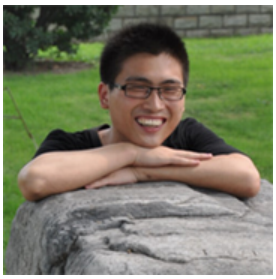
zhanlijun



个人经历
2015 至今 阿里巴巴
2013-2015 美团
2010-2013 中科院（硕士）
2006-2010 浙大（本科）
阿里巴巴RDC长期招聘Java研发工程师，有意者站内联系！
昵称：zhanlijun
园龄：4年10个月
粉丝：664
关注：5
+加关注

GeoHash核心原理解析

http://www.cnblogs.com/LBSer/p/3310455.html

引子

机机是个好动又好学的孩子，平日里就喜欢拿着手机地图点点按按来查询一些好玩的东西。某一天机机到北海公园游玩，肚肚饿了，于是乎打开手机地图，搜索北海公园附近的餐馆，并选了其中一家用餐。

饭饱之后机机开始反思了，地图后台如何根据自己所在位置查询来查询附近餐馆的呢？苦思冥想了半天，机机想出了个方法：计算所在位置P与北京所有餐馆的距离，然后返回距离<=1000米的餐馆。小得意了一会儿，机机发现北京的餐馆何其多啊，这样计算不得了，于是想了，既然知道经纬度了，那它应该知道自己在西城区，那应该计算所在位置P与西城区所有餐馆的距离啊，机机运用了递归的思想，想到了西城区也很多餐馆啊，应该计算所在位置P与所在街道所有餐馆的距离，这样计算量又小了，效率也提升了。

机机的计算思想很朴素，就是通过过滤的方法来减小参与计算的餐馆数目，从某种角度上讲，机机在使用索引技术。

一提到索引，大家脑子里马上浮现出B树索引，因为大量的数据库（如MySQL、oracle、PostgreSQL等）都在使用B树。B树索引本质上是对索引字段进行排序，然后通过类似二分查找的方法进行快速查找，即它要求索引的字段是可排序的，一般而言，可排序的是一维字段，比如时间、年龄、薪水等等。但是对于空间上的一个点（二维，包括经度和纬度），如何排序呢？又如何索引呢？解决的方法很多，下文介绍一种方法来解决这一问题。

思想：如果能通过某种方法将二维的点数据转换成一维的数据，那样不就可以继续使用B树索引了嘛。那这种方法真的存在嘛，答案是肯定的。目前很火的GeoHash算法就是运用了上述思想，下面我们就开始GeoHash之旅吧。

一、感性认识GeoHash

首先来点感性认识，http://openlocation.org/geohash/geohash-js/ 提供了在地图上显示geohash编码的功能。

1）GeoHash将二维的经纬度转换成字符串，比如下图展示了北京9个区域的GeoHash字符串，分别是WX4ER，WX4G2、WX4G3等等，每一个字符串代表了某一矩形区域。也就是说，这个矩形区域内所有的点（经纬度坐标）都共享相同的GeoHash字符串，这样既可以保护隐私（只表示大概区域位置而不是具体的点），又比较容易做缓存，比如左上角这个区域内的用户不断发送位置信息请求餐馆数据，由于这些用户的GeoHash字符串都是WX4ER，所以可以把WX4ER当作key，把该区域的餐馆信息当作value来进行缓存，而如果不使用GeoHash的话，由于区域内的用户传来的经纬度是各不相同的，很难做缓存。

2）字符串越长，表示的范围越精确。如图所示，5位的编码能表示10平方千米范围的矩形区域，而6位编码能表示更精细的区域（约0.34平方千米）

3）字符串相似的表示距离相近（特殊情况后文阐述），这样可以利用字符串的前缀匹配来查询附近的POI信息。如下两个图所示，一个在城区，一个在郊区，城区的GeoHash字符串之间比较相似，郊区的字符串之间也比较相似，而城区和郊区的GeoHash字符串相似程度要低些。

