

# 研究生毕业论文

## (申请工程硕士学位)

论 文 题 目 地理轨迹相似性分析服务的设计和实现

作 者 姓 名

学 科、专 业 名 称 软件工程

研 究 方 向 软件工程

指 导 教 师

2019年4月10日

学              号 : **MF1732038**  
论文答辩日期 : 年 月 日  
指 导 教 师 :              (签字)

# **Design and Implementation of Geographic Trajectory Similarity Analysis Service**

By  
**(Author Name)**

Supervised by  
(Supervisor's position)**(Supervisor's Name)**

A Thesis  
Submitted to the XXX Department  
and the Graduate School  
of XXX University  
in Partial Fulfillment of the Requirements  
for the Degree of  
**Master of Engineering**

May 2019

# 研究生毕业论文中文摘要首页用纸

毕业论文题目： 地理轨迹相似性分析服务的设计和实现

软件工程 专业 2017 级硕士生姓名： \_\_\_\_\_

指导教师（姓名、职称）： \_\_\_\_\_

## 摘要

地理数据，是直接或间接关联着地球上某个地点的数据，也是表示地理位置、分布特点的自然现象和社会现象的诸要素数据。在使用地理数据的用户业务中，一个新颖而有实用价值的研究方向是针对地理轨迹的相似性分析。在进行犯罪同伙分析时，一般的犯罪同伙，都有着相同或者类似的运动轨迹。通过对相似轨迹的检索和相似度的排序，可以迅速找出最有可能协同犯罪的嫌疑人，从而极大提高侦察的效率。但是目前业界相关系统基本上都是关注轨迹的存储、展示和跟踪，对于轨迹相似性检索，并没有良好而稳定的解决方案。

针对以上问题，本文设计和实现了一个高性能的地理轨迹相似服务。该服务包括两个主要的数据子服务，轨迹数据子服务和地图瓦片数据子服务，具体的实现方法是基于优先点树结构扩展实现全文检索库Lucene的VpTree索引，以此作为轨迹数据检索的核心数据结构，并围绕该数据结构开发了轨迹数据插入和初始建立索引的功能。而地图瓦片数据子服务则是基于Mapbox通用地图瓦片标准，WGS-84坐标系和Web墨卡托投影法等地理数学知识实现了地图瓦片的读取和局部更新等功能。地理轨迹相似服务正是从以上这两个子服务中分别获取轨迹数据和地图瓦片数据，然后通过前端库的整合，实现在浏览器端的可视化，从而使用户方便快捷地达成相似轨迹检索的目标。

本文的主要工作就是介绍以上两个子服务的设计和实现。首先，本文介绍了目前在地理轨迹相似服务中涉及到的，主要的技术概念和相关标准。然后对整个系统进行了需求分析，并使用软件架构4+1视图模型进行整体架构的设计。接着分别介绍了轨迹数据子服务和地图瓦片数据子服务这两个主要功能模块的详细设计和实现。在该过程中还讨论并分析了系统中出现的一些性能以及模型上的问题，提供了更好的解决方案。在这之后介绍了整个服务的测试情况，包括单元测试和功能测试的流程细节，并通过折线图和条形图展示了服务性能测试的结果。最后，本文详细分析了整个服务在地理轨迹相似分析方面的效果提升以及当前版本的局限性，并对服务未来的发展前景进行了分析和展望。

**关键词：** 地理数据，轨迹数据，地图瓦片数据，优先点树，轨迹相似性

# 研究生毕业论文英文摘要首页用纸

THESIS: Design and Implementation of Geographic Trajectory Similarity Analysis Service

SPECIALIZATION: Software Engineering

POSTGRADUATE: (Author Name)

MENTOR: (Supervisor's position)(Supervisor's Name)

## **Abstract**

Geographic data is data that directly or indirectly relates to a certain place on the earth. It is also the data of the natural phenomena and social phenomena that represent the geographical location and distribution characteristics. In the user business using geographic data, a novel and practical research direction is to analyze the similarity of geographic trajectories. In the analysis of criminal accomplices, the general criminal accomplices have the same or similar trajectories. Through the retrieval of similar trajectories and the ranking of similarities, the suspects who are most likely to cooperate with crimes can be quickly identified, thus greatly improving the efficiency of reconnaissance. However, the current industry-related systems are basically concerned with the storage, display and tracking of trajectories. There is no good and stable solution for trajectory similarity retrieval.

In response to the above problems, this thesis designs and implements a high-performance geographic trajectory similar service. The service includes two main data sub-services, trajectory data sub-service and map tile data sub-service. The specific implementation method is based on the vantage point tree structure extension to realize the full-text search library Lucene's VpTree index, which is used as trajectory data retrieval. And the function of trajectory data insertion and initial indexing is developed around this core data structure. The map tile data sub-service is based on Mapbox universal map tile standard, and the geo-mathematics knowledge such as WGS-84 coordinate system and Web Mercator projection method realizes the functions of reading and partial updating of map tiles. The geographic trajectory similar service obtains the trajectory data and the map tile data from the above two sub-services respectively, and

then realizes the visualization on the browser side through the integration of the front-end library, thereby enabling the user to conveniently and quickly achieve the target of similar trajectory retrieval.

The main work of this thesis is to introduce the design and implementation of the above two sub-services. First of all, this thesis introduces the main technical concepts and related standards currently involved in geographic trajectories similar services. Then the requirements analysis of the entire system was carried out, and the overall architecture design was carried out using the software architecture 4+1 view model. Then the detailed design and implementation of the two main functional modules of the trajectory data subservice and the map tile data subservice are introduced respectively. In the process, some performance and model problems appearing in the system are also discussed and analyzed, providing a better solution. After that, the test of the entire service is introduced, including the process details of unit test and functional test, and the results of the service performance test are shown by a line chart and a bar chart. Finally, this thesis analyzes in detail the improvement of the geographic service trajectory similarity analysis and the limitations of the current version, and analyzes and prospects the future development of the service.

**Keywords:** Geographic data, trajectory data, map tile data, vantage point tree, trajectory similarity retrieval

# 目 录

## 表 目 录

# 图 目 录

# 第一章 引言

## 1.1 项目背景

轨迹数据指的是运动实体在空间中所经过点的集合，包括交通工具运动轨迹，人类活动的轨迹，动物季节性迁徙的轨迹等等[?]。每时每刻都在发生的事情实际上都在产生各自相应的轨迹数据[?]。这些数据在技术不成熟的过去几乎都被忽略掉了，而在移动互联网和卫星定位技术高速发展成熟的今天，稳定地捕获、记录、存储轨迹数据成为了可能。这为基于海量轨迹数据的商业分析应用奠定了基础。

当前，轨迹大数据已成为数据挖掘和地理空间数据提取等领域最有潜力的发展方向[?]。海量的轨迹数据潜在性地暴露了实体的活动特征、行为倾向和环境关系等信息[?]。这些细节信息的暴露，很大程度上是由于轨迹数据本身存在位置特征和时空特征上的关联性，这种关联性蕴藏着巨大的商业价值。而在所有这些的关联性中，最直观，最有商用价值的，就是轨迹的相似性。

与轨迹相似性有关的高价值应用场景很多，例如基于轨迹相似的用户分类，交通路线预测，犯罪同伙分析等等。因此，一个稳定高效的，支持轨迹数据相似检索和索引更新的轨迹数据系统是符合商业发展要求的必需产品。而要构建这样一个轨迹数据系统，必需考虑以下三个方面的问题。

第一，海量轨迹的存储。轨迹本身是具有时空特性的几何图形，而轨迹数据的量级一般都在千万级甚至亿级以上。这对于轨迹数据存储提出了很高的要求，传统的关系型数据库对于轨迹数据的存储和查询存在明显的性能瓶颈[?]。因此以什么样的结构存储轨迹数据是首先要考虑的重要问题。

第二，检索的数据结构和检索行为的定义。由于数据量很大，传统的预处理、排序、过滤等方法即使发挥到最大效应，也很难提供让用户满意的检索性能。因此，必须为轨迹数据建立定制的索引结构，并根据这种索引结构定义检索行为，才能让检索性能满足商用需求[?]。

第三，可视化运行。在当今大数据行业的发展背景下，数据可视化几乎是所有商业用户的共同需求。数据可视化很大程度上降低了系统的使用门槛，提高了系统的可用性，扩大了系统的适用范围[?]。而在地理轨迹数据系统的应用场景下，需要可视化的，除了轨迹数据之外，地图数据也必须要实现可视化[?]，否则的话，只有轨迹数据，没有其在对应地图上的状态显示，轨迹本身失去

了空间特性，退化为简单的几何图形，那么轨迹可视化本身也失去了意义。因此轨迹+地图的展示模式，才是一个完整的轨迹数据服务的可视化模式。

针对以上三个方面问题，本文设计并实现了地理轨迹相似服务，为用户提供高效，稳定的轨迹相似性相关功能服务。

## 1.2 国内外相关系统的发展概况

### 1.2.1 国内(外)轨迹数据系统发展概况

目前国内外轨迹相关系统所提供的轨迹分析功能着重于时空速度变化的捕获和分析，本文主要调查了百度鹰眼和ArcGis这两个比较成熟的系统。

百度鹰眼是一个开放的轨迹数据服务，它支持轨迹追踪，存储，运算和查询等功能，可帮助开发者管理百万级别的人类运动轨迹[? ]。百度鹰眼所支持的持续轨迹追踪功能是由鹰眼SDK提供的。鹰眼SDK可以实时地获取移动终端的空间地理位置，并持续回传轨迹数据，采集回传的频率一般在2s到5min这个区间内。轨迹回传的频率越高，对轨迹数据的完整性保留越好，轨迹数据越贴近真实。

百度鹰眼支持海量轨迹的存储，它提供了轨迹数据访问隔离和分布式存储，能够有效地保证数据安全。支持存储的内容包括坐标、速度、图片、视频和用户自定义字段等。在数据存储量上，百度鹰眼目前支持100万终端，储存1年的轨迹数据[? ]。

除了轨迹跟踪和存储之外，百度鹰眼还支持轨迹查询和展示功能。开发者可以很低延时地查询到终端的实时位置或者历史轨迹，并可以选择性地对历史轨迹进行回放。

最后，百度鹰眼支持轨迹数据分析功能。目前已提供的轨迹分析功能有驾驶行为分析，具体包括驾驶急速和超速判断，违法停车行为评估等方面。这些分析功能主要针对轨迹时间特性和速度特性提供服务。

ArcGis是由美国ESRI公司开发的地理信息系统系列软件，严格地讲，ArcGis1.0是世界上第一个现代意义上的，商业化的GIS软件。在轨迹数据服务方面ArcGis 提供了路径图层创建，障碍创建，停靠点编辑，轨迹运动方向生成，运动轨迹展示，轨迹运动分析等功能。其中轨迹运动分析包括所有停靠点最佳访问方式路径生成和展示。

综上可知，目前针对轨迹相似性的相关分析功能，在国内外已有系统中并不成熟和完备。

### 1.2.2 国内（外）地图瓦片数据系统发展概况

TileServer-GL是一个针对地图瓦片数据的开源地图服务器。它能够在服务器端使用MapBox GL内置引擎对矢量瓦片进行栅格化，进而为web应用和移动应用提供地图数据。它支持Mapbox GL JS,Android SDK,IOS SDK,Leaflet,OpenLayers, HighDPI/Retina, GIS via WMTS等众多前端库的数据调用[? ]。

TileServer-GL不仅能够提供地图瓦片数据，还提供了基于Mapbox GL Style的地图风格渲染。用户只要提供了有效的Mapbox GL style文件，TileServer就能够按照指定风格渲染地图数据，并返回给浏览器或移动端。

尽管TileServer-GL拥有代码开源，部署方便，使用简单等优点，但是它在商业应用领域存在以下两方面明显的短板。

首先，它的数据是保存在mbtiles文件中，而mbtiles是sqlite数据库的一种文件格式。TileServer-GL强耦合了这种数据库文件格式使得其对不同数据源缺乏扩展性，面对那些数据保存在传统关系型数据库或是列数据库中的用户，TileServer-GL将无能为力。

其次，TileServer-GL只能提供瓦片读取服务，而不能提供地图数据的实时动态更新，也就是说，整个服务从始至终都是无状态的。而在某些商业场景下，地图数据发生更新变化的可能性是非常大的。对于这种有更新要求的商业场景，TileServer也无法胜任。

综上可知，TileServer-GL能力有限，不能应用于本文所述的地理轨迹数据服务的实现中。

## 1.3 本文的主要工作

本文设计和实现了基于优先点树和地图瓦片的地理轨迹数据服务，其核心功能包括针对千万级别的轨迹数据的索引建立，相似轨迹的检索和新轨迹数据的插入，地图可视化以及地图数据更新等。本文主要工作有以下两个方面。

第一，设计和实现了Lucene Geometry vp-tree这个索引结构，也就是优先点树索引结构。这个数据结构是原生Lucene没有的。其主要功能是在JTS Geometry数据之上，建立一个多分查找的树结构，以最大限度地做到搜索剪枝，将搜索时间控制在 $n \log(n)$ 这个级别上。同时，为了满足用户使用场景中对索引本身的更新和初始化需求，本文还设计和实现了优先点树索引的初始化建树和插入新的轨迹数据的功能。

第二，设计和实现了地图瓦片服务Map Tile Service(简称MTS)。MTS是一个提供了瓦片读取和更新的地图瓦片服务器，能够为OpenLayer、Leaflet、GIS

via wmts 等多个前端库提供瓦片数据。相比于开源的TileServer-GL，MTS不仅能够提供地图更新功能，还具有良好的数据扩展性，提供多种数据存储方式的支持，能够实现与多种数据库的无缝功能对接。

## 1.4 本文的组织结构

本文的组织结构如下：

第一章 引言。介绍了项目背景和国内外相关系统的研究现状。

第二章 相关技术概念综述。介绍了项目中使用到的优先点树结构、lucene引擎结构与原理、地图瓦片相关概念、nodejs express框架、MapBox标准等相关技术概念。

第三章 系统需求分析与概要设计。通过需求列表分别展示了两个主要子服务的功能需求和非功能需求，通过用例图介绍了需求分析的结果，并针对重要的，操作复杂的用例使用用例描述表进行重点介绍。还通过软件架构4+1模型对整个项目进行了概要设计。

第四章 轨迹数据子服务详细设计与实现。在概要设计的基础上，对轨迹数据子服务进行详细设计并阐述具体的设计细节，以类图展示了类关系，以流程图和示意图介绍算法实现思路，并展示了关键部分代码和相似轨迹检索功能的运行效果图。

第五章 地图瓦片数据子服务详细设计与实现。在概要设计的基础上，对地图瓦片数据子服务进行详细设计并阐述具体的设计细节，以类图展示了类关系，以流程图和示意图介绍算法实现思路，并展示了关键部分代码。

第六章 系统测试。介绍了整个服务的测试环境，测试数据以及单元测试，功能测试，性能测试的情况，以测试列表描述测试方法和过程，以折线图和条状图描述测试结果。

第七章 总结与展望。总结论文期间所做的工作，并就轨迹数据服务的未来方向作了进一步展望。

## 1.5 本章小结

本章首先介绍了项目背景，阐述了建立地理轨迹数据服务的价值和必要性，以及主要面临的技术问题和挑战。随后介绍了国内外相关系统的发展情况，包括百度鹰眼、ArcGis和TileServer-GL，指出了这些已有系统的优缺点，分析了地理轨迹数据服务需要解决的需求点。最后介绍了本文的组织结构，包括各个章节的主要内容和行文思路。

## 第二章 相关技术概念综述

### 2.1 优先点树

#### 2.1.1 NN问题

Nearest Neighbour 问题，即最近邻居问题。指的是针对空间中的一个点集合，定义一个距离函数d（这里的距离函数d包括但不仅限于欧几里得距离），那么对于一个给定的目标搜索点q，找到距离q点的距离最小一个点[? ]，这就是最近邻居问题。相对应的，要找到与q点距离最近的K个点，就是KNN问题。

针对NN问题，最直观的做法，就是采用线性遍历的方式考虑所有点。这种做法不仅会因为大量空间点的遍历消耗很大，而且多维空间点之间的距离计算也非常复杂，其运行代价将会非常高昂。而一个比较合理的思路是，将二分查找的逻辑应用于点集合的检索，从而能将时间损耗降低为 $\log(N)$ 级别。也就是说，如果能够将点集合建立成某种有序的、树状索引结构。那么在搜索时，就可以通过类似于二分查找的方式达到 $\log(N)$ 级别的速度[? ]。

#### 2.1.2 优先点树概述

vp-tree(vantage point tree)，中文名称，优先点树，正是上述思路的一种实现。vp-tree的运行原理是基于三角不等式进行递归分解的剪枝，从而缩小检索范围，提高速度。其核心思想保证了两种情况下剪枝行为的正确性。第一种，是在检索过程中，对于那些远远超出搜索范围的分支，就不需要进行搜索了。第二种，是当搜索目标点显然在某一个范围内的时候，外部的其他分支就都不必搜索了[? ]。基于这两个原则，搜索点的数量和点之间距离计算的次数都被大幅度地减少，从而显著提升了搜索性能。

#### 2.1.3 优先点树基本原理与节点结构

vp-tree的基本思路就是对点集合进行空间划分。第一步，要选择一个点作为vantage point，也就是优先点，作为空间划分中心点。第二步，集合中的其他所有点要计算自己与优先点之间的距离。第三步，根据距离值的大小升序排序，然后将点集合均分为两支，距离小于等于中值的点组成left/inside子集合，距离大于中值的点组成right/outside子集合[? ]。第四步，以left/inside集合作为左子树的根节点，right/outside集合为右子树的根节点，再针对这两棵子树分别递归地

