

zhanlijun

首页

新随笔

联系

订阅

管理

随笔 - 49 文章 - 0 评论 - 370

Visitors

CN

177,205

US

15,556

HK

5,780

JP

3,721

TW

3,008

SG

1,356

CA

816

FR

805

GB

775

AU

695

DE

670

KR

336

Pageviews: 377,090

Flags Collected: 83

FLAG counter



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

最新随笔

1. 一个复杂系统的拆分改造实践

2. mysql死锁问题分析

3. 近期code review几处小问题集锦

4. 你应该知道的RPC原理

5. 如何健壮你的后端服务？

6. 如何用消息系统避免分布式事务？

7. 一个故事讲清楚NIO

8. 地图匹配实践

9. 利用模拟退火提高Kmeans的聚类精度

10. 空间插值文献阅读（Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall）

深入浅出空间索引：为什么需要空间索引

<http://www.cnblogs.com/LBSer/p/3392491.html>

一、问题

先思考个常见的问题：如何根据自己所在位置查询来查询附近50米的POI（point of interest，比如商家、景点等）呢（图1a）？

每个POI都有经纬度信息，我用图1b的SQL语句在mySQL中建立了POI_spatial的表，其中lat和lng两个字段来代表纬度和经度。为后续分析方便起见，我人造了40万个POI数据。




图1a

```
CREATE TABLE poi_spatial(  
  id int(30) not null,  
  name varchar(100),  
  Lat float,  
  Lng float,  
  ...  
);
```

图1b

二、传统的解决思路

方法一：暴力方法

该方法的思路很直接：计算位置与所有POI的距离，并保留距离小于50米的POI。

插句题外话，计算经纬度之间的距离不能像求欧式距离那样平方开根号，因为地球是个不规则的球体（图2a），按最简单的完美球体假设，两点之间的距离函数应该如图2b所示。




图2a

```
double Pi = 3.1425926;  
double Ri = 6371; //地球半径km  
  
double distance=(Ri*acos(sin(Lat1*(Pi/  
180.0))*sin(Lat2*(Pi/180.0)) +  
cos(Lat1*(Pi/180.0))*cos(Lat2*(Pi/  
180.0))*cos((Lon1 - Lon2)*(Pi/180.0))));
```

图2b

该方法的复杂度为：40万*距离函数。我们将球体距离函数写为mysql存储过程distance，之后我们执行查询操作（图3），发现花费了4.66秒。

该方法耗时的原因显而易见，执行了40万次复杂的距离计算函数。

```
mysql> select id,name from poi_spatial where  
distance(lng,lat,116.3290,39.9688)<0.05;
```

id	name
32	[REDACTED]
37	[REDACTED]

2 rows in set (4.66 sec)

图3

方法二：矩形过滤方法

该方法分为两部：

- a) 先用矩形框过滤（图4a），判断一个点在矩形框内很简单，只要进行两次判断（LtMin<lat<LtMax; LnMin<lng<LnMax），落在矩形框内的POI个数为n（n<<40万）；
- b) 用球面距离公式计算位置与矩形框内n个POI的距离（图4b），并保留距离小于50米的POI
- 矩形过滤方法的复杂度为：40万*矩形过滤函数 + n*距离函数（n<<40万）。

随笔分类(57)

java(3)

LBS(10)

paper阅读笔记(2)

大数据(6)

定位原理/算法(3)

发表的SCI/SSCI(4)

服务治理(4)

空间索引原理(7)

数据库(5)

推荐相关(1)

线上问题定位及解决(2)

消息系统(2)

信息检索算法/实践(6)

应用服务器(2)

积分与排名

积分 - 115075

排名 - 2612

最新评论

1. Re:如何设计实现一个地址反解析服务？

如果仅仅是为了将用户坐标解析到道路级别的话，也未必需要用栅格。对于任意一条道路，根据历史记录，可以得到定位于这条道路的所有点，根据这堆点可以得到一个外包多边形，以后所有落在这个多边形内的点都可以认为是.....

