，排序算法，输入文件，输入文件

-reorder -lexicographic zipfiantable1 zipfianoutputlex1

-reorder -reflectedgray zipfiantable1 zipfianoutputgray1

-reorder -freq zipfiantable1 zipfianoutputfreq1

-reorder -lexicographic zipfiantable2 zipfianoutputlex2

-reorder -reflectedgray zipfiantable2 zipfianoutputgray2

-reorder -freq zipfiantable2 zipfianoutputfreq2

-reorder -lexicographic zipfiantable3 zipfianoutputlex3

-reorder -reflectedgray zipfiantable3 zipfianoutputgray3

-reorder -freq zipfiantable3 zipfianoutputfreq3

得到了三组文件可以比较三种算法在不同size文件的排序情况

1. 比较排序之前和之后的run数 CountRunsBinary.java ? 数字为啥出不来？二个输入参数分别是，计算Run数的依据，待计算的文件；但是我发现排序数字出不来？是不是因为文件太大了，我以前用小文件是可以输出排序数字的。

-m3 zipfiantable1

-m3 zipfianoutputlex1

-m3 zipfianoutputgray1

-m3 zipfianoutputfreq1

-m3 zipfiantable2

-m3 zipfianoutputlex2

-m3 zipfianoutputgray2

-m3 zipfianoutputfreq2

-m3 zipfiantable3

-m3 zipfianoutputlex3

-m3 zipfianoutputgray3

-m3 zipfianoutputfreq3

1. 可以比较2-opt算法排序和3-opt算法排序的压缩率

EWAH包找到了，但是不会带入使用。