进行上述划分，从而形成一颗平衡的二叉树，如图??中①所示。而根据上述的点集合划分过程，可以得出一棵最简单的两路优先点树的结构如图??中②所示，其每个非叶子节点都包括一个用于标识优先点的VP-ID，一个中值mu和分别指向左右子树的两个指针。通过以上这样的平衡二叉树树结构，优先点树实际上

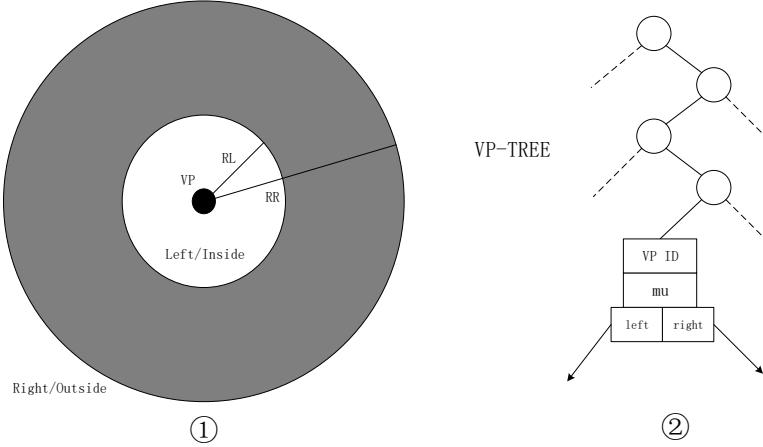


图 2.1: vp-tree点集合分割示意图

实现了对整个空间中的点集合进行了连续的球状二分。在整个数据空间中，大量的数据点集合被以不同的优先点为中心划分成了大量的相互交错层叠的球型子空间，如图??所示，[图片来自\[? \]](#)。实际上，优先点树的搜索过程，就是将搜索范围不断缩小到更小的球性空间内，最终确定距离最相近的结果的。关于优先点树的搜索过程，详见下文。

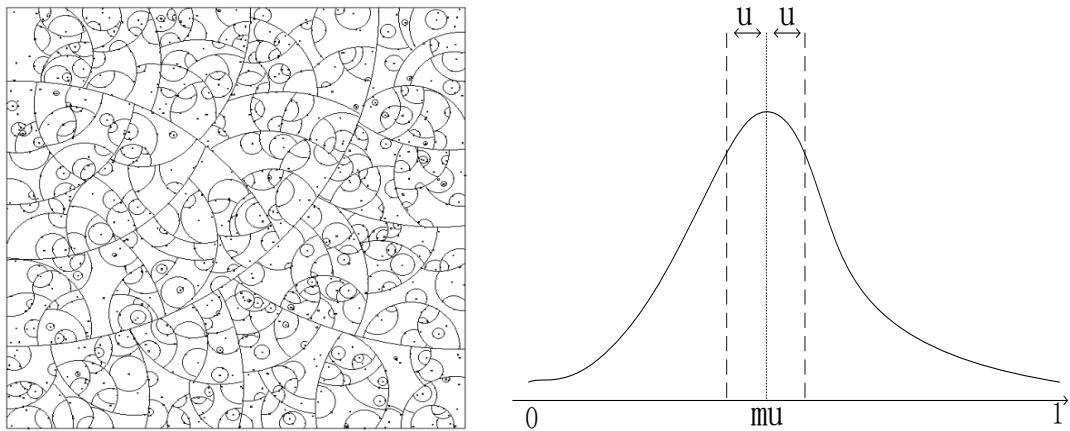


图 2.2: 球状空间分割

图 2.3: 优先点距离分布折线图

### 2.1.4 两路优先点树的搜索过程

优先点树的搜索算法的运行思路如算法??所示。对于一次检索的目标点target和当前优先点树的节点node，我们会设置一个容忍距离u。首先计算target与当前节点的优先点之间的距离distance。如果距离 $distance \geq mu + u$ ,那么基于三角不等式，左子树中不可能存在距离target点距离小于容忍距离u的点，从而舍弃左子树，只搜索右子树。反之，如果 $distance \leq mu - u$ ,就舍弃右子树，只搜索左子树。如果 $mu - u < distance < mu + u$ , 那么无法完成剪枝，左右子树都要搜索。

---

#### Algorithm 1: 最简单vp-tree的搜索过程search

---

```

1 node ← currentNode;
2 if node==null then
3   return
4 distance ← distance(target,node.vp);
5 if distance < u then
6   u ← distance;
7   best ← node.vp;
8 if distance ≥ mu-u then
9   search(node.right);
10 if distance ≤ mu+u then
11   search(node.left);

```

---

假设任意点集合的距离分布都符合如图??所示的正态分布。图中正中虚线为中值 $mu$ 的位置,左右两条虚线分别代表 $mu-u$ 和 $mu+u$ 的位置，那么显而易见的是，容忍阈值u越小，区间 $(mu-u, mu+u)$ 越短，目标点落入这一区间的可能性越小，剪枝成功的可能性越大，搜索性能也就越好。因此容忍距离u的选择至关重要，在实际的算法实现中，一般的做法都是随着递归搜索的过程推进而不断以当前距离target最小的距离取代容忍距离u，因为既然u是目前最小的距离，那么比u距离更大的点也就不必考虑了。因此，如何实现算法使得容忍阈值以较快的速度收敛，是算法实现的重点。在后文的详细设计中，本文会展开介绍具体方法。

## 2.2 Lucene

Lucene是一套用于全文检索的开放源代码Java程序库，由Apache软件基金会支持和提供。所谓全文检索，指的是针对无结构的，纯字符串内容的检索，而不仅是针对类似于日期，地点这样的结构化内容。Lucene 提供了一个简洁强大的应用程序接口，能够做全文索引和搜索。“在Java开发环境里Lucene是一个成熟的免费开放源代码工具；就其本身而论，Lucene 是现在并且是这几年，最受欢迎的免费Java 信息检索程序库[?]。”

这里要明确的是，“Lucene并不是一个完整有形的搜索引擎，而只是一个Java类库，对于不同的索引和搜索内容它是通用的，从而赋予了应用程序极大的灵活性和实现空间，而且Lucene 的设计紧凑而简单，能够很容易地嵌入到各种应用环境中[?]。”

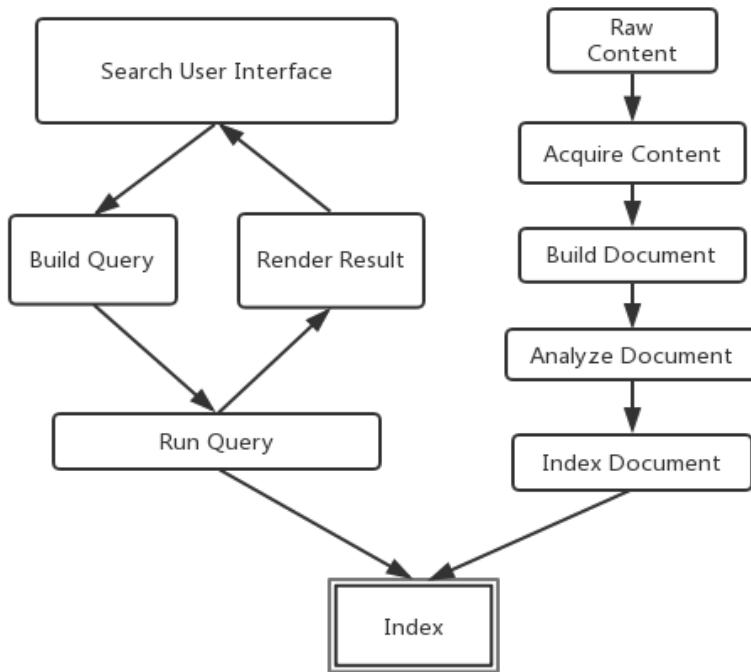


图 2.4: lucene检索组件图

Lucene的检索行为是基于索引进行的。其核心流程如图 ??所示，用户通过Search User Interface来与Lucene库进行交互，Lucene是基于索引Index进行检索的，其创建索引的过程是针对文档内容进行抽取，分词，索引的过程，而检索行为封装为Lucene Query,不同语义的Query作用于Lucene Index，再经由Render Result返回检索结果给用户[?]。注意：图中Index 的含义是Lucene各种检索数据结构的概称，并不单指倒排或正排索引。在本文的设计与实现中，轨迹相似检索功能就是封装成LuceneQuery来供用户使用的。

本服务涉及到的Lucene核心数据结构是Doc-Value。Doc-Value是Lucene针对文档内容建立的正排索引，可以将其理解为以Doc-Id为key的键值对的正序组合结构。这一结构与传统数据库中B-tree的索引结构是相似的。主要用于对一些特殊的结构化数据，比如图形，大数字，日志等不适合进行全文检索的数据进行顺序存储。本文实现中所使用的几何图形JTSGeometry就是存储在Doc-Value中的。

本服务之所以选择lucene作为基本存储框架，主要原因在于lucene本身是一个全文检索的代码库，其代码简洁，优雅，扩展性良好，而且运行稳定。扩展Lucene的索引结构，再应用于Lucene的Query语义是非常可控的设计行为。而且Lucene与本文所使用的几何体类库JTS.Geomtry是兼容的，因此可以在lucene环境中直接使用JTS的几何类库，降低了编码的难度。

## 2.3 地理相关技术概念介绍

### 1.地理坐标系

WGS-84坐标系（World Geodetic System—1984 Coordinate System），是一种通用的世界地心坐标系。该坐标系颁布于1984年，它充分利用了当时所有的技术和工艺，全面考虑了地球重力场，地理几何和地理测量等多方面的因素，制定了一个某一时刻的，按照某一恒定速率旋转的标准地球坐标体系。WGS-84坐标系的坐标原点为地球质心，Z轴方向与BIH（国际时间服务机构）1984.O定义的协议地球极（CTP）方向相同，X轴是协议地球子午面与WGS基准子午面的交线，Y轴与Z轴，X轴构成右手正交坐标系。WGS84坐标系也被称为1984年世界大地坐标系统，是目前最新的地球模型[?]。在本文的地图坐标应用场景下，所有的初始经纬度坐标都是WGS-84坐标值。

### 2.地图投影法

地图投影法，是遵循一定的数学规则，把不规则的地球表面的复杂地理信息映射到平面地图的理论方法。地图投影出于表达和理解上的正确性，必须遵循三个主要的不变原则，即角度不变，面积不变，距离不变。

墨卡托投影法（英语：Mercator projection），又称麦卡托投影法，是一种保证角度不变的圆柱形地图投影方式，因此也被称为等角正切圆柱投影。它以本初子午线与赤道交点为投影坐标原点，将赤道投影为X轴，将本初子午线投影为Y轴，从而构成墨卡托平面直角坐标系。此投影法之所以被称为墨卡托投影，是因为其作者为法兰德斯地理学家杰拉杜斯·墨卡托。1569年，他以此投影法为坐标基础绘制了一幅长202公分、宽124公分的世界地图。由于角度不变形的

保证，该地图经纬线于任何位置皆垂直相交。从而大大降低了地图信息的复杂性。

如公式??所示为墨卡托投影法的计算公式，其中a为地球椭球的长轴，b为短轴， $\theta$ 为经度弧度值，取值区间为 $(-\pi, +\pi)$ ，正值为东经，负值为西经。 $\phi$ 为纬度值弧度值，取值区间为 $(-\pi/2, +\pi/2)$ ，北纬取正值，南纬取负值，e为地球椭球体第一偏心率[? ]。

$$\begin{cases} x = a \times \theta \\ y = a \times [\ln \tan(\frac{\pi}{4} + \frac{\phi}{2}) + \frac{e}{2} \ln(\frac{1 - e \times \sin \phi}{1 + e \times \sin \phi})] \end{cases} \quad (2.1)$$

原生墨卡托投影法由于考虑到地球的椭球体事实，使得公式比较复杂，不利于计算机网络传输和计算，由此产生了简化的需求。Web墨卡托投影法就是为了方便地图数据在网络上计算传输渲染的，对原生墨卡托投影的一种简化变形的坐标投影法。它与原生墨卡托投影的一个重要区别是，Web 墨卡托投影忽略了地球作为一个椭球体的客观事实，出于计算简单，直接将地球视作一个标准球体，这就造成了Web 墨卡托投影法与原生墨卡托投影法存在0.33% 的精度差别。但是它的计算公式较之于原生墨卡托投影简化了很多，极大地方便了地图信息在网络间的传输和在浏览器上的渲染。在本文的详细设计章节的局部地图更新功能设计中，就是对公式??的等价变换实现，完成了WGS-84坐标到Web 墨卡托坐标的变换，从而将经纬度转换为瓦片坐标，作为局部更新的位置依据。

$$\begin{cases} x = a \times \theta \\ y = a \times \ln \tan(\frac{\pi}{4} + \frac{\phi}{2}) \end{cases} \quad (2.2)$$

### 3.地图瓦片

地图瓦片指的是经过Web墨卡托投影为平面的世界地图，在不同的地图分辨率(整个世界地图的像素大小)下，通过正方切割的方式将世界地图划分为像素为 $256 \times 256$ 的地图单元，划分成的每一块地图单元称为地图瓦片[? ]。每一个瓦片在地图平面内对应的横轴、纵轴坐标就是瓦片坐标。如图??所示，在zoom=1级别下，整张世界地图被切分成4块，每一块都是一张地图瓦片，整张地图在zoom=1级别的展示效果就是0-3这四张瓦片拼接起来的结果，以此类推，可知zoom=2和3的情况。显而易见的是，瓦片等级zoom 和瓦片坐标(tileX,tileY)一起唯一确定了一个二进制数据，也就是地图瓦片数据[? ]。在通用的数据存储格式中，一般都是把level,x,y 作为一张数据库表格的唯一主键，

瓦片数据作为列值进行存储。所有通用的地图瓦片数据系统，都是基于这样类型的数据库存储开发的[? ]。

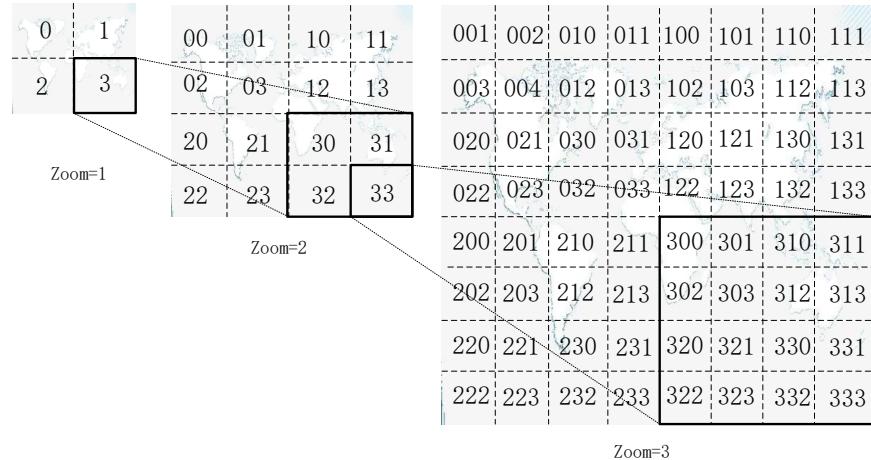


图 2.5: 地图瓦片层级示意图

#### 4. 瓦片金字塔

瓦片金字塔模型是在地图瓦片概念的基础上，自然产生的一种多分辨率层次模型[? ]。在图??中，实际上zoom=1,2,3这三种情况下的地图瓦片集就构成了一个瓦片金字塔。图中在zoom=2情况下，比例尺变为原来的2倍，分辨率变为原来的4倍，原本的zoom=1时的3号瓦片被切割成30,31,32,33这四个瓦片。而在zoom=3情况下，再次经过切割，原本的33号瓦片又被切割成330,331,332,333四个瓦片。也就是说，zoom精度每升一级，瓦片的数量就变为上一级的四倍，分辨率也变为原来的四倍，而由于每个瓦片的大小始终保持 $256 \times 256$ 像素平方大小，整张地图的地理范围也保持不变，因此每个瓦片所能展示的地理空间只有上一级的1/4。以此类推，可知zoom=4,5,6....的情况。对于一个特定地理空间的，若干zoom层级的瓦片数据，共同构成一个自顶向下的塔状结构，这就称为瓦片金字塔[? ]。

## 2.4 Nodejs Express框架

Express 是一个基于Node.js 平台的极简的、灵活的web 应用开发框架。它提供一套非常简洁的Rest风格服务接口，可以非常方便地将数据服务的接口暴露给调用者，使其获得想要的数据。本文应用Express 框架作为地图瓦片数据服务的基础框架，封装了地理数据空间转换的计算逻辑和与瓦片数据库的交互逻辑，降低了开发难度，提高了系统的可维护性。

## 2.5 本章小结

本章主要介绍开发地理轨迹相似服务所涉及到的相关技术概念，首先介绍了优先点树的概念、原理、结构和基本搜索算法，然后介绍了全文检索库Lucene的组件和检索原理，接着介绍了地理相关的技术概念，包括地理坐标系、地图投影法、地图瓦片和瓦片金字塔等概念。最后介绍了Nodejs Express框架的优势和在项目中的作用。