城区

郊区

通过上面的介绍我们知道了GeoHash就是一种将经纬度转换成字符串的方法，并且使得在大部分情况下，字符串前缀匹配越多的距离越近，回到我们的案例，根据所在位置查询来查询附近餐馆时，只需要将所在位置经纬度转换成GeoHash字符串，并与各个餐馆的GeoHash字符串进行前缀匹配，匹配越多的距离越近。

二、GeoHash算法的步骤

下面以北海公园为例介绍GeoHash算法的计算步骤

2. 你应该知道的RPC原理(30598)

3. 如何用消息系统避免分布式事务？(23677)

4. mysql死锁问题分析(22275)

5. 位图索引:原理（BitMap index）(21132)

评论排行榜

1. 地图匹配实践(82)

2. 如何用消息系统避免分布式事务？(42)

3. 你应该知道的RPC原理(23)

4. GeoHash核心原理解析(22)

5. 地理围栏算法解析（Geo-fencing）(20)

2.1. 根据经纬度计算GeoHash二进制编码

地球纬度区间是[-90,90]， 北海公园的纬度是39.928167，可以通过下面算法对纬度39.928167进行逼近编码:

1）区间[-90,90]进行二分为[-90,0),[0,90]，称为左右区间，可以确定39.928167属于右区间[0,90]，给标记为1；

2）接着将区间[0,90]进行二分为 [0,45),[45,90]，可以确定39.928167属于左区间 [0,45)，给标记为0；

3）递归上述过程39.928167总是属于某个区间[a,b]。随着每次迭代区间[a,b]总在缩小，并越来越逼近39.928167；

4）如果给定的纬度x（39.928167）属于左区间，则记录0，如果属于右区间则记录1，这样随着算法的进行会产生一个序列1011100，序列的长度跟给定的区间划分次数有关。

根据纬度算编码

bit	min	mid	max
1	-90.000	0.000	90.000
0	0.000	45.000	90.000
1	0.000	22.500	45.000
1	22.500	33.750	45.000
1	33.7500	39.375	45.000
0	39.375	42.188	45.000
0	39.375	40.7815	42.188
0	39.375	40.07825	40.7815
1	39.375	39.726625	40.07825
1	39.726625	39.9024375	40.07825

同理，地球经度区间是[-180,180]，可以对经度116.389550进行编码。

根据经度算编码

bit	min	mid	max
1	-180	0.000	180
1	0.000	90	180
0	90	135	180

1	90	112.5	135
0	112.5	123.75	135
0	112.5	118.125	123.75
1	112.5	115.3125	118.125
0	115.3125	116.71875	118.125
1	115.3125	116.015625	116.71875
1	116.015625	116.3671875	116.71875

2.2. 组码

通过上述计算，纬度产生的编码为10111 00011，经度产生的编码为11010 01011。偶数位放经度，奇数位放纬度，把2串编码组合生成新串：11100 11101 00100 01111。

最后使用用0-9、b-z（去掉a, i, l, o）这32个字母进行base32编码，首先将11100 11101 00100 01111转成十进制，对应着28、29、4、15，十进制对应的编码就是wx4g。同理，将编码转换成经纬度的解码算法与之相反，具体不再赘述。

Decimal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Base 32	0	1	2	3	4	5	6	7	8	9	b	c	d	e	f	g
Decimal	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Base 32	h	j	k	m	n	p	q	r	s	t	u	v	w	x	y	z

三、GeoHash Base32编码长度与精度

下表摘自维基百科：http://en.wikipedia.org/wiki/Geohash

可以看出，当geohash base32编码长度为8时，精度在19米左右，而当编码长度为9时，精度在2米左右，编码长度需要根据数据情况进行选择。

geohash length	lat bits	lng bits	lat error	lng error	km error
1	2	3	±23	±23	±2500
2	5	5	± 2.8	± 5.6	±630
3	7	8	± 0.70	± 0.7	±78
4	10	10	± 0.087	± 0.18	±20
5	12	13	± 0.022	± 0.022	±2.4
6	15	15	± 0.0027	± 0.0055	±0.61
7	17	18	±0.00068	±0.00068	±0.076
8	20	20	±0.000085	±0.00017	±0.019