--张可纯biubiu

2. Re:GeoHash核心原理解析

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？

--casterQL

阅读排行榜

1. GeoHash核心原理解析(43980)

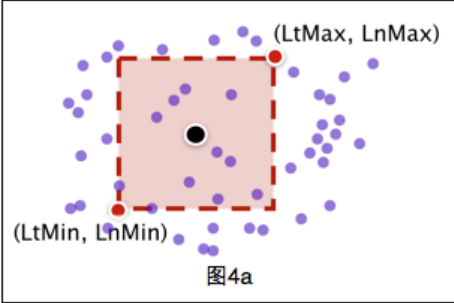


图4a

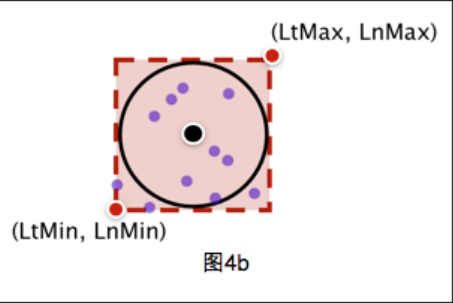


图4b

根据这个思路我们执行SQL查询（图5）（注：经度或纬度每隔0.001度，距离相差约100米，由此推算出矩形左下角和右上角坐标），发现过滤后正好剩下两个POI。

此查询花费了0.36秒，相比于方法一查询时间大大降低，但是对于一次查询来说还是很长。时间长的原因在于遍历了40万次。

```
mysql> select id, name from poi_spatial where lng between 116.3284 and 116.3296 and lat between 39.9682 and 39.9694;
```

id	name
32	[REDACTED]
37	[REDACTED]

2 rows in set (0.36 sec)

图5

方法三：B树对经度或纬度建立索引

方法二耗时的原因在于执行了遍历操作，为了不进行遍历，我们自然想到了索引。我们对纬度进行了B树索引。

```
mysql> alter table poi_spatial add index latindex(lat);
```

此方法包括三个步骤：

- 通过B树快速找到某纬度范围的POI（图6a），个数为m（m<40万），复杂度为Log(40万)*过滤函数；
- 在步骤a过滤得到的m个POI中查找某经度范围的POI（图6b），个数为n（n<m），复杂度为m*过滤函数；
- 用球面距离公式计算位置与步骤b得到的n个POI的距离（图6c），并保留距离小于50米的POI

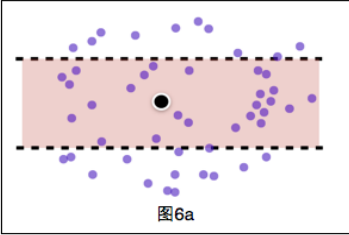


图6a

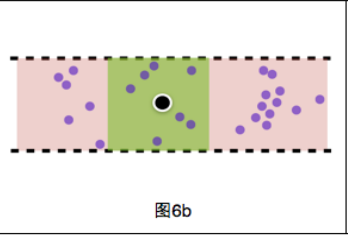


图6b

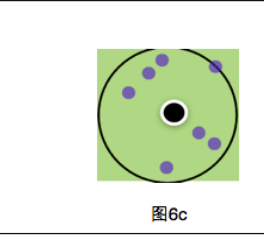


图6c

执行SQL查询（图7），发现时间已经大大降低，从方法2的0.36秒下降到0.01秒。

```
mysql> select id, name from poi_spatial where lng between 116.3284 and 116.3296 and lat between 39.9682 and 39.9694;
```

id	name
32	[REDACTED]
37	[REDACTED]

2 rows in set (0.01 sec)

图7

三、B树能索引空间数据吗？

http://www.cnblogs.com/LBSer/p/3392491.html

2/5

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)

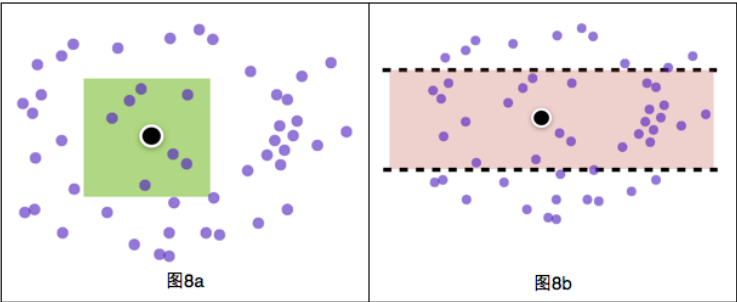
评论排行榜
1. 地图匹配实践(82)
2. 如何用消息系统避免分布式事务？(42)
3. 你应该知道的RPC原理(23)
4. GeoHash核心原理解析(22)
5. 地理围栏算法解析（Geo-fencing）(20)