## 第三章 系统需求分析与概要设计

### 3.1 概述

地理轨迹相似服务总体上分为三个部分，轨迹数据服务，地图瓦片数据服务和可视化服务。其中可视化服务是用户直接操作的前端，其主要职责是调用轨迹数据服务获取轨迹数据和调用地图瓦片数据服务获取瓦片数据，然后利用前端库整合两部分数据，实现整个服务的可视化。这三者间的调用关系如图??所示。

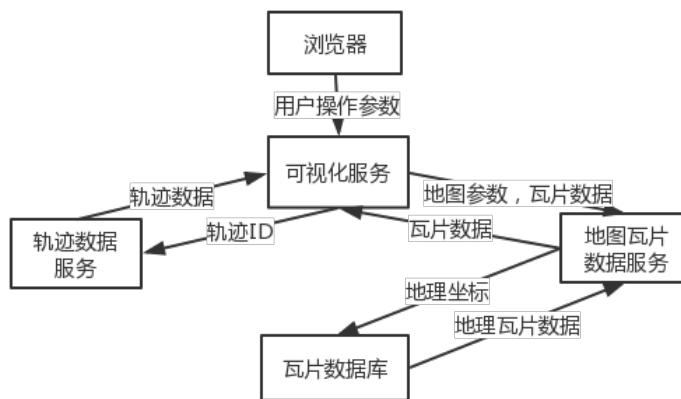


图 3.1: GTDS总体结构概览

本文设计和实现的重点是轨迹数据服务和地图瓦片数据服务这两个主要的，并列的子模块。其中轨迹数据服务的职责，是接收可视化服务的轨迹数据参数和轨迹ID，进行轨迹数据索引的创建，更新和相似检索功能。而地图瓦片数据服务的职责是接收可视化服务的地理数据参数或瓦片数据，对地图瓦片数据库进行相应的读写功能，从而为可视化服务提供地图背景的展示和动态更改等功能。

### 3.2 地理轨迹相似服务需求分析

#### 3.2.1 轨迹数据服务的需求分析

##### 轨迹数据服务功能需求

整个地理轨迹数据服务的核心功能有三个。第一个是对批量轨迹数据导入建立索引的功能。第二个是相似轨迹KNN检索功能，即根据目标轨迹的ID，执

行检索算法，找出与目标轨迹最相近的K条轨迹。第三个功能是对已有索引新加入数据点，即插入功能。注意，在本文所实现的轨迹数据服务中，只提供了增，查，初始化等功能，并没有提供删除功能。这样做是因为在大数据量的应用场景下，用户往往只需要动态地增加和查询轨迹数据，轨迹数据总体量一般都达到百万级，每次做KNN 检索的K值也能达到几百，用户并不在意索引中存在某一些冗余的轨迹，相关的删除操作优先级很低。所以本文放弃了删除轨迹功能的实现。

表 3.1: 轨迹数据服务功能需求列表

需求ID	需求名称	需求描述
R1	轨迹索引批量初始化	服务调用方能够通过批量上传轨迹数据，在规定时间内完成轨迹索引的建立，并返回结果。
R2	相似轨迹knn检索	服务调用方能够通过传递目标轨迹ID和检索量K，在规定时间内返回K个与目标轨迹最相似的轨迹。
R3	新轨迹数据的插入	服务调用方能够通过传递新轨迹ID和轨迹数据，将新轨迹插入到原索引中并返回插入结果。

### 轨迹数据服务的非功能需求

轨迹数据服务的三个主要功能都涉及到用户的直接操作。由于索引本身只是距离对比关系的存储，而并不直接存储轨迹数据，所以磁盘空间的消耗不是主要问题。而索引的建立，更新和检索，都需要大量的距离计算和区间比对，因此这显然是一个计算密集型的应用。因此用户的非功能需求主要体现在时间和准确率上。如表格??所示，时间特性和负载特性都是运行时保证的属性，而精度特性其实是在测试阶段保证的。因为在用户实际使用中，无法比对结果是否准确，只能在测试阶段通过比对结果集合和答案集合来确认准确率。

表 3.2: 轨迹数据服务非功能需求列表

时间特性	对于十万级别的轨迹数据量，服务应该在5s之内返回检索结果
负载特性	服务应该能应对10w以上的并发访问
精度特性	轨迹检索的结果准确度应该达到90% 以上

### 轨迹数据服务用例图

本文主要关注初始建立轨迹索引，插入新轨迹数据和相似轨迹检索这三个重要的功能用例，如图??所示。由图中可见地理轨迹相似服务的主要参与者包括普通用户和轨迹数据库管理员。

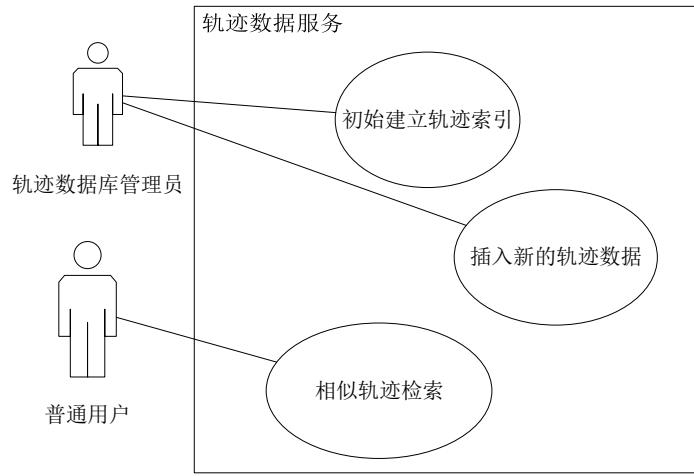


图 3.2: 轨迹数据服务用例图

### 初始建立轨迹索引用例描述

初始建立轨迹索引是轨迹数据库管理员的使用案例，也是整个轨迹数据服务必备的功能。本文选择在命令行环境下运行此功能。使用该功能最容易出现

表 3.3: 初始建立轨迹索引用例描述表

ID	UC1
名称	初始建立轨迹索引
参与者	轨迹数据库管理员
目的	将目标的轨迹数据集建立成优先点树索引结构，并返回操作结果。
描述	数据库管理员通过命令行参数设定轨迹数据表和数据条目区间，由服务完成索引建立。
优先级	高
触发条件	数据库管理员需要建立轨迹索引以备普通用户使用轨迹相似检索功能
前置条件	服务正常运行，数据库管理员打开服务命令行工具。
后置条件	优先点树索引正确建立，持久化到磁盘中，并返回结果给数据库管理员。
正常流程	1. 打开命令行工具。 2. 在命令行中输入轨迹数据表格ID、目标区间和优先点树的元数据，包括扇出度、节点数据量等。 3. 回车确认建立索引。
异常流程	2a. 表格ID不存在，服务端发现错误并告知用户。 2b. 筛选条件存在逻辑错误，比如以字符串值筛选整形列。服务器发现错误，要求重新输入。 2c. 目标区间的前后界大小矛盾，前端发现错误，告知数据库管理员重新输入。 4a. 由于网络，磁盘等各种可能原因导致的初始化失败，服务器应明确告知数据库管理员。

的问题是，对轨迹数据的元信息配置和对优先点树的元信息配置出现矛盾和错误。例如区间设定错误，文件格式错误，数据库表格配置错误等。具体如表格??所示，要充分考虑到用户设置的元信息本身错误可能严重影响索引性能的

情况，对于这些情况，系统必须予以更正或警告。除此之外，还要为关键参数设置默认值，以保证在数据库管理员遗漏配置项目时进行补充运行。

### 插入新轨迹数据用例描述

表 3.4: 插入新数据用例描述

<b>ID</b>	UC2
名称	插入新轨迹数据功能用例描述
参与者	轨迹数据库管理员
目的	新的，单个数据的到来，数据库管理员试图将新数据加入到原有索引中
描述	数据库管理员通过命令行工具提供要插入的新数据，由服务完成将新数据插入到原有索引中的操作
优先级	高
触发条件	数据库管理员需要向原有索引中插入新的轨迹数据，以备普通用户使用轨迹相似检索功能。
前置条件	服务正常运行，数据库管理员打开服务命令行工具。
后置条件	新轨迹数据正确插入到原索引中，并持久化在磁盘中，返回结果给数据库管理员。
正常流程	1. 打开命令行工具。 2. 数据库管理员提供新的轨迹数据，并选定要插入的目标轨迹数据集合。 3. 回车确认，完成新轨迹插入。
异常流程	2a.. 数据库管理员要插入的数据量不大，可以通过直接导入数据文件的方式进行。 2b. 用户要插入的数据量较大，则可以通过先导入到数据库中，再通过配置数据库表格元信息的方式来指定要插入的数据。 2c. 目标轨迹数据表不存在，服务器告知用户，重新输入参数。 2d. 要插入的轨迹数据已经存在于目标轨迹集合中，服务器赋予该轨迹新的唯一ID，直接插入。 3a. 由于网络，磁盘等各种可能原因导致的插入失败。服务器应明确告知用户，插入失败。

插入新轨迹数据功能是轨迹数据库管理员的使用案例。在真实得使用场景中，对于新数据插入操作，有少量的单个插入和批量插入这两种可能场景。针对这两种情形，轨迹数据服务工具应该分别支持直接导入数据进行插入和先导入数据，再配置数据元信息进行批量插入这两种方式，如表格??所示。这一过程中，我们不考虑轨迹重复的问题，即任何新插入的轨迹都被赋予一个独一无二的ID，被插入到优先点树索引中。而地理上相同的轨迹数据的插入位置可能在同一个叶节点，也由于叶节点分裂和数据重分布而在不同叶节点，这一点重复并不影响最终的检索结果，因此予以忽略。

### 相似轨迹检索用例描述

相对于初始建立索引和插入新的数据，轨迹相似检索功能不涉及到索引状态的改变。用户操作的变数主要在于检索轨迹的输入，目标集合的选定和K值的设定，如表格??。

表 3.5: 相似轨迹检索用例描述表

<b>ID</b>	UC3
<b>名称</b>	相似轨迹检索用例描述
<b>参与者</b>	普通用户
<b>目的</b>	用户想找出轨迹集合中与某一特定轨迹最相近的若干其他轨迹。
<b>描述</b>	用户通过浏览器选定要做相似性检索的轨迹，服务进行检索，并返回检索结果。
<b>优先级</b>	高
<b>触发条件</b>	用户要做轨迹相似检索。
<b>前置条件</b>	服务正常运行，用户进入操作界面。
<b>后置条件</b>	服务正确地根据索引进行检索，并将与目标轨迹最相似的若干条轨迹被返回给用户。
<b>正常流程</b>	<ol style="list-style-type: none"> <li>1.进入轨迹数据服务界面</li> <li>2.用户提供目标轨迹数据，设置检索匹配数目K，并选定要搜索的目标轨迹数据集合。</li> <li>3.点击确认，进行检索。</li> <li>4.返回检索结果集合。</li> </ol>
<b>异常流程</b>	<ol style="list-style-type: none"> <li>2a.用户要检索的轨迹如果是目标集合中已有的轨迹，此时用户只需要提供目标轨迹的ID即可。</li> <li>2b.用户检索的轨迹不在目标轨迹集合中，此时用户应该首先插入数据到目标集合中或者上传轨迹数据文件，是否插入数据应该由用户决定。</li> <li>2c.用户上传的轨迹数据文件错误，服务器端校验后返回错误结果并通知用户。</li> <li>3a. 目标轨迹数据集合不存在，服务器告知用户，重新选定。</li> <li>3b. 用户设定的K值过大，检索结果集中数量不足，服务器应该明确告知用户，结果集比期望数目不足。</li> <li>3c. 目标轨迹数据目标轨迹数据集合尚未建立轨迹索引，服务器告知用户，并询问是否新建索引。</li> <li>4a. 由于网络，磁盘等各种可能原因导致的检索失败。服务器应明确告知用户，检索失败。</li> </ol>

### 3.2.2 地图瓦片数据服务需求分析

#### 地图瓦片数据服务的需求综述

在轨迹检索服务中，对于地图瓦片数据的要求有增加，删除，修改，查询，数据源配置等。其中以更新功能的操作细节较多。而读取功能一般是直接被前端库调用的，不与真实用户直接交互，其逻辑也比较简单，所以本文予以省略，只用用例图详细介绍操作比较复杂的，与瓦片数据库管理员直接交互的地图瓦片更新功能的用例。注意：在GTDS的功能需求中，对于瓦片的增加和删除都是以整张地图为单位的，局部的增加和删除这种需求并不存在，所以此处不列为功能需求。

对于非功能需求，瓦片数据服务涉及到的计算主要是坐标转换，请求解析，缓存处理。这些操作的计算量不大，所以并不是计算密集型应用，而是IO密集型应用，高并发和快速响应是其必须满足的非功能特性。除此之外，用户的瓦片数据可能是存在于各种不同的数据库中的。因此服务还应该独立于不同的数据库，做到高可用性，高扩展性。

## 地图瓦片数据服务功能需求

表 3.6: 地图瓦片数据服务功能需求列表

需求ID	需求名称	需求描述
R1	新增地图瓦片数据	服务调用方能够通过瓦片服务，参数为地图名称和地图瓦片数据，增加一个地区的完整地图瓦片数据.
R2	删除地图瓦片数据	服务调用方能够通过瓦片服务，参数为地图ID，删除一个地区的全部地图瓦片数据.
R3	局部更新地图瓦片数据	服务调用方能够通过瓦片服务，参数为地图ID和地图数据，更新一张大地图中某一个小地区的地图瓦片数据.
R4	bounding-box更新地图瓦片数据	服务调用方能够通过瓦片服务，参数为地图ID，经纬度范围，地图数据，更新一张大地图中某一个地区某一纬度矩形范围内的地图瓦片数据.
R5	地图瓦片数据全量查询	服务调用方能够通过瓦片服务，通过获取全量数据的JSON文件，作为输入的数据源，获取整张地图的全量数据.
R6	单个地图瓦片数据查询	服务调用方能够通过瓦片服务，参数为地图ID和栅格坐标zxy，获取指定的瓦片.
R7	局部bounding-box的渲染查询	服务调用方能够通过瓦片服务，参数为地图ID和经纬度矩形范围，获取这一部分的地图渲染结果.

## 地图瓦片数据服务非功能需求

表 3.7: 瓦片数据服务非功能需求列表

时间特性	在正常负载情况下，服务的平均响应时间应在1s之内
负载特性	服务能稳定应对十万级别的并发访问，不会出现延迟超过10s或服务崩溃的情况
高可用性	服务能通过设置中间件的方式，便捷地对接到各种不同的数据库，并保证运行正常

如表格??所示，地图瓦片数据服务本身是服务于业务展示层，是面向地理相关用户的，也就是专用用户的。因此其负载只需能满足十万级别即可。而其时间特性，指的是业务展示服务调用瓦片数据的时间，即读取操作的响应时间。对于瓦片数据更新操作，其运行时间与数据量，网络传输，数据库写性能有关，无法明确定义。对于高可用性，目前用户的地理数据库包括Sqlite和Hbase,这两种数据库都应该能够通过配置中间件的方式，与地理数据服务对接起来运行。

## 地图瓦片数据服务用例图

以上用例中的功能并不是逻辑完备的，某些逻辑功能，比如局部地图瓦片非渲染查询，这种需求在实际应用中没有使用场景，这里就没有列举。而上图中的部分功能，例如地图瓦片的全量更新，删除整张地图，都是直接的瓦片数据库的操作，逻辑比较简单，所以本文后续章节对这些内容都予以忽略，不加介绍。

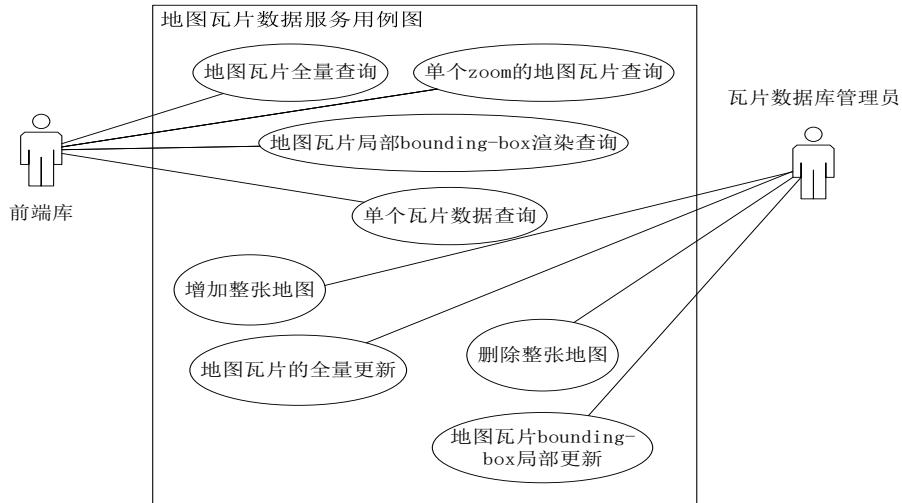


图 3.3: 瓦片数据服务用例图

### 地图瓦片数据服务用例描述

本文开发的地图瓦片数据服务，主要是服务于整个地理轨迹数据服务的可视化功能的。它的某些功能直接供前端库调用，而没有与人交互的场景。比如

表 3.8: 地图瓦片bounding-box局部更新用例描述表

ID	UC1
名称	地图瓦片bounding-box局部更新
参与者	地图瓦片数据库管理员
目的	更新某张地图中一个矩形范围内的瓦片数据，以改变轨迹展示背景
描述	数据库管理员上传新的瓦片数据，以实现对瓦片数据库中数据的修改，进而改变业务展示的结果
优先级	高
触发条件	某一地图中某一部分数据发生变更，需要更新为新数据。
前置条件	服务正常运行，数据库管理员进入服务界面
后置条件	地图瓦片数据完成更新，业务展示服务可以经过刷新可以看到地图变化
正常流程	1.进入瓦片数据更新界面 2.设置north,south,west,east,4个经纬度值 3.点击浏览本地文件并上传文件数据 4.点击更新瓦片数据
异常流程	2a. 经纬度大小值错误，前端应自动检测并警告 3a. 数据库管理员选择的文件格式错误，则前端对数据库管理员发出警告 4a. 由于网络，磁盘等各种可能原因导致的更新失败。服务器应明确告知数据库管理员，更新失败。