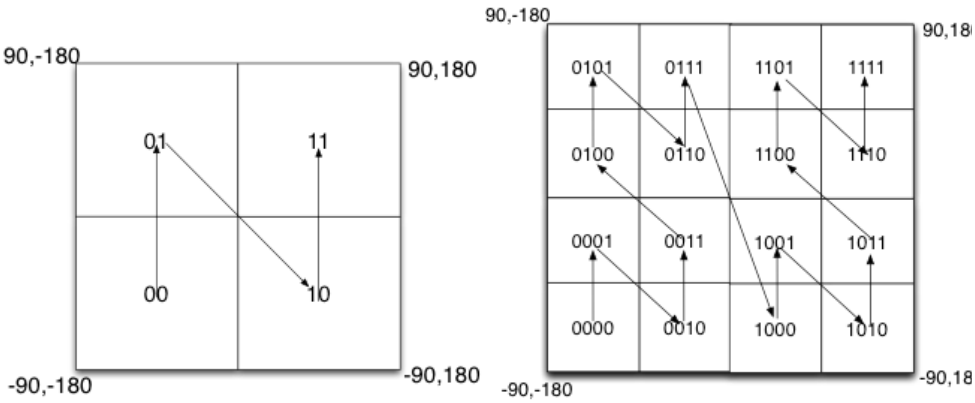
三、GeoHash算法

上文讲了GeoHash的计算步骤，仅仅说明是什么而没有说明为什么？为什么分别给经度和维度编码？为什么需要将经纬度两串编码交叉组合成一串编码？本节试图回答这一问题。

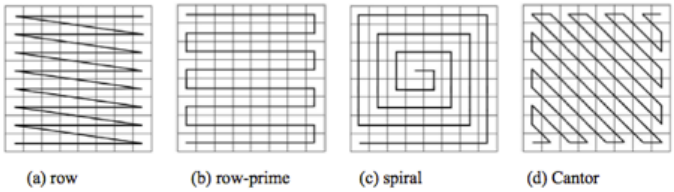
如图所示，我们将二进制编码的结果填写到空间中，当将空间划分为四块时候，编码的顺序分别是左下角00，左上角01，右下脚10，右上角11，也就是类似于Z的曲线，当我们递归的将各个块分解成更小的子块时，编码的顺序是自相似的（分形），每一个子快也形成Z曲线，这种类型的曲线被称为Peano空间填充曲线。

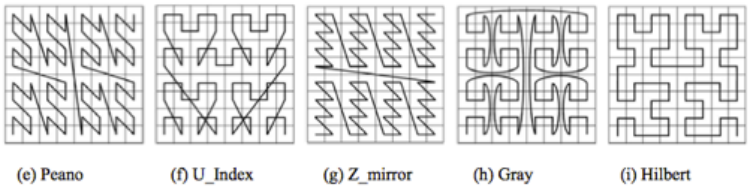
这种类型的空间填充曲线的优点是将二维空间转换成一维曲线（事实上是分形维），对大部分而言，编码相似的距离也相近， 但Peano空间填充曲线最大的缺点就是突变性，有些编码相邻但距离却

相差很远，比如0111与1000，编码是相邻的，但距离相差很大。



除Peano空间填充曲线外，还有很多空间填充曲线，如图所示，其中效果公认较好是Hilbert空间填充曲线，相较于Peano曲线而言，Hilbert曲线没有较大的突变。为什么GeoHash不选择Hilbert空间填充曲线呢？可能是Peano曲线思路以及计算上比较简单吧，事实上，Peano曲线就是一种四叉树线性编码方式。