这时候有人会说了：“方法三效果如此好，能够满足我们附近POI查询问题啊，看来B树用来索引空间数据也是可以的嘛！”

那么B树真的能够索引空间数据吗？

1）只能对经度或纬度索引（一维索引），与期望的不符

我们期待的是快速找出落在某一空间范围的POI（如矩形）（图8a），而不是快速找出落在某纬度或经度范围的POI（图8b），想象一下，我要查询北京某区的POI，但是B树索引不仅给我找出了北京的，还有与北京同一纬度的天津、大同、甚至国外城市的POI，当数据量很大时，效率很低。



2）当数据是多维，比如三维（x，y，z），B树怎么索引？

比如z可能是高程值，也可能是时间。有人会说B树其实可以对多个字段进行索引，但这时需要指定优先级，形成一个组合字段，而空间数据在各个维度方向上不存在优先级，我们不能说纬度比经度更重要，也不能说纬度比高程更重要。

3）当空间数据不是点，而是线（道路、地铁、河流等），面（行政区边界、建筑物等），B树怎么索引？

对于面来说，它由一系列首尾相连的经纬度坐标点组成，一个面可能有成百上千个坐标，这时数据库怎么存储，B树怎么索引，这些都是问题。

既然传统的索引不能很好的索引空间数据，我们自然需要一种方法能对空间数据进行索引，即空间索引。

下节将对空间索引分类体系、原理、优缺点及数据库支持情况进行阐述（正在写）。

转载请标明源地址：<http://www.cnblogs.com/LBSer>

分类: 空间索引原理

好文要顶

关注我

收藏该文

zhanlijun

关注 - 5

粉丝 - 664

±加关注

« 上一篇：[分布式追踪系统dapper](#)

» 下一篇：[深入浅出空间索引：2](#)

80

posted @ 2013-10-28 15:16 zhanlijun 阅读(8120) 评论(7) 编辑 收藏

评论列表

#1楼 2014-11-25 00:45 见吻戏哦

b树的确不适合多个字段的索引，期望空间索引是怎样解决的，楼主的写法很通俗，非常支持支持(0) 反对(0)

#2楼[楼主] 2014-11-25 20:34 zhanlijun

@ 见吻戏哦 thks	支持(0) 反对(0)
#3楼 2014-11-25 20:35 见吻戏哦	
楼主什么时候分布四叉树的文章啊	支持(0) 反对(0)
#4楼[楼主] 2014-11-25 20:56 zhanlijun	
@ 见吻戏哦 争取周末吧	支持(0) 反对(0)
#5楼[楼主] 2014-11-25 20:57 zhanlijun	
@ 见吻戏哦 近期的一些精力主要放在模型算法上	支持(0) 反对(0)
#6楼 2016-09-17 17:00 文轩云阁	
@ 见吻戏哦 R树家族	支持(0) 反对(0)
#7楼 2016-12-10 13:50 hheedat	
“我们期待的是快速找出落在某一空间范围的POI（如矩形）（图8a），而不是快速找出落在某纬度或经度范围的POI（图8b），想象一下，我要查询北京某区的POI，但是B树索引不仅给我找出了北京的，还有与北京同一维度的天津、大同、甚至国外城市的POI，当数据量很大时，效率很低。”这个效率还好吧	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](#)。

- 【推荐】超50万VC++源码：大型组态工控、电力仿真CAD与GIS源码库！
- 【缅怀】传奇谢幕，回顾霍金76载传奇人生
- 【推荐】业界最快速.NET数据可视化图表组件
- 【腾讯云】买域名送解析+SSL证书+建站
- 【活动】2050 科技公益大会 - 年青人因科技而团聚



- 最新IT新闻:
- 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
 - 手机厂商群撩小程序，是隔靴搔痒还是釜底抽薪？
 - 外卖平台该不该将“准时送达”服务变成增值业务？
- » 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...](#)

Copyright ©2018 zhanlijun