地图瓦片全量查询功能，是在地理轨迹数据服务运行时提供地图背景的功能，普通用户并不需要单独使用这个功能。而还有一些功能是直接操作瓦片数据，面向瓦片数据库管理员的功能。因此，瓦片数据服务的功能大都操作简单明确而且重复，无需使用用例描述。本文只对操作比较复杂的，与瓦片数据库管理

员直接交互的“地图瓦片bounding-box局部更新”进行了具体描述。如表格??所示。

### 3.3 地理轨迹相似服务架构设计

#### 3.3.1 架构设计综述

地理轨迹相似服务是一个包括两个主要功能模块，涉及到底层数据存储和逻辑计算的较复杂的数据库系统。它所面向的风险承担者，包括普通用户，轨迹数据库管理员，地图瓦片数据库管理员和系统的开发者和维护者，并且由于是实际的商业软件，对系统性能，稳定性，可用性都有相应的要求。因此，出于对架构设计全面性的考虑，本文选择使用软件架构蓝图4+1模型，从逻辑视图，过程视图，物理视图和开发视图四个角度并结合重要的使用场景，对系统架构进行详细和全面的设计。

#### 3.3.2 逻辑蓝图

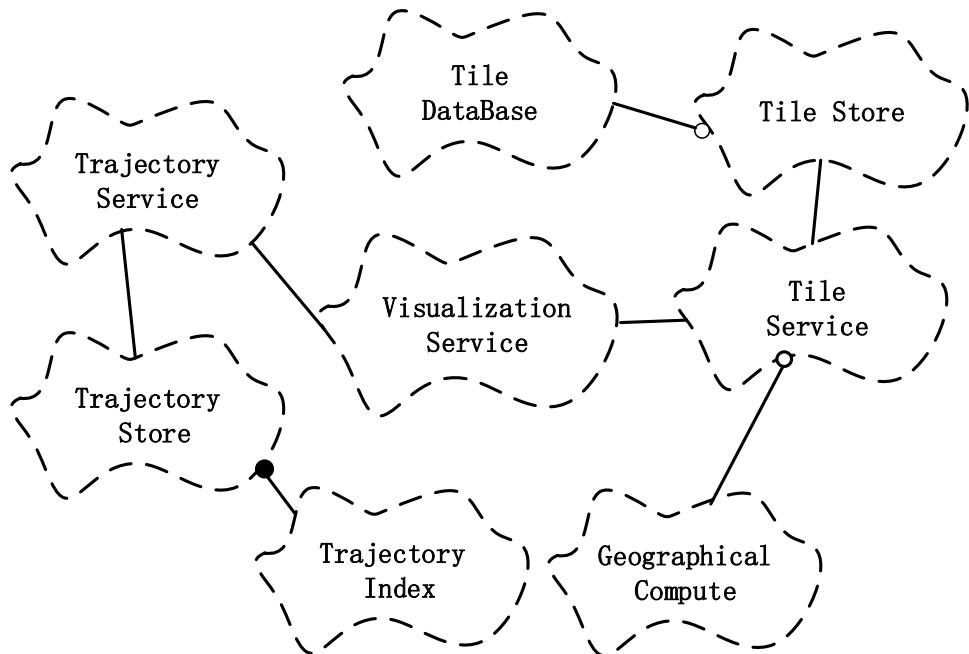


图 3.4: 地理轨迹数据服务逻辑蓝图

如图 ??所示，此为轨迹数据服务在对象视角下的逻辑视图，也就是逻辑蓝图。其中Visualization Service是可视化服务，也就是直接与用户交互的界面。它依赖两个主要的功能模块，Trajectory Service和Tile Service。

Trajectory Service是轨迹数据服务，它关联了Trajectory Store模块，也就是轨迹存储模块，Trajectory Store在存储轨迹数据的同时，还要为轨迹建立相似索引，因此Trajectory Service聚合了Trajectory Index模块。Trajectory Index模块负责轨迹数据索引的建立，更新和搜索功能，是本文设计和实现的重点。

Tile Service使用了Geographical Compute模块进行地理逻辑计算，两者只是调用关系，Geographical Compute模块本身是无状态的，可配置的。除此之外，Tile Service还关联了Tile Store模块，注意，Tile Store是地图瓦片存储处理逻辑的实现模块，其本身并不存储地图瓦片数据。Tile Store模块是连接Tile Service与Tile DataBase的桥梁，该模块通过对Tile Service所需接口的抽象化，实现各种不同地图瓦片数据库与Tile Service的无缝对接，从而为系统提供良好的可用性和可扩展性。

整个地理轨迹数据服务，就是以Visualization Service为中心，分别调用Trajectory Service和Tile Service中的轨迹数据和地图瓦片数据，通过前端库的渲染整合，为用户提供稳定的，高可用的轨迹数据服务。

#### 3.3.3 过程蓝图

如图 ??所示，此为地理轨迹数据服务的过程视图，也就是过程蓝图。在整个服务的运行周期内，Visualization Damon Process作为可视化服务的守护进程，时刻保持服务接口开放。Visualization Damon Process分别依赖Tile Damon Process和Trajectory Damon Process。

Tile Damon Process是地图瓦片数据服务的守护进程，当它收到可视化服务守护进程的瓦片数据调用消息之后，会首先开启Geographical Compute Task这样一个单线程进行地理坐标转换计算等处理逻辑，然后再根据地理计算结果，启动一个Tile DB Task作为瓦片数据库的数据操作线程。该线程从瓦片数据库中读取或者写入数据，从而完成用户的操作需求。

Trajectory Damon Process是轨迹数据服务的守护进程，在整个服务运行周期内，为可视化服务提供轨迹数据的检索和更新功能。当它收到消息之后，首先会单独启动一个轨迹索引进程Trajectory Index Process，该进程通过开启三个不同的线程任务，Search Task,Insert Task和Init Task来分别负责用户的轨迹相似检索，轨迹插入和轨迹初始建树的功能要求。之所以单独启动轨迹索引进程，是因为除了以上三个主要的线程之外，轨迹索引运行逻辑中还包括诸多并行的子过程，例如定位单个最相似点，路径回溯等。这些子过程的运行需要并发开启多个单独的线程运行，并维护各个线程的状态，所以此处必须启动一个轨迹索引进程。但是该线程并不需要周期性运行，在请求处理之后，即可关闭，以

避免进程长期占用过多系统资源。

综上可知，在整个地理轨迹数据服务的运行周期内，通过守护进程产生临时进程，再由临时进程产生并发的对应线程，来实现整个服务的正常运行。

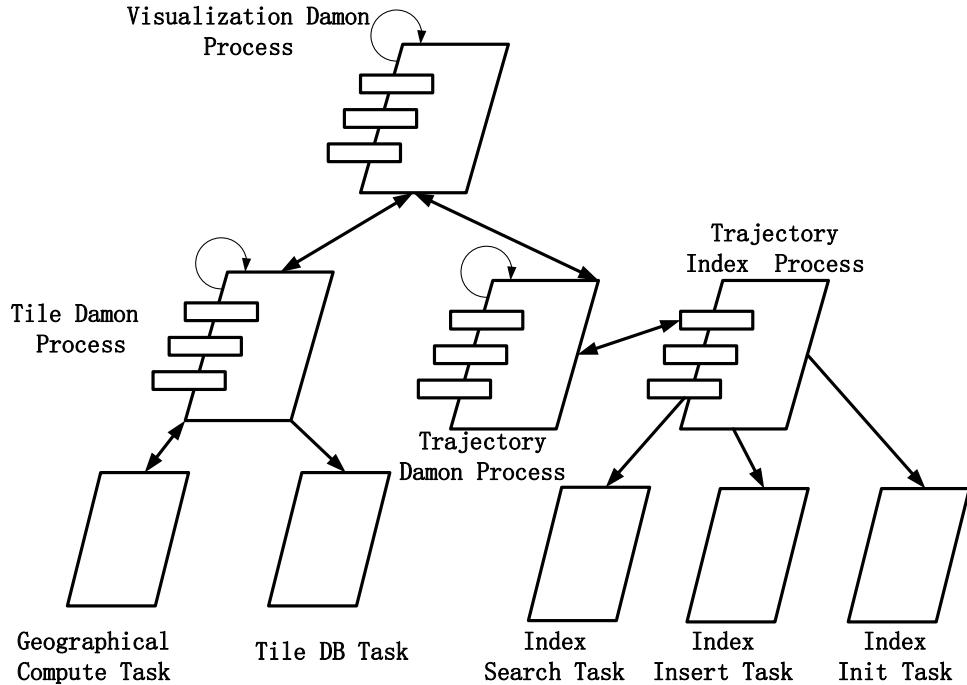


图 3.5: 地理轨迹数据服务过程蓝图

### 3.3.4 开发蓝图

如图 ??所示，此为地理轨迹数据服务的开发视图，也就是开发蓝图。本文按照各个子模块的职责和相互依赖关系，将整个服务划分为4个分层的开发组织结构。自底向上，第1层是运行于硬件和操作系统之上的基本要素，包括Trajectory Index轨迹索引和Tile DataBase瓦片数据库。第2层是基于JVM和Nodejs V8引擎实现的轨迹存储组件Trajectory Store 和瓦片存储组件Tile Store。第1层和第2层都是独立于用户使用域需求的职责模块，不论上层应用如何变化，都不影响这两层的设计与实现。第3层是轨迹数据服务Trajectory Service 和瓦片数据服务Tile Service，这两个组件是直接面向用户使用域需求的服务模块，它们分别依赖Trajectory Store和Tile Store实现轨迹数据和瓦片数据的存储。第1层Visualization Service 是直接与用户交互的可视化服务模块，它依赖于第2层的Trajectory Service 和Tile Service对用户操作做出响应。

综上，本文通过将地理轨迹数据服务划分成4层结构，明确和简化了各个组件之间的依赖关系，降低了开发复杂度。

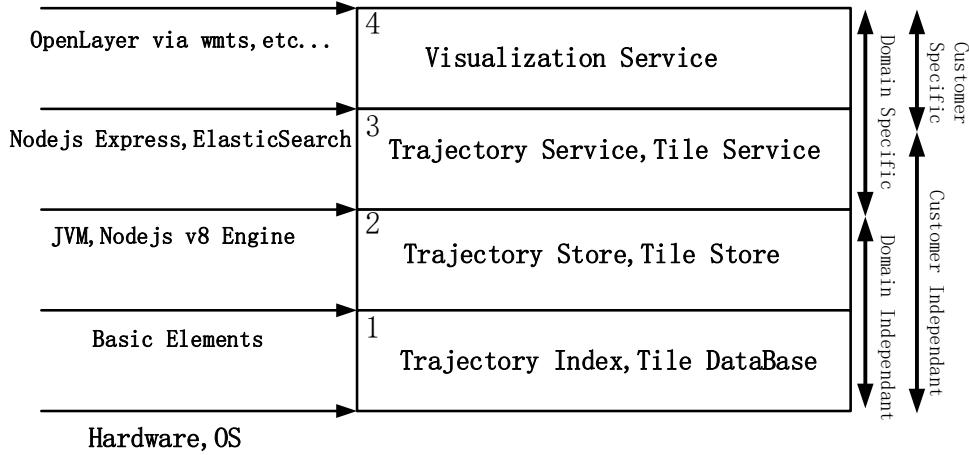


图 3.6: 地理轨迹数据服务开发蓝图

### 3.3.5 物理蓝图

如图 ??所示，此为地理轨迹数据服务的物理视图，也就是物理蓝图。它描述了软件总体结构与硬件的映射关系。本文构建的系统各组件之间关系简单清晰，因此在物理蓝图中保留了各个组件进程的映射。

图中 **Visualization Node** 是可视化子服务的运行节点，其上运行了可视化守护进程。 **Tile Service Node** 是运行地图瓦片数据子服务的节点，其上运行着瓦片守护进程以及地理计算任务线程。本文出于实现简单，并没有为地图瓦片数据服务设计HA功能，所以图中没有 **Tile Service Node** 的备用节点。 **Tile DataBase Node** 是地图瓦片数据库节点，其上运行着瓦片数据处理任务线程，并直接将数据处理结果持久化到硬盘中。

图中 **Trajectory Service Node** 是轨迹数据子服务的运行节点，其上运行了轨迹守护进程和轨迹索引进程。 **Trajectory Service Node** 既是逻辑处理节点，又是轨迹数据存储节点，所以并不依赖外部的数据库节点。

### 3.3.6 场景

本服务面向用户的功能场景有多个，本节选取最具代表性的相似轨迹检索功能使用场景，阐述各个组件之间协调合作关系。如图 ??所示就是相似轨迹检索场景图解，其各个步骤细节如下。注意，由于可视化服务对瓦片数据子服务和轨迹数据子服务的数据调用是并行的，所以图中编号分别采用了阿拉伯数字和罗马数字加以区分。

(1).当用户请求到来的时候， **Visualization Service** 首先向 **TileService** 发送地图区域参数和风格参数。

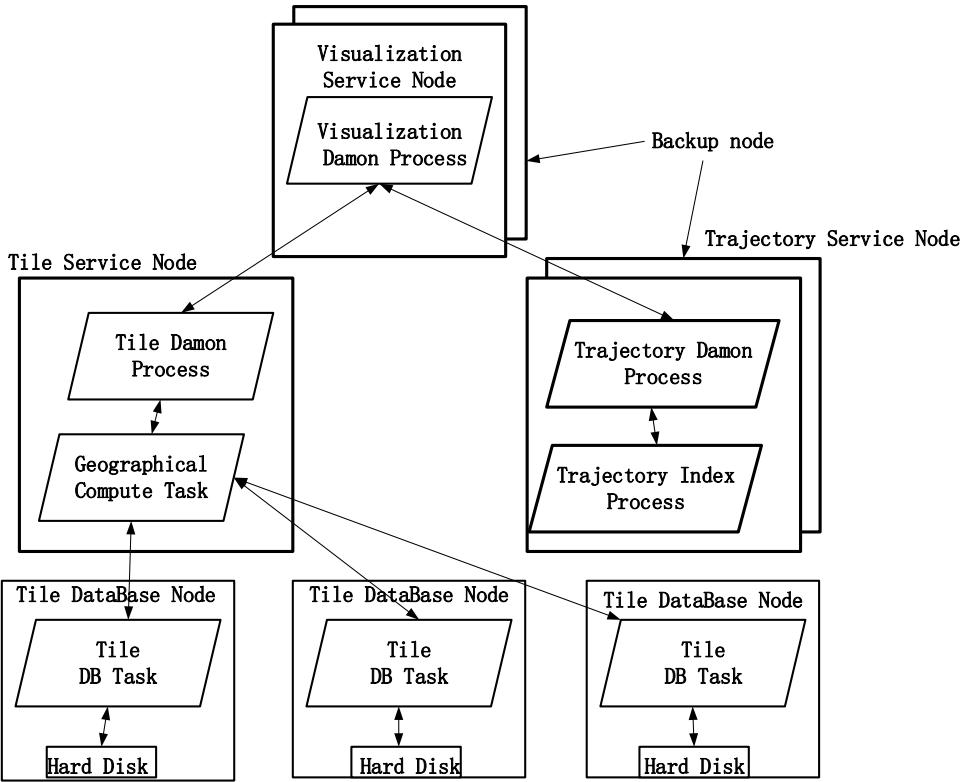


图 3.7: 地理轨迹数据服务物理蓝图

- (2). Tile Service 将 Area 参数转换为对应的瓦片坐标参数列表传递给 Tile Store 模块。
- (3). Tile Store 模块将瓦片坐标列表转换为瓦片数据库表的行主键列表，传递给瓦片数据库。
- (4). Tile DataBase 将查询到的瓦片数据返回给 Tile Store。接着在(5)(6)处一路回传瓦片数据。
- (I). Visualization Service 向 Trajectory Service 发送目标轨迹数据 Id。
- (II). Trajectory Service 发送目标轨迹数据 Id 给 Trajectory Index，进行相似轨迹检索。
- (III). Trajectory Index 将检索结果集中的轨迹数据 Id 列表返回给 Trajectory Service。
- (IV). Trajectory Service 向 Trajectory Store 模块发送轨迹数据 Id 列表，以获取对应的轨迹数据列表。
- (V). Trajectory Store 将轨迹数据列表返回给 Trajectory Service。
- (VI). Trajectory Service 返回轨迹数据列表给 Visualization Service，供可视化服务进行结果展示。