四、使用注意点

1）由于GeoHash是将区域划分为一个个规则矩形，并对每个矩形进行编码，这样在查询附近POI信息时会导致以下问题，比如红色的点是我们的位置，绿色的两个点分别是附近的两个餐馆，但是在查询的时候会发现距离较远餐馆的GeoHash编码与我们一样（因为在同一个GeoHash区域块上），而较近餐馆的GeoHash编码与我们不一致。这个问题往往产生在边界处。

解决的思路很简单，我们查询时，除了使用定位点的GeoHash编码进行匹配外，还使用周围8个区域的GeoHash编码，这样可以避免这个问题。

2）我们已经知道现有的GeoHash算法使用的是Peano空间填充曲线，这种曲线会产生突变，造成了编码虽然相似但距离可能相差很大的问题，因此在查询附近餐馆时候，首先筛选GeoHash编码相似的POI点，然后进行实际距离计算。

geohash只是空间索引的一种方式，特别适合点数据，而对线、面数据采用R树索引更有优势（可参考：**深入浅出空间索引：为什么需要空间索引**）。

参考文献：

http://en.wikipedia.org/wiki/Geohash

http://openlocation.org/geohash/geohash-js/

Cantor空間填充曲線之演算法探討.pdf

转载请标明源地址：http://www.cnblogs.com/LBSer

分类: 空间索引原理

标签: lbs 定位, 索引

zhanlijun
关注 - 5
粉丝 - 664
+加关注

12 0

« 上一篇 : Geohash距离估算
» 下一篇 : 位图索引:原理（BitMap index）

posted @ 2013-09-09 19:03 zhanlijun 阅读(43982) 评论(22) 编辑 收藏

评论列表

#1楼 2013-09-09 20:02 lreis

好文！收藏了

支持(0) 反对(0)

#2楼 2013-11-18 10:32 tanshaohua

你好，请问下，如果要选择2公里，3公里等，应该怎么处理，geohash没有指名具体每公里的情况

支持(0) 反对(0)

#3楼[楼主] 2013-11-18 19:16 zhanlijun

@ tanshaohua
您好，您可以参考下这个，会有大致的估算http://www.cnblogs.com/LBSer/p/3298057.html

支持(0) 反对(0)

#4楼 2013-11-19 08:57 tanshaohua

嗯，谢谢，这篇文章之前也看过，我就是想用户可以根据选择距离来查询多少公里内的东西，不知道除了geohash还有什么比较适用

支持(0) 反对(0)

#5楼[楼主] 2013-11-19 11:58 zhanlijun

@ tanshaohua
其实你需要的是空间范围查询，为了更高效地进行空间查询需要使用空间索引，geohash只是其中一种，还有四叉树、R树等索引。四叉树和R树这种树类型的空间索引支持你查询具体距离以内的东西，目前postgreSQL、mysql都有相应的空间扩展支持

支持(0) 反对(0)

#6楼 2014-06-23 14:22 Alexia(minmin)

请问下http://segmentfault.com/q/1010000000586274#a-1020000000586281 这个需求能不能用geohash

支持(0) 反对(0)

#7楼[楼主] 2014-06-25 07:47 zhanlijun

@ Alexia(minmin)
这个需求应该不能用geohash，原因是geohash的x,y指的是二维平面上的一个点，而你需求的x,y

指的是区间范围，意义是不一样的。
用hash来解决这种没有规律的范围问题应该很难。

支持(0) 反对(0)

#8楼 2014-07-16 15:31 ginger_jiang

geohash值可以区分精度，位数越多，精度越高，表达的地理位置越精细；如如一位的geohash值把地球划分为32个矩形，8位的geohash值把地球划分为32^8个小矩形

1. 适合根据某个经纬度坐标position计算出geohash值，然后和数据库中精度更高的geohash值做前缀比较

2. 由于有边界突变，在做大小比较时会有如下问题，比如：
A （1, 1) --> s00twy01
B (-1, -1) --> 7zz631zy
C (1, 20) --> s2njxn59
虽然A与B距离更接近，但geohash值相差更大；为了修复边界的问题，常用的方法是找出该position周围的8个点，然后和数据库中的值比较