以上即为相似轨迹检索场景下的使用脚本。

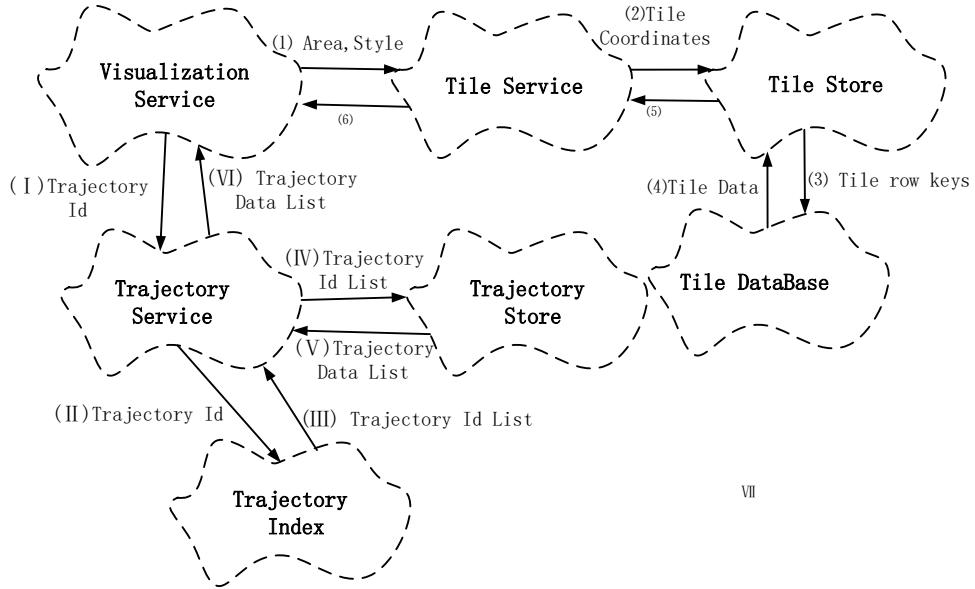


图 3.8: 相似轨迹检索场景脚本

### 3.4 本章小结

本章先是通过概述说明了整个系统主要组件的职责和相互之间的关系。然后分别针对轨迹数据服务和地图瓦片数据服务这两个主要组件进行了需求分析。需求分析的过程中，使用了需求列表来展现主要的功能需求和非功能需求，还使用了用例图的形式对各项功能进行划分，并对其中操作较为复杂的功能使用用例描述进行了详细介绍。最后，使用架构4+1模型方法，设计了整个服务在各个视图下的架构，并结合相似轨迹检索场景串联了各个组件的相互协作关系，为接下来按照模块进行详细设计和实现做准备。

## 第四章 轨迹数据子服务详细设计与实现

### 4.1 概述

轨迹数据服务是在全文检索引擎Lucene的基础上，扩展实现了单独的优先点树索引结构来实现的。因此，部署轨迹数据服务的方法，是通过将扩展后的Lucene库集成到ElasticSearch中，再基于ElasticSearch的服务节点运行的。

在轨迹数据服务中，选择以JTS-Geometry作为轨迹的存储结构。只考虑轨迹的几何属性，舍弃轨迹的方向性，也就是说轨迹线段AB和轨迹线段BA被认定是完全相同的路径。这种牺牲轨迹方向属性的做法可能会造成一定的精度损失，但是却能极大地简化设计和实现。

本服务选择使用豪斯多夫距离度量不同Geometry之间的相似性。两个Geometry之间的豪斯多夫距离越短，就判定两个Geometry越相似，也就是两个轨迹越相似。基于这样的前提，我们将Geometry作为度量空间中的数据点，实际上把相似路径检索问题转化为Geometry集合的KNN问题，然后针对Geometry集合，基于豪斯多夫距离值的大小关系建立优先点树索引，并定义检索行为，从而实现对轨迹相似检索问题的求解。

优先点树结构的功能模块主要分三个，分别是初始建树功能，相似轨迹检索功能和插入新轨迹功能。本章节将通过算法流程图、算法示意图和代码实现等方式详细阐述以上三个功能的设计和实现。

### 4.2 初始建树功能模块详细设计与实现

#### 4.2.1 优先点树索引类图

如图??所示，整个优先点树结构以VpTree为核心。其中VpTree直接实现了GeometryNNIndex这个接口。之所以要这样设计，是因为解决NN问题的索引结构并不是只有vpTree，另外还有Kd-Tree，R-Tree，mqr-tree等等[?]，本文所设计的VpTree只是其中一种。未来可能会有其他结构的实现对接到Lucene的索引逻辑，因此出于扩展性的考虑，必须预留接口。

Node是VpTree的实现的核心，是VpTree存储数据点Id，距离信息，子树边界，父节点指针，是否为叶节点等元信息的主要数据结构。Vptree的搜索算法和插入算法都是基于Node的上溯，遍历，分裂和元数据移动进行的。

DistanceCache是对距离的缓存结构，以减少数据点之间距离的计算次数。它其实是一个针对特定轨迹的缓存结构，一个DistanceCache保存其他若干轨迹

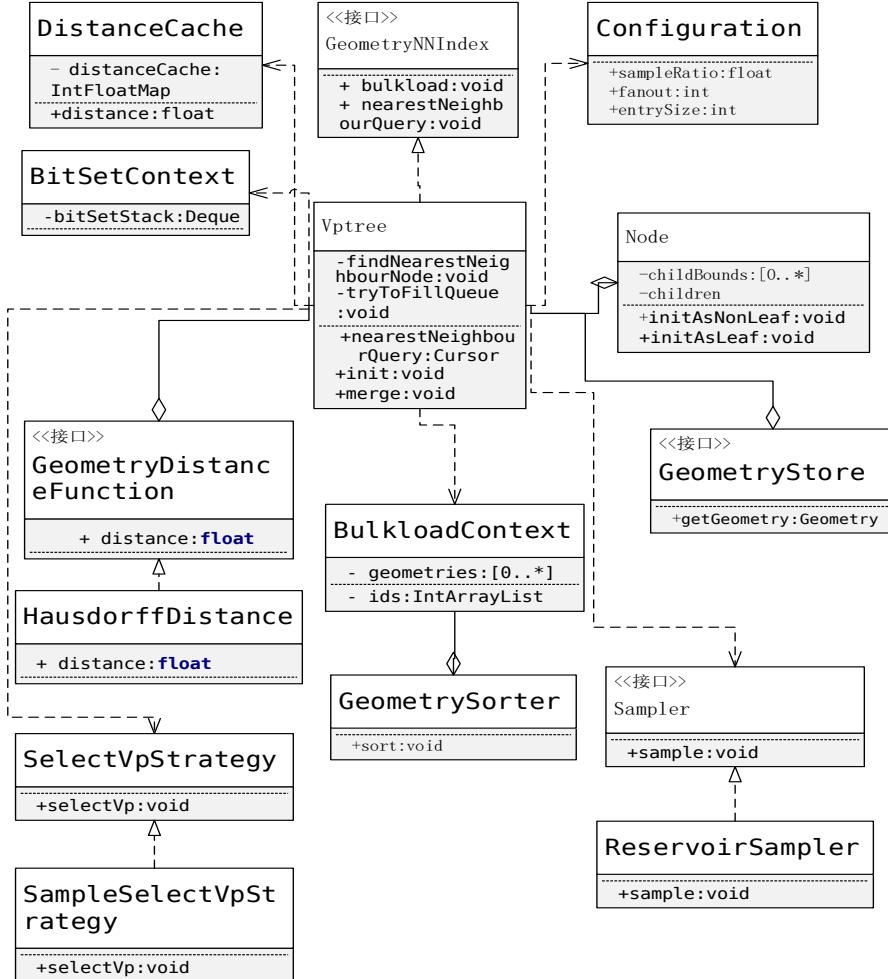


图 4.1: 优先点树索引类图

相对于目标轨迹的距离。所以在代码实现中，会有多个DistanceCache保存在内存里。为了避免在使用时一遍遍寻找特定轨迹的DistanceCache，本文在检索过程中保持DistanceCache 的栈与轨迹栈一一对应，从而顺序读取各个轨迹的cache,这在运行时消耗了相当数量的内存。

BitSetContext用于记录VpTree一条检索路径中各个分支节点的已搜索状态，用于避免重复检索。与distanceCache一样，BitSetContext通过顺序上的一一对应关系来维护和节点之间的映射关系。而且目前BitSetContext 的实现中，依赖于外部的释放内存空间，BitSetContext本身没有设计垃圾回收机制。

Configuration类用于保存VpTree的元信息，例如扇出数，最小扇出数，取样

方法，节点尺寸，采样率等。这个参数是VpTree的固有属性，在检索和插入操作中都被使用到。

BulkLoadContext是VpTree初始建树时的输入参数，主要包括一个docIdList和一个Geometrylist，这两者的主要功能是在初始建树的过程中提供轨迹数据。GeomtrySorter是一个排序工具类，用于对BulkLoadContext中的轨迹进行特定区间内的排序，以应对初始化建树时的多路切分，这个类是对快排的扩展实现。

GeomtryDistanceFunction是VpTree数据点间距离计算的接口，本文实现的算法HausdorffDistance采用豪斯多夫距离，但这显然不是唯一一种距离衡量方式。因此预留接口，以备扩充。

SelectVpStrategy是VpTree用于选择优先点的策略接口，本文目前只支持最大标准差的选取方法。其实，随机取优先点也是一种可选方式。Sampler是取样器的实现类，在最大标准差SampleSelectVpStrategy的实现中，使用了采用器进行采样，目前的采样器的实现只有一种ReservoirSampler。

综上所述，整个VpTree结构的设计，是充分面对扩展的。对于未来可能出现的新的距离计算函数，采样方式，节点结构，数据点形式，有良好的适应性。只是在内存使用上，存在一定量的冗余，这在未来版本的优化中会得到解决。

### 4.2.2 优先点树节点结构

本文所设计实现的优先点树的结构是对最简单vp-tree结构的改良，将原生vp-tree的两路结构改为多路结构。相应地，就需要把按中值进行二分修改为以边界值进行多分，并且为了提高检索的速度，减少检索时的距离运算量，改良后的vp-tree的非叶节点不仅存储了Vantage Point ID和每棵子树的指针，还存储了每个子树的距离值的上界和下界以及最大距离值。具体结构如图 ??所示为一个4路vp-tree的内部结构示意图。图中d1-d4分别是每棵子树轨迹集合的最大距离值，lower和upper是距离的上界值和下界值，child为子树根节点指针。

### 4.2.3 初始建树的流程

初始建树的输入数据是一系列的docId+Geometry，初始建树的输出是一棵完整的多路vp-tree。具体流程见 ??。

如图 ?? 所示初始建树实际上是一个递归的过程。但是为了避免出现内存溢出，本文选择了用循环+栈的模式。流程开始时，首先将根节点压栈，然后判断数据点的个数是不是小于叶节点的数据量标准。如果是，就意味着已经达到终止这一分支的条件，则直接初始化为叶节点，然后通过空流程返回循环判断。

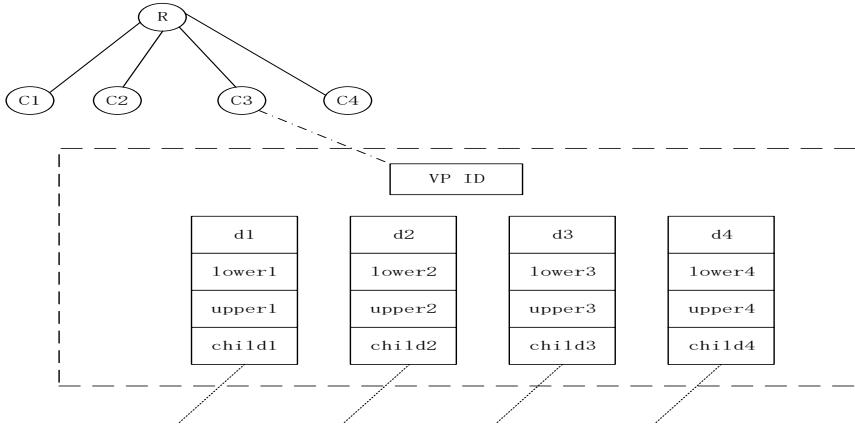


图 4.2: 4路vp-tree的内部结构示意图

如果数据量仍然大于叶节点的数据量，就使用选取函数选择优先点，计算其他各个数据点到优先点的距离，再根据距离对数据点进行升序排序。注意：由于此时有一个点被选做优先点，所以数据点总量要减一，然后判断剩余的数据量是否大于扇出数，如果数据量已经不能满足全部的扇出，那么就初始化为单个非叶节点再入栈，从而进入下一次循环。如果数据量依然足够分割全部的扇出，就按照距离排序的结果进行多路切分，用新建的子节点代表新的子树，并入栈。将距离值，上下界值和子节点指针分别填入对应的数组中，再进入下一次循环。以此类推，循环往复，完成所有数据点的建树操作。

#### 4.2.4 初始建树算法实现

如代码??所示，①处是初始化三个配置栈，包括偏移量栈offsetStack，长度栈lengthStack和节点栈nodeStack。②处是分别从三个栈顶部获取配置量，其中currentOffset 指示了当前节点在数据全集数组中的起始数据点的偏移位置，currentLength 指示了当前节点所包括的数据集合从当前偏移位置开始有多少个数据点，currentNode指向了当前节点。初始建树的输入数据是一段很长的数组，保存了要建立索引的Geometry。为了减少内存使用，本文采用偏移量+长度这样的组合量来记录每个节点所涉及的数据点集合，从而避免了输入数据的内存复制，后续的分裂过程是依靠这offset,length这两个当前节点的配置量作为分支处理依据。如图??所示，数据点数组为d0-d9，要生成一颗三路的优先点树，则在切分数据点数组时，除去vp点之外，其余10个数据点分为4,4,2三份，分别顺序对应n1,n2,n3的数据子集合，对应的偏移量分别是0,4,8。当进行下一层子树生成时，如图中虚线所示，将同一节点的偏移量和长度对应起来，即可在数据点数组中获得对应区间的数据，然后在下一层递归中进行相同分裂处理操作。

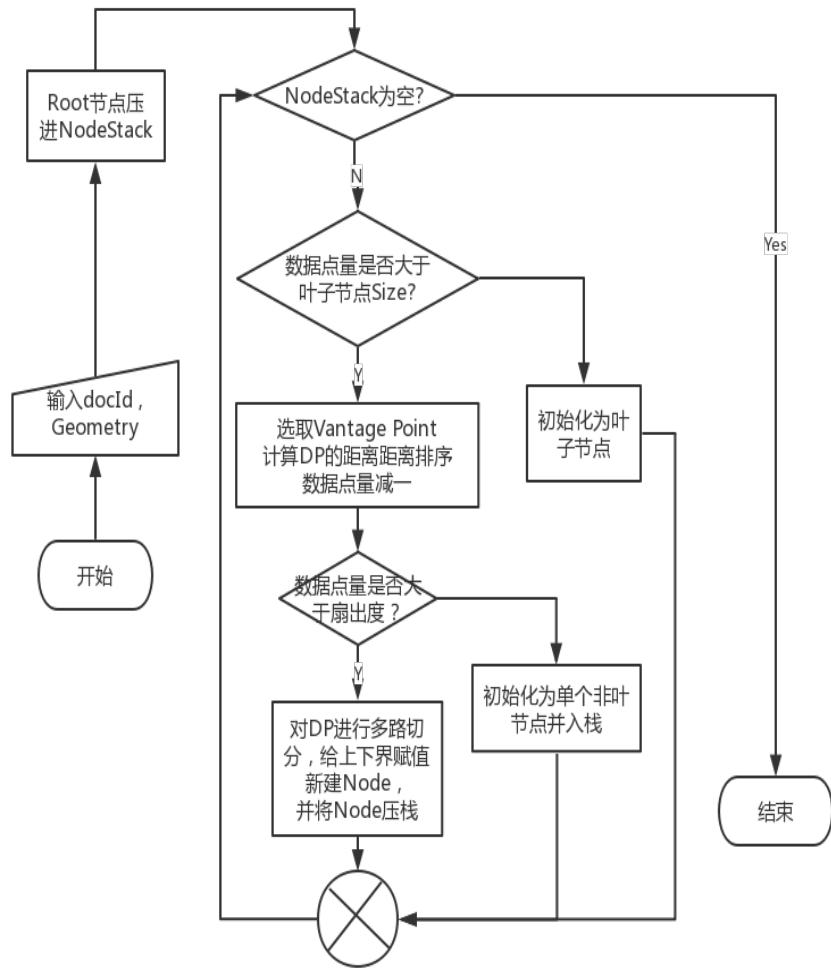


图 4.3: 初始建树流程图

此处显而易见的是，由于使用了栈实现从左向右的多路划分，所以优先级是从右向左，深度优先的，也就是说优先点树的最右边一个分支会最早完成创建，即n3所在子树最早被建立完成。此外，由于vp-tree 的多路划分是均衡的，所以vp-tree 自然是一个平衡树，分支初始化的顺序与最终结果没有关系。

代码??中③处，是在判定当前数据区间长度大于节点最小尺寸的条件下，先将其初始化为非叶节点并压入节点栈中。而由于当前长度因为选取优先点而要再减去1，所以在下一轮循环中，长度可能不再大于节点最小尺寸，而被重新初始化为叶节点。④处的逻辑与此类似，如果当前长度小于等于扇出度，同样先初始化为非叶节点，然后在下一层分裂处理时修正为叶子节点。④之后的else分支是对应正常的子树分裂逻辑，当新生成子树的时候，对应的子树指针和距离上下界等元数据都要进行设置。

```

private void createNode(Node root, BulkloadContext bulkloadContext,
SelectVpStrategy selectVpStrategy) {
    /.../ ①初始化节点栈，偏移量栈，长度栈，选择优先点策略
    while (!nodeStack.isEmpty()) {
        /.../ ②三个栈分别 pop 出栈顶，获取当前偏移
        if (currentLength > configuration.getEntrySize()) {
            currentNode.initAsNonLeaf(configuration);
            // ③初始化为单个非叶节点，选取优先点，排序
            selectVpStrategy.selectVp();
            if (currentLength <= fanout) {
                /.../ ④初始化为单个非叶节点，并入栈
            } else {
                int childSize = (int) Math.ceil(currentLength / fanout);
                //计算每个子树应有的数据点量
                for (int i = 0, start = 1, end; i < fanout; i++) {
                    end = Math.min(start + childSize - 1, currentLength);
                    /...../设置子树指针并分别为每路子树，设置最大距离值，距离上下界值
                }
            } else {/..../初始化叶节点，结束一个分支}
        }
    }
}

```

图 4.4: 初始建树代码

#### 4.2.5 设计要点：优先点的选择算法

在初始建树的过程中，优先点选取是非常关键的一步。选取的好坏取决于一个优先点能否让切分出来的各路子树的边界值相差足够大，因为各路子树的边界值相差越大，在检索的时候，与优先点的距离值和容忍距离构成的检索区间完全落入某一单个子树的边界内的可能性越大，剪枝成功的可能性越高，性能就越好。反之，如果优先点选择很差，导致多个子树的上下界非常接近，就很容易出现一个检索区间横跨多个子树区间的情况，这种情况下无法进行有效剪枝，多棵子树必须都被检索，导致性能下降。因此，如何选择尽可能好的优先点，是初始建树算法设计和实现的重点。本文针对优先点选择的实现是基于随记取样和标准差结果的。本文认为，一个点与其他所有数据点距离的标准差越大，越能将数据点空间均匀切分，作为优先点的性能越好。而由于数据全量很大，不可能都计算，就采用随机抽样的方式进行部分计算。其设计思路是，在数据点全集中随机取样K个点，作为候选的优先点。针对这K个候选的优先点进行循环遍历，每个候选优先点再随机取K个点作为参照点，然后计算每个候选优先点和参考点之间距离的标准差，最后取标准差最大的那个候选点作为真正的优先点。这样的选举过程涉及到取样，距离和标准差的计算，相当于用一部分初始建树性能的损耗来换取了检索性能的提高。

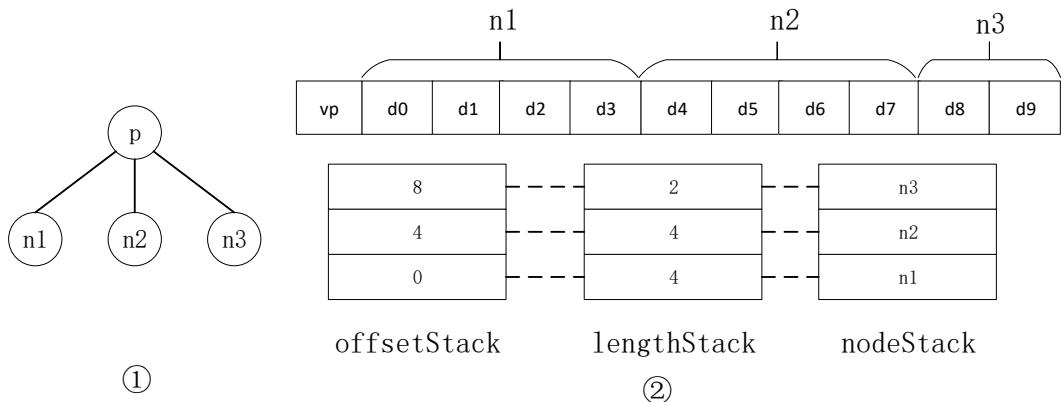


图 4.5: 节点分裂配置栈示意图

如代码 ??所示。①处是根据用户配置的取样率计算取样数量，②处设置一内一外两个取样结果对象以及使用取样器外取样，③处是使用取样器进行内取样，④处是遍历内取样每一数据点，计算与外取样对应数据点的标准差。最后在⑤处判断新的标准差是否大于原有最大标准差，并更新最大标准差数据点的记录。

### 4.3 相似轨迹检索功能模块详细设计与实现

#### 4.3.1 KNN问题的常规方法

在我们的轨迹数据服务的作用域内，KNN问题，即K Nearest Neighbour的含义是，找到与目标Geometry距离最近的K个Geometry。

原生vp-tree的搜索算法是面对NN问题，也就是nearest Neighbour问题，只找一个距离最近的点。那么面对KNN问题，显然不能通过简单地运用K次原生搜索算法来解决，那样会导致三次方级别的计算消耗，这是用户绝不会接受的。

比较直观的想法是，使用一个大小为K的最小堆，在搜索过程中实时更新这个最小堆的状态，那么在搜索算法走完的时候，这个最小堆中的结果就是K个距离最近的Geomtry。这是[?]中所阐述的思路。本文实现的算法的借鉴了这一思路，同样是用堆来动态维护检索实时状态，但采取了一些措施应对这一思路的明显短板以获取更好的检索性能。详见下节。

#### 4.3.2 设计焦点与解决办法

在KNN问题的搜索过程中，对于每一个多路节点，除了考虑与优先点的距

```

public void selectVp(BulkloadContext bldCtx, int curOff, int
curLen, int[] values, float[] disBuf, SelectVpResult result) {
    int spSize = Math.max((int) (curLen * conf.ratio), 1); // ①
    SampleResult sampleResult = new SampleResult(...),
    SampleResult sampleResultInner = new SampleResult(...); // ②
    sampler.sample(...);
    float maxStdev = -1;
    for (int i = 0; i < spSize; i++) {
        Geometry candidate = bldCtx.geometries[bldCtx.spBuf[i]];
        sampler.sample(...); // ③
        for (int j = 0; j < spSize; j++) {
            ....// ④
        }
        float current = computeStdev(disBuf, 0, spSize); // ⑤
        if (current > maxStdev) {
            //更新标准差最大的点
        }
    }
}

```

图 4.6: 优先点选取代码

离之外，还要考虑一个容忍距离T，也就是超出这个容忍距离T的数据点不再予以考虑[? ]。

这是距离范围剪枝的基本依据。如图??所示，某一数据空间基于优先点VP被分为S1,S2,S3三个子空间，目标轨迹target落在子空间s2中，T为容忍距离。在①情况下，以target点为中心，容忍距离T的半径范围覆盖了S1,S2,S3，无法剪枝，使得检索必须遍历全部三个分支，而实际上S1,S3很可能没有匹配的结果。而对于②而言，T半径仅仅只在S2子空间内，所以可以认定，S1 和 S3 中不可能包括与target 距离小于T 的点，从而排除S1,S3，实现剪枝。由以上两种情况可见，容忍距离越小，成功剪枝的概率越高，检索性能越好。但是因为数据点与VP 点距离集合本身是离散的，容忍距离也不能太小，否则过度剪枝造成检索不到结果。所以容忍距离的初始值设置和收敛速度是直接影响检索性能的关键[? ]。

在上文所述的单纯用最小堆动态更新检索结果的算法中，其最大问题在于，容忍距离是从正无穷开始更新的。这使得检索从一开始完全不可能做到剪枝，大量的分支都被搜索了。容忍距离的收敛速度会非常慢，导致剪枝的效率很低，性能也比较差[? ]。

基于上面所提到的容忍距离收敛较慢问题，本文采用了结果堆预填+分支

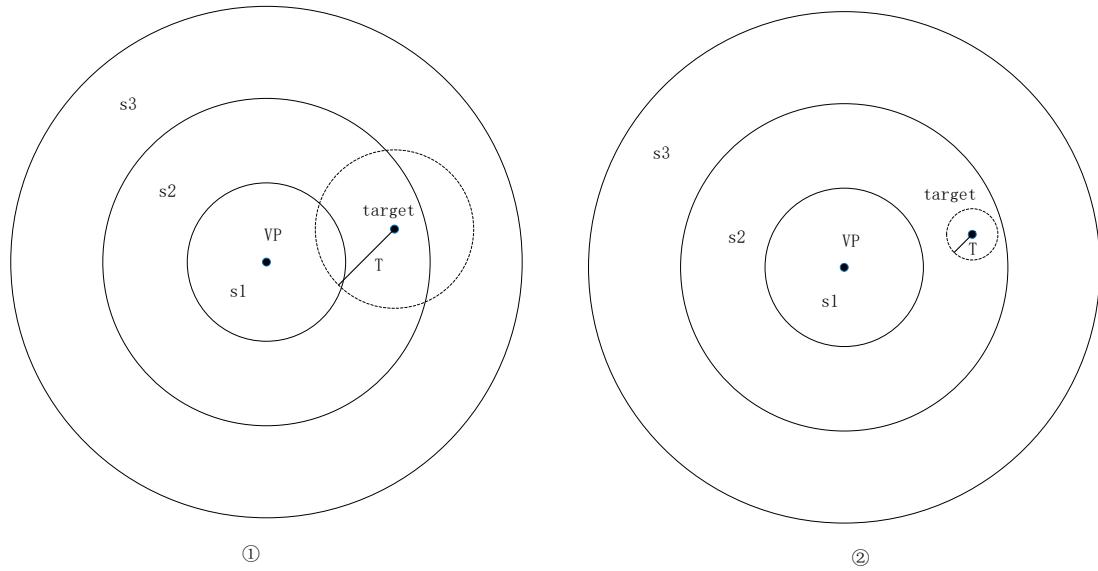


图 4.7: 容忍距离对剪枝的作用示意图

回溯的方式进行优化处理。即通过vp-tree原生搜索算法，先找到single nearest neighbour，也就是单个最相似数据点，并且在这个过程中预填结果堆，并保存从root到single nearest neighbour的整条路径的所有节点。然后以结果堆中最大距离作为容忍距离的收敛起点，回溯整条路径中的节点，完成检索。这样通过预填结果堆，使得容忍距离以更低的起点收敛，很多分支在检索开始阶段就被剪掉，容忍距离收敛的速度更快，性能更好。

#### 4.3.3 相似检索功能实现要点：避免重复访问

由上文所述，通过预填结果集的方式加快容忍距离的收敛速度，这种做法带来部分分支重复访问的问题。这是因为某些节点的一部分分支在预填过程中就已经被搜索过了。如果不对搜索进行标记，重复的搜索不会贡献新的结果，从而浪费了搜索时间。关于这一点，本文使用了BitSet 栈的方式来实现，在预填结果集的过程中，与节点栈同步保存一个Bitset 栈，两者的顺序一致。每个bitset 的大小与vp-tree 的扇出数一致，用于保存若干分支中，哪些已经被访问过了。这样，在回溯发生的时候，首先检查对应的bitset，对于那些已经被置位的分支直接跳过，只访问尚未遍历过的分支，避免重复的检索和距离计算。

#### 4.3.4 相似检索算法流程图

如图??所示为检索算法的运行流程图。图中显示了检索过程中最关键的三个子流程，即检索最相似点，预填结果堆和回溯更新结果堆。其中检索最相似

点是通过递归实现的，节点栈在其过程中的作用是保存搜索路径。而另外两个子流程都是使用栈+循环的方式实现的。详细说明，见后文代码实现。

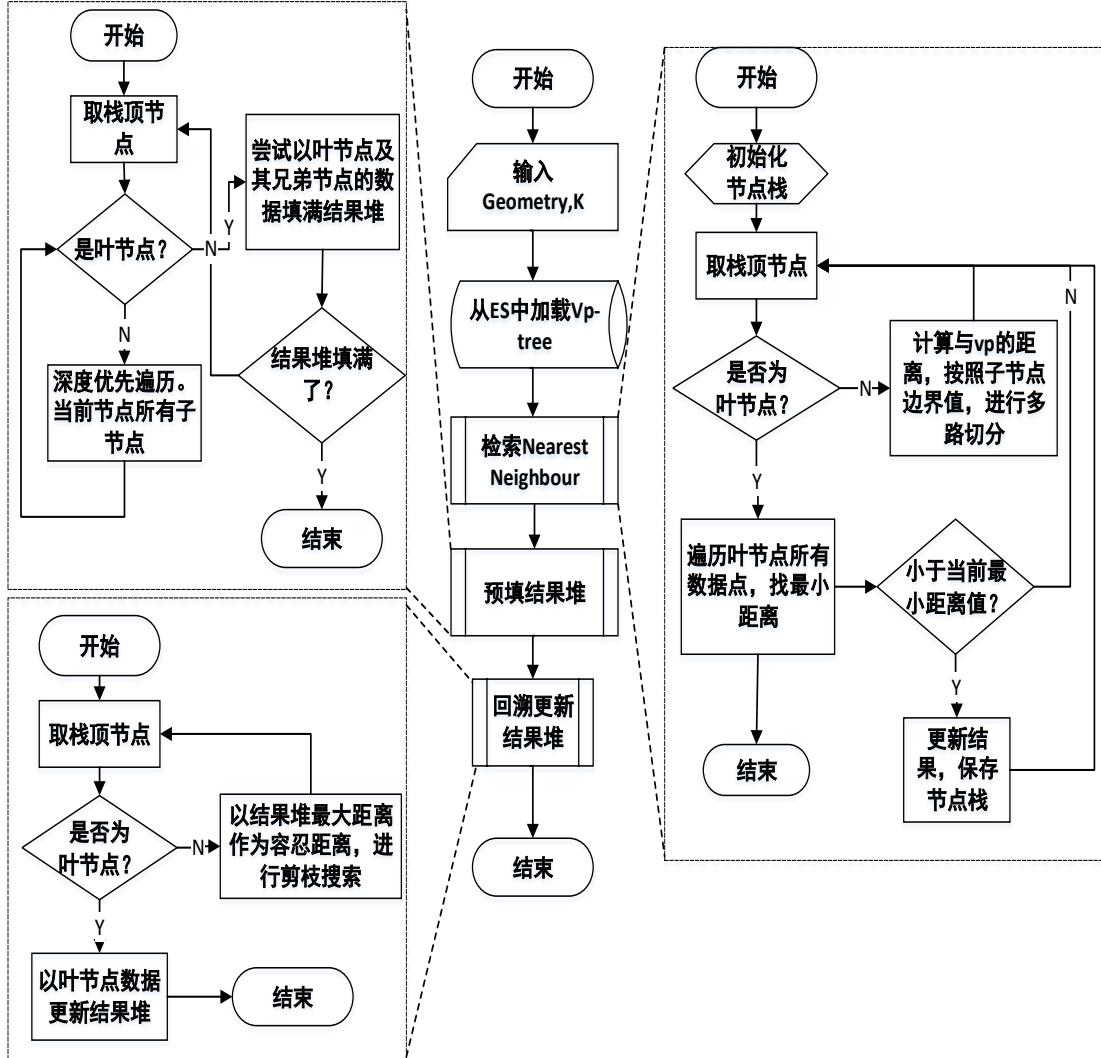


图 4.8: 相似检索算法流程图

#### 4.3.5 相似检索功能各子流程代码实现

本节将会依次展示??中所示的各个子流程的代码实现。

**检索单个最相似轨迹代码实现：**如代码??所示，①处是当节点为叶节点遍历所有数据点，如果有距离比当前最小值更小的数据点，则在②处更新当前最相似数据点搜索路径上所有点的序列。注意，此时nodestack 中的点序列即为当前路径下所有节点，因此清空nearest neighbour中原来的最近点路径，全部更新为nodestack中的节点序列。③处为判断被选定的优先点是否符合最近点条

件，如果符合也要予以更新。④处为获取非叶节点的子树配置信息，包括距离上下界，作为剪枝的条件。⑤处为计算每棵子树的上界和下界，注意，此处要用nearest neighbour当前最小距离为容忍距离，使其成为是否剪枝的标准条件。⑥处为递归搜索下一分支。

```

private void findNnNode(node, nodeStack, nn, distanceCache) {
    nodeStack.push(node);
    if (node.isLeaf()) {
        boolean update = false;
        IntArrayList children = node.getChildren();
        for (int i = 0; i < children.size(); i++) {
            /.....①遍历叶子节点
        }
        if (update) {
            /.....②更新搜索路径点序列
        }
    } else {
        //计算与 vp 点的距离 distance
        if (distance < nn.distance) {/.....③}
        /.....④获取非叶节点分支元信息，距离上下边界
        float low, high;
        for (int i = 0; i < size; i++) {
            low = cBounds.get(i * 2) - nn.distance;
            high = cBounds.get(i * 2 + 1) + nn.distance; //⑤
            if (distance >= low && distance <= high) {
                findNnNode(...); //⑥
            }
        }
    }
    nodeStack.pop();
}

```

图 4.9: 检索单个最相似轨迹代码实现

**预填结果集代码实现：**如代码??所示，①处首先要判断结果堆是否已经满了，然后在②处，从之前搜索的最相似单个数据点过程中保留的路径节点栈中，取出最后一个节点last及其父节点parent。如果last是叶子节点，③处就尝试用该叶子节点的所有数据点填满结果堆。如果该叶子节点填不满，则⑥处尝试用该叶子节点的兄弟节点填结果堆，直到填满为止。如果节点last 不是叶子节点，则使用fillQueueIfNonLeaf方法填满结果堆。此处要使用tmpNodeStack和tmpBitSetStack两个栈记录深度搜索路径上经过的节点