支持(0) 反对(0)

#9楼[楼主] 2014-07-17 07:42 zhanlijun

@ ginger_jiang
说得很对，从数学上讲geohash本质上是Peano空间填充曲线，将二维空间降维成分形曲线，此分形曲线能保持一定的空间局部性，从而可以用于排序。
Peano曲线突变比较厉害，空间局部性不是特别好，相比之下Hilbert空间填充曲线空间局部性非常不错。

支持(0) 反对(0)

#10楼 2015-03-18 22:12 superpipix

如果我想搜周围8个格子，甚至再外一圈12个格子应该怎么做呢？有什么办法可以算出周边格子的hash值吗？

支持(0) 反对(0)

#11楼[楼主] 2015-03-19 10:09 zhanlijun

@ superpipix
可以的，你是否是想做这样的工作，比如给个范围（可以是矩形或者圆），查出包含在里面的geohash？

支持(0) 反对(0)

#12楼 2015-03-19 10:51 superpipix

@ zhanlijun
恩，对的。支持根据范围查geohash列表，出来的geohash如果是根据距离排序就更好了。

支持(0) 反对(0)

#13楼 2015-03-23 16:49 LIUSANNITY

为了修复边界的问题，怎么找出该position周围的8个点呢？有啥好的方法？

支持(0) 反对(0)

#14楼 2015-07-23 14:15 newjueqi

请问文章上的截图，是在geohash-js弄出来的吗？现在这个网站没法访问了

支持(0) 反对(0)

#15楼[楼主] 2015-08-03 15:41 zhanlijun

@ newjueqi
是的，貌似看不了

支持(0) 反对(0)

#16楼 2015-09-12 08:27 时之沙漠

纬度产生的编码为10111 00011，经度产生的编码为11010 01011。偶数位放经度，奇数位放纬度，把2串编码组合生成新串：11100 11101 00100 01111

根据得到的新串看，应该是奇数位放经度，偶数位放纬度吧？

支持(0) 反对(0)

#17楼 2015-12-24 12:58 OliverMann

你好我问下，给定一个geohash块，查找周围8块甚至周围25块的算法有吗，如果有的话能提供下吗 谢谢

支持(0) 反对(0)

#18楼 2016-07-25 21:13 谁说我不是会员

@ zhanlijun
请教个问题，为什么要去掉a, i, l, o

支持(0) 反对(0)

#19楼 2016-09-18 22:04 asbai

按照维基百科的截图，geohash length 为 5 的时候，km error 为 ±2.4，那对应矩形覆盖的面积应该是 23.04 平方公里呀？

支持(0) 反对(0)

#20楼 2017-11-14 11:12 N3verL4nd

@ 谁说我不是会员
l长得像1
o长得像0
就这么简单

支持(0) 反对(0)

#21楼 2017-12-10 16:44 金亚大王

@ OliverMann @LIUSANNITY
https://github.com/chenjinya/php-geohash 我写了一个供参考，里面有计算附近的8块的方法

支持(0) 反对(0)

#22楼 2018-03-16 11:06 casterQL

lucene里面使用了geohash，但是计算距离的时候貌似还是用经纬度计算距离，那使用geohash还有什么意义呢？

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

最新IT新闻:

· 美团打车：已拿下上海1/3市场份额
· 孙宏斌：投资乐视网肯定是失败了，对财务和团队判断有失误
· 蓝色光标陷劳资纠纷 6年人员成本增逾10倍
· 投资育碧、腾讯财报 这两件事应该一起看
· 7小时通宵大搜查！英国隐私监管机构进驻剑桥分析可查服务器
» 更多新闻...

最新知识库文章:

· 写给自学者的入门指南
· 和程序员谈恋爱
· 学会学习
· 优秀技术人的管理陷阱
· 作为一个程序员，数学对你到底有多重要
» 更多知识库文章...