以及每个非叶节点的分支访问情况。当fillQueueIfNonLeaf方法返回之后，要将tmpNodeStack中的节点和tmpBitSetStack中BitSet对应压进节点栈和访问记录栈。这里要注意的是⑥处，节点栈要把最后一个节点先弹出，因为最后一个节点一定是叶节点，它没有对应的bitset。如果不将其弹出，之前的节点与对应bitset将会错1位。

```

private void tryToFillQueue(NNqueue, nodeStack, bitSetContext,
disCache) {
    /.../①判断结果堆是否已经填满
    Node parent, last; /.../ ②路径节点栈中获当前节点及其父节点
    if (last.isLeaf()) {
        /.../③尝试用叶节点的所有数据点填满结果堆
        int siblingSize = parent.getBounds().size() / 2;
        //④判断结果堆是否已经填满
        /.../⑤以兄弟节点数据填满结果堆
    }
    } else {
        /...../ 初始化 tmpBitSetStack, tmpNodeStack
        fillQueueIfNonLeaf(tmpBitSetStack, tmpNodeStack, ...);
        //尝试用非叶节点填满结果堆
        while (!tmpNodeStack.isEmpty()) {
            nodeStack.push(tmpNodeStack.removeLast());
            } //如果填满了，要将所经之节点都压入 nodestack
            nodeStack.pop(); //⑥
            while (!tmpBitSetStack.isEmpty()) {
                bitSetContext.bitSetStack.
                push(tmpBitSetStack.pop());
            }
        }
    }
}

```

图 4.10: 预填结果集代码

## 4.4 插入新轨迹功能模块详细设计与实现

### 4.4.1 Insert算法概述

在系统运行过程中，经常出现新的轨迹数据被插入已有索引的情况。这种情况下，如果每次都是将整个数据点集合重新进行初始建树，消耗会非常大。因此，有必要实现优先点树插入新数据点的相关算法。

对于新数据点插入操作，首先要进行深度搜索，找到正确位置的叶节点，然后对应以下几种情况分别采取不同措施。

第一，新插入的数据点所匹配的叶节点未满，则直接插入到正确的叶节点的空槽位中。如图??所示，新数据点e被插入叶节点L的空槽位中。这种情况下，只有数据赋值操作，没有距离计算，性能最好，之后几种情况的处理原则都是使后续插入尽可能满足第一种情况。注意，必须保证叶节点中数据点的顺序性，所有相关元信息的对应位置都要向后移动。

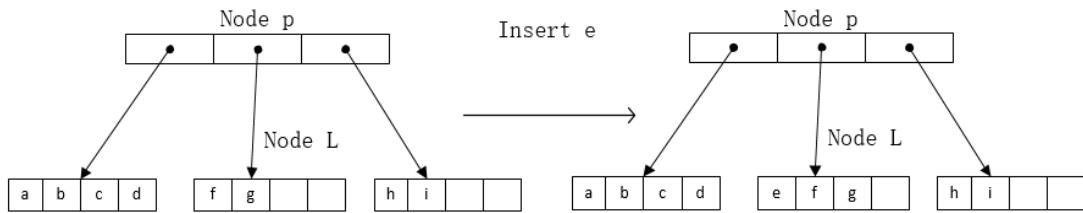


图 4.11: 直接插入对应叶节点算法示意图

第二，新插入的数据点所匹配的叶节点已满，但是其父节点分支数未满。则分裂匹配的叶节点，为父节点新增子叶节点，并且进行数据移动。注意，分裂之后的数据重分布应该秉持均分的原则，让空槽位尽可能均匀地分配在两个叶节点中，从而使得后续的插入操作尽可能多地符合叶节点不满的情况，避免发生连续多次分裂的情况。如图??所示，数据点e本应插入在节点L的f之前，但是因为节点L已经满了，而节点P的分支数未满，此时将节点L分裂成两个新节点L2和L3，将节点L中原本相距节点P的优先点距离最远的数据点h,i分配给L3。以此实现了数据点e的插入。同时保证L2,L3两个叶节点都有空槽位，以备后面的插入操作。

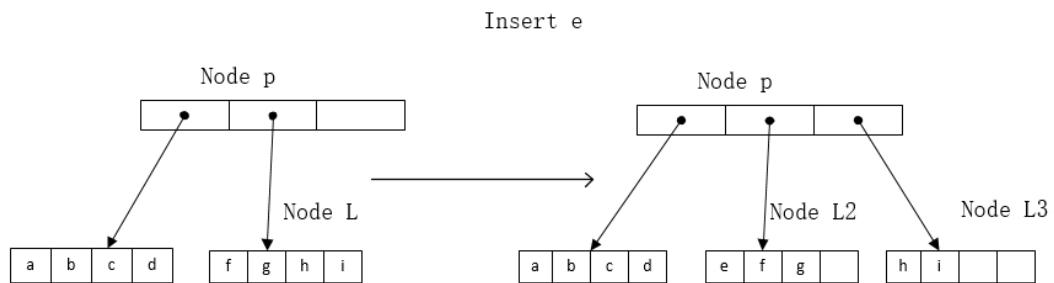


图 4.12: 叶节点分裂算法示意图

第三，新插入的数据点所匹配的叶节点已经满了，而且父节点的分支数也已经满了，不能再分裂出新的叶节点。这种情况下，采用中心扩散的方式寻找最临近未满的兄弟节点，进行数据重分布。注意，这里要考虑距离和空槽位数

量两种因素。空槽位数量越多的叶节点，重分布之后空槽位分布越均匀，越适合进行数据重分布。但是重分布不只是涉及目标叶节点和兄弟叶节点，而是涉及到两者之间的所有叶节点。两者之间的距离越远，需要重新分布的节点数就越多，消耗越大，反之，消耗越小。所以此处应该以距离最近为优先，以最小化指针移动和元数据更新的消耗。在距离相同的情况下，由叶节点空槽位的数量决定，空槽位越多的兄弟节点，越优先。如图??所示数据点以本应插入到节点L2中，但L2已满，则以中心扩散的方式，同时发现了两侧距离相同的未满节点L1和L3。此时发现L3中的空槽位比L1中多，因此选择L3进行数据重分布，移动数据h,i,j到L3中，然后完成e在L2中的插入。

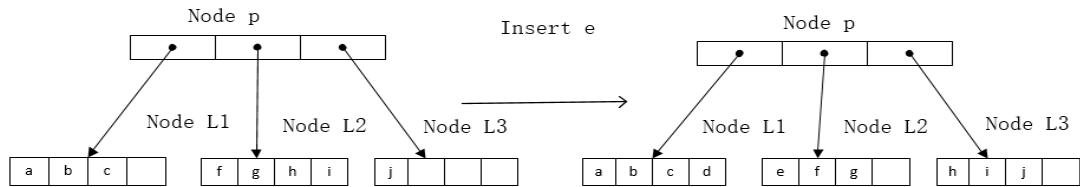


图 4.13: 叶节点数据重分布算法示意图

第四，新插入的数据点所匹配的叶节点，父节点以及所有的兄弟节点都已经满了，则向上回溯，找到第一个未满的祖先节点，如祖先节点的分支数未满，则优先进行目标分支的分裂，形成新的分支。注意，在这种情况下，不能像叶节点数据进行简单的移动就解决问题，由于产生新分支意味着必须重新建立子树，才能将原分支的数据点和新数据均匀分开，以降低接下来插入操作的负载。如图??所示，当插入数据点u的时候，分裂节点lowerAncestor，产生节点newBranch，并将均分的数据集合分别建立新的叶节点。而新子树的建立过程中必然需要选取新的优先点并且进行相应的距离计算。这种情况下的性能消耗，只比相同数据子空间的初始建树操作少了一轮优先点选取和距离计算而已。

第五，如果祖先分支数已经满了，则寻找到分支数据未满的子树的祖先节点，注意这里是分支数据未满，不是分支数未满。所谓分支数据未满，是指一个分支所能覆盖的最多的数据点的数量，这里也就包括了分支数不满和分支数满但是叶节点槽位不满两种情况。如果是后一种情况，分支之间的数据只需要进行数据重分布而并不涉及新的数结构的生成，与叶节点的数据重分布类似，只是数据节点和分支元信息的移动，但是多了一个深度搜索匹配叶节点的过程。如图?? 所示，当要插入节点 $\Psi$ 时，发现upperAncesor 的分支notFullBranch 的数据点数量不满，也就是说notFullBranch 的叶节点有空槽位。此时将lowerAncestor 和notFullBranch进行数据重分布，以相对于UpperAncestor的优先点的距离作为挪动数据点的依据，将那些lowerAnce

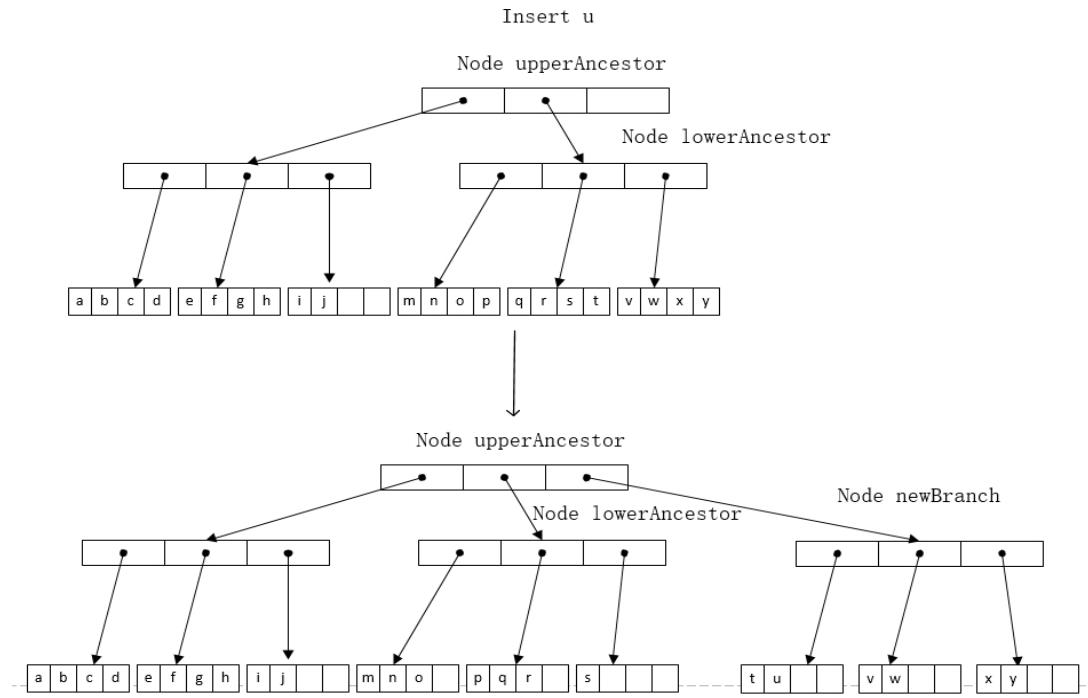


图 4.14: 分支分裂算法示意图

stor 中距离较大的 k 项移动到 notFullBranch 中，这里的 k 是两个分支节点总数的一半与 lowerAncestor 现有数据点量的差值。通过将 k 个数据点移动到 lowerAncestor，完成 insert 的同时，也尽可能地均匀分布了数据，为后续的 insert 留出空位，以尽可能地减少成本较高的数据重分布操作。而如果兄弟分支的子分支的分支数不满，那么就要进行类似于情况 2 的分支分裂操作，以实现目标数据点的插入。

第六，如果所有的位置都已经满了，则此时优先点树已经是一棵满多叉树。这种情况下，进行重新建树。尽管确实可以添加一个新的根节点，然后将旧根节点作为新根节点的一个分支，再以情况四的做法类似进行分支分裂。但这些操作的消耗等同于重新建树。本文出于简单实现，直接选择重新建树。

#### 4.4.2 Insert 算法设计细节

##### Insert 算法细节 1：以最小扇出度作为初始建树的分裂终止条件

这样做的目的在于推迟节点分裂的结束，使得更多的分支数目被创建出来。那么在初始建树之后，就会预留出较多的叶节点空位。当进行插入新数据点时，叶节点的空位置将会显著减少数据重新分布和节点分裂的次数，提高性能。但是也要注意的是，最小扇出度不能设的太小，因为节点元数据的存储都是以数组方式的线性存储，如果预留过多的空位，将会导致内存消耗大大增加。

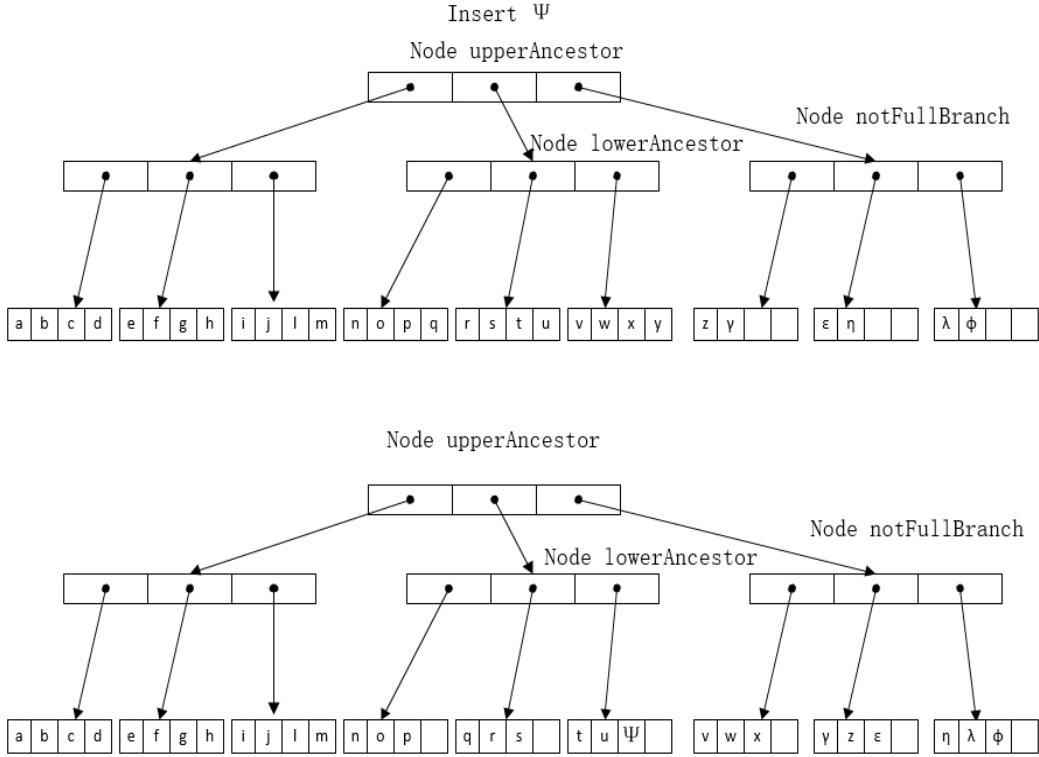


图 4.15: 分支重分布算法示意图

所以最小扇出度的设置要综合考量插入操作性能与内存容量而定。

### Insert算法细节2：分裂与数据重分布的优先级问题

节点分裂能够更好地创造更多的叶节点空槽位，对于连续的插入操作非常友好。而且对于分支数较多的VpTree，节点分裂将对节点元数据的更改降到2个分支以内，实际上可以做到以少量的距离计算代替大量的内存移动和比较操作，从而明显节省时间。

而数据重分布的好处在于，能够通过在节点元数据保存数据点集合的信息来实现简单的移动就能完成数据点重新平衡的效果，从而减少了新的距离计算的消耗。这在分支数少的VpTree中性能良好。但是对于那些分支数多，跨分支范围很大的VpTree，则可能造成巨量的比较和移动操作。比如一个10路的VpTree的，重分布的区间是从第二个分支到第八个分支，那就意味着这七个分支之间要进行数据移动。大量的比较和移动操作带来的消耗，同样可能带来性能灾难。而且如果分支的层级较高，重分布本身的距离计算也不会太少。

综上所述，节点分裂与重分布的性能消耗对比，与VpTree的分支数，分支距离，分支所在的层数都有关系，应该是一个动态计算的衡量结果。采用静态

设置的方式，很有可能造成优先级上的偏颇，但本文出于实现的简单，静态地采用了优先分裂的原则，这显然不是性能最佳的实现。

#### **Insert算法细节3：存储数据点距离顺序**

在初始建树的同时，在节点中顺序保存当前节点数据点集合相对于上一级优先点的距离排序结果。这样做是为了在数据重分布的时候，避免重新的距离计算和排序，可以通过简单的数组元素移动和界标更新来完成，从而明显减少计算量，提升性能。但是这样做的代价是，重复保存了大量的id列表，消耗了额外的内存，而且在重分布的过程中，必须将顺序集合的内容进行同步更改，也带来时间消耗。而在数据点较少，距离计算量少的情况下，数据点顺序列表的存储消耗和维护成本，可能比直接进行距离计算还要严重，因为每次插入都必须更新整条路径上的所有祖先节点的数据点顺序列表。因此，数据点顺序列表其实也应该是一个动态衡量的策略。

#### **Insert算法细节4：分支节点数据未满的判断**

与分支数相比，一个非叶节点所能包含的最大数据点的量与节点在整个优先点树中位置有关。位置越高，子树层数越多的节点，其所包含的数据点的量就越大。所以要判断一个非叶节点的数据点量是不是已经满了，要通过节点子树的高度，最大扇出度和叶节点最大的数据量进行计算。

本文通过在节点中保存父节点的指针，来访问父节点的高度，这样在初始建树的过程中，树高会从根节点的0一定递增下去，直到叶节点。而非叶节点最大的数据量通过如公式??进行计算。其中 $FANOUT$ 是VpTree的扇出度， $HEIGHT$ 是子树的高度， $ENTRYSIZE$ 是节点最大数据点量。当节点的数据量小于 $MAXDPCOUNT_e$ 时，就可以认定该节点数据量未满。

$$MAXDPCOUNT_e = (FANOUT)^{HEIGHT} \times (ENTRYSIZE + 1) - 1 \quad (4.1)$$

由以上几点细节可知，由于优先点树是基于不同优先点的轨迹距离索引，其各个数据点之间相对于优先点是偏序关系，而不存在整个数据集合上的全序关系。所以，不能像一般数据库的B-Tree 索引结构那样简单地比较索引的关键值就行了。对于分支分裂和子树重建，高成本的距离计算无法避免，导致其插入操作本身消耗很大，很难找到轻量级的办法来实现。而在较高位置的分支分裂，其计算消耗几乎可以等同于重新建树。

除此之外，为了保护检索性能，必须维持优先点树的天然平衡属性，不能简单地增长某棵子树的高度而破坏平衡，那样将会使搜索向一侧剧烈倾斜甚至在极端情况下等同于线性遍历，这将极大地破坏检索性能。

基于以上几点，本文所采用的分裂优先重分布，存储额外元数据，延长建树分支等策略，都是以空间换取时间的做法。而这些静态策略的实际运行效果和具体的数据集合的情况有关，所以在某些情况下可能并不是最佳的性能策略。

#### 4.4.3 Insert算法流程图

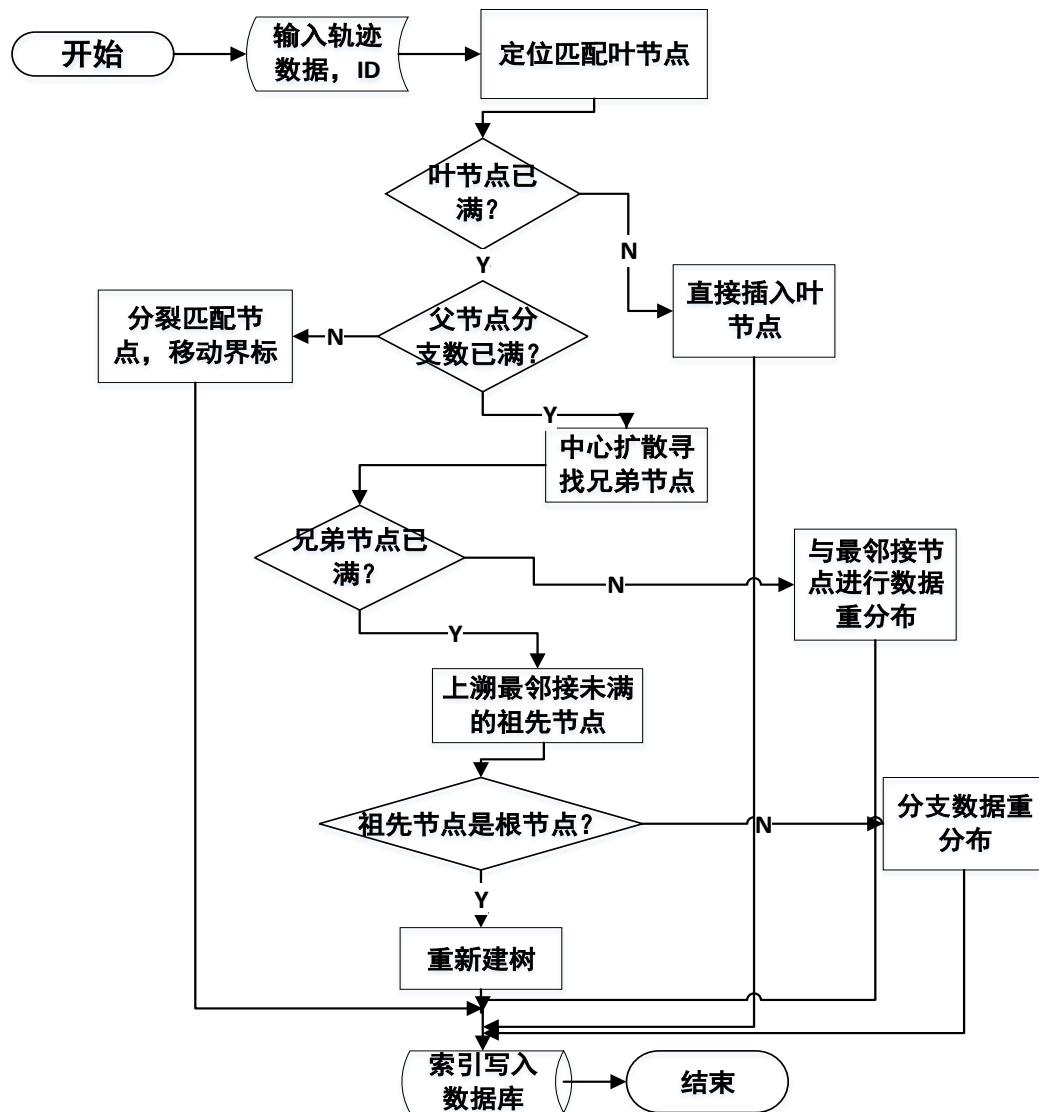


图 4.16: Insert算法流程图

#### 4.4.4 Insert算法实现

已满叶节点的分裂代码实现:如代码??所示, 已满叶节点分裂在获取正确的叶节点之后, 首先在①处更新了distances数组。随后将新轨迹数据id插入到childIds列表中, 此时childIds列表的长度比用户配置的叶节点尺寸多一个数据点, 紧接着在②处计算新的左右子节点的数据点数目, 即数据点总数除以2, 再做上取整, 这样使得数据点数左多右少。然后在③处初始化新的右侧叶子节点, 并将childIds中超出左叶子节点新尺寸的数据点id移除, 然后将这些数据点赋值到右叶子节点的childIds中。完成数据点转移之后, 在④处对parent节点的分支信息进行更新, 并且移动后续所有分支的槽位。parent节点的各种配置信息列表也要进行更新, 包括分支距离上下界列表branchbounds, 分支最大距离列表branchDistance, 子节点列表childNodes等, 最后在⑥处将新的右侧叶子节点加入到节点池中。

```
private void splitLeafNode(Node correctLeafNode, Node parent){
    /...../获取子节点列表
    /...../①更新 distances 数组
    childIds.insert(pos, id); //将 id 插入对应位置
    /...../②计算新的左右子节点的尺寸, 原则:左大右小
    /...../初始化新的右侧的叶节点
    childIds.removeRange(leftSize, childCount); //③清空分裂出去的位置

    /..../ ④移动 parent 的后续分支的槽位

    childrenNodeIds[branchPos + 1] = rightLeaf.getId();
    branchDistances[branchPos + 1] =
        rightLeaf.distances[rightSize - 1];
    branchBounds.insert(branchPos * 2, rightLeaf.distances[0]);
    branchBounds.insert(branchPos * 2 + 1,
        rightLeaf.distances[rightSize - 1]);
    //⑤更新 parent 节点的 branchBounds, branchDistance, childNodes
    nodePool.addNode(rightLeaf);
    //⑥将 rightLeaf 加入节点池中
}
```

图 4.17: 已满叶节点的分裂代码实现

**叶节点数据重分布代码实现:** 如代码??所示, ①处为记录最佳分支位置和拥有最大未满值的叶节点, 我们寻找重分布分支的依据以距离最近优先, 在距

离相同的情况下，以空槽位较多的分支为优，这里的largestDis记录的就是最大的未满值，也就是最多的空槽位值。②处left和right两个变量分别是分支搜索的位置坐标，由于是中心扩散搜索，因此是从当前分支的左右两侧的第一个位置开始，向两边扩散进行分支遍历。③处两个变量leftDPCount和rightDPCCount分别记录了向左数据重分布涉及的数据点数量和向右涉及的数据点数量。④处为中心扩散的循环中，持续跟踪更新leftDPCount和rightDPCCount，当找到bestBranch的时候，基于距离优先原则，尽快跳出循环。⑤处是计算重分布后，所有涉及到的分支，应该持有的平均数据点量，注意，此处展示了向左数据重分布的伪代码，实际还有向右的对称情况，并且平均数量要有取整操作。基于⑤处计算出来数据点量，在⑥处依次挪动所有涉及分支中包含的多于dpCount的数据点，实现数据重分布。在⑦处，更新parent节点对应的分支元信息，包括界标数组和距离数组。

```

private void redistributeLeafNode(correctLeafNode, parent) {
    int bestBranch = -1;
    int largestDis = -1; //①记录最大未满值和最佳目标重分布分支

    int left = branchPos - 1;
    int right = branchPos + 1; //②搜索起始位置，中心点两侧

    int leftDPCount = 0;
    int rightDPCount = 0; //③记录左右重分布区间的数据点数量

    while( right < branchCount || left >= 0){
        /...../ ④跟踪更新 leftDPCount 和 rightDPCCount
        if(bestBranch != -1) break; //尽快结束搜索
        left--; right++; //左右指针向两侧移动
    }

    if(bestBranch != -1){
        int redisBranchCount = branchPos - bestBranch + 1; // ⑤
        int dpCount = (leftDPCount * 1.0) / redisBranchCount;
        /...../ ⑥最好分支在左侧，各分支数据点依次向左移动，元数据的更新
        /...../ 最好分支在右侧，分支数据点依次向右移动，元数据的更新
        /...../ ⑦更新 parent 节点的界标数组
    }
}
}

```

图 4.18: 叶节点数据重分布代码实现

分支分裂代码实现：如代码??所示，在①处计算左右分支的数据点量，

```

private void splitNonLeafNode(Node ancestorUpper, Node
ancestorLower, IntArrayList dpList){
    /.../①计算分裂后左右分支数据点量
    IntArrayList rightList = new IntArrayList(rightSize);
    for(int i = leftSize; i < dpCount; i++){
        rightList.add(dpList.get(i));
    }
    dpList.removeRange(leftSize, dpCount - 1);
    //②更改左右分支的数据点集合

    Node leftNode = new Node(nextNodeId(), false);
    leftNode.setParent(ancestorUpper);
    leftNode.setNodeHeight(ancestorUpper.getNodeHeight() + 1);
    makeVpTree(leftNode, dpList);
    //③设置左节点，并以左节点为根节点，分裂新的子树

    /...../④针对 rightNode 执行与 leftNode 相同的操作，代码同③

    int pos = locateChildPos(ancestorUpper, ancestorLower);
    //⑤定位下方祖先节点的分支位置

    shiftBranchInfo(ancestorUpper, pos + 1, 1);
    //⑥将后续分支的配置信息向后挪动
    /.... /更新上方祖先节点的元数据
}

```

图 4.19: 分支分裂代码实现

与前文类似。②处是将左右分支的数据点重新分配，也就是把左分支，也就是把旧分支多出来的数据点挪给右分支，也就是新分支。③处为设置左节点，并以左节点为根节点进行子树创建的过程，这里也是分支分裂与叶节点分裂最大的不同，就是必须对分支所拥有的数据点子集合进行重新建树，也就是说，针对子集合的选举优先点和基于优先点的距离计算要重新进行，因此③处的makeVpTree方法其实是createNode方法的封装实现，其目的都是以某一节点为根节点建立子树。④处为设置右节点的代码，与设置左节点代码类似。⑤处为确定当前分支节点与祖先分支节点的相对位置关系，因为在新建一个新的分支之后，必须要把原来分支后面的所有分支的配置信息向后移动，如⑥处shiftBranchInfo 方法。而对应的祖先分支节点的其他元数据，包括分支上下界数组和距离数组都要对应更新和移动。

## 4.5 轨迹数据子服务运行效果展示

如图??所示，图中的地图背景为中国东海岸，两个红色群集和以浅蓝色群集构成了3个船舶轨迹的数据集合。右侧框中显示了相似检索结果，轨迹的ID是数据采集的日期。

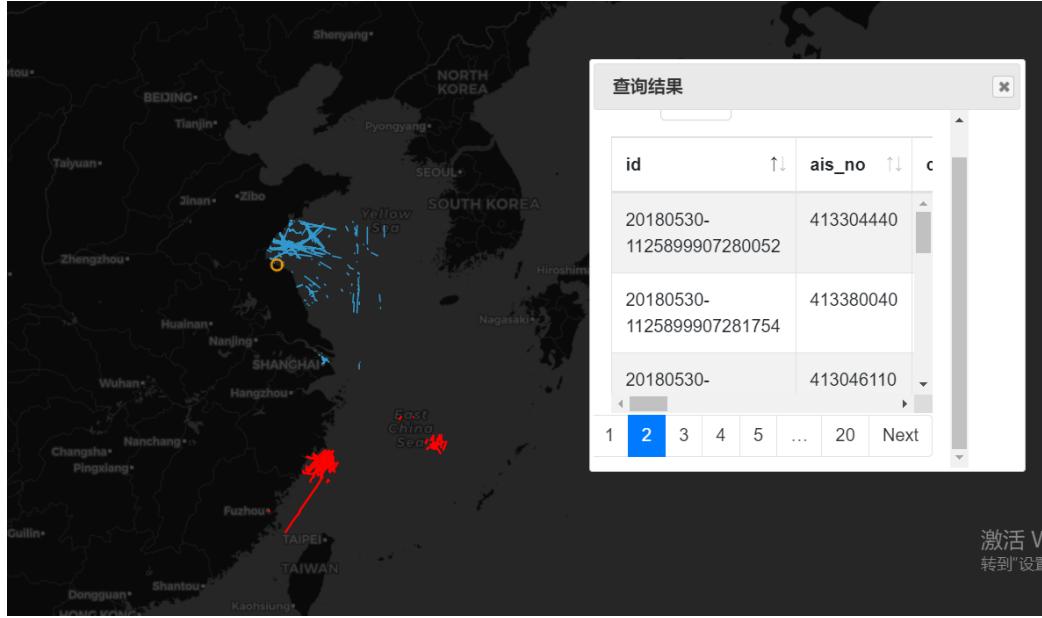


图 4.20: 轨迹检索运行效果

## 4.6 轨迹数据子服务Rest调用示例

```
curl -XPOST -s "localhost:9201/${index}/_search?pretty&size=${size}" -d '
{
  "query":{
    "trajectory_nearest_neighbour_query": {
      "size": 1000,
      "targetID":201804231025899907280022,
      "index":train_set
    }
  }
}'
```

图 4.21: 相似轨迹检索Rest调用示例

如图??所示为相似轨迹检索功能的HTTP curl命令的Rest调用示例，图中的ip是本文实验所用IP， trajectory\_nearest\_neighbour\_query是对应的LuceneQuery名称，其参数中有目标结果集合尺寸size、目标轨迹ID和目标索引id。轨迹数据

服务的另两个功能与此接口的调用方式相似，只是参数略有不同，此处不再赘述。

## 4.7 本章小结

本章描述了轨迹数据子服务的详细设计和实现，主要介绍了VpTree初始建树，相似轨迹检索以及新轨迹数据插入这三个功能模块的设计要点和实现原理。通过流程图，类图来说明设计的思路，然后通过具体代码展示关键的实现细节。最后展示了轨迹数据子服务的相似轨迹检索功能的运行效果图。

## 第五章 地图瓦片数据子服务详细设计与实现

### 5.1 服务架构概述

地图瓦片数据子服务的功能是为整个系统提供可视化地图数据。它是基于Nodejs Express框架实现的，符合MapBox数据标准的数据微服务。**MapBox**是业内通用标准，符合**Mapbox** 标准是为了确保通用性。它的所有数据服务都通过Rest接口暴露给调用方。

如图 ?? 所示，地图瓦片服务主要分为data、style、cache、logic和db这5个模块，以及对接db的各种数据库驱动组成，其各部分功能如下。

data\_service模块是直接决定数据操作和数据状态的模块，主要负责提供瓦片数据的读取，更新功能。这一模块是整个服务的核心逻辑，它依赖logic模块进行地理坐标计算，依赖db 模块对接各种外部瓦片数据库，依赖cache模块进行缓存。data模块通过Express Route接口直接对外提供Rest访问服务。

style\_service模块是提供风格数据的模块，负责向调用方提供指定的风格数据，用于页面渲染。style是地理瓦片服务的特有标准，其作用类似于CSS，是决定地图数据以何种样式渲染在浏览器上。style模块作用于MapBox的style文件，直接与文件系统交互，并不涉及数据库操作，本质上只是静态文件的读取。

logic\_service是支持与地理相关的运算逻辑模块，例如坐标转换，bounding-box换算等。这一部分的实现是无状态的，与data模块完全解耦，可以通过插件的形式，随时增加或者变换逻辑计算的功能。

cache\_service是与瓦片数据缓存相关的功能模块，可以为瓦片数据配置不同的缓存策略。与一般的缓存系统不同的是，瓦片数据访问的局部性除了有水平局部之外，还存在垂直局部，也就是用户在连续滑动鼠标滚轮时，同一个水平区域的不同zoom级别的瓦片会被连续加载。正是基于这种缓存策略可变的情况，cache模块也被设置为无状态，可替换的模块。目前的实现是对水平局部瓦片的缓存。

db\_service是面向各种不同数据库的通用访问接口层，其是data\_service与其他各种地图瓦片数据库交互的媒介。db\_service的接口是固定的，所有要对接到瓦片数据服务上的数据库驱动必须实现db\_service要求的接口，才能正常运行。目前实现的数据库驱动有sqlite和hbase两种。

地图瓦片服务本身接口较多，本文只挑选其中比较复杂的地图局部更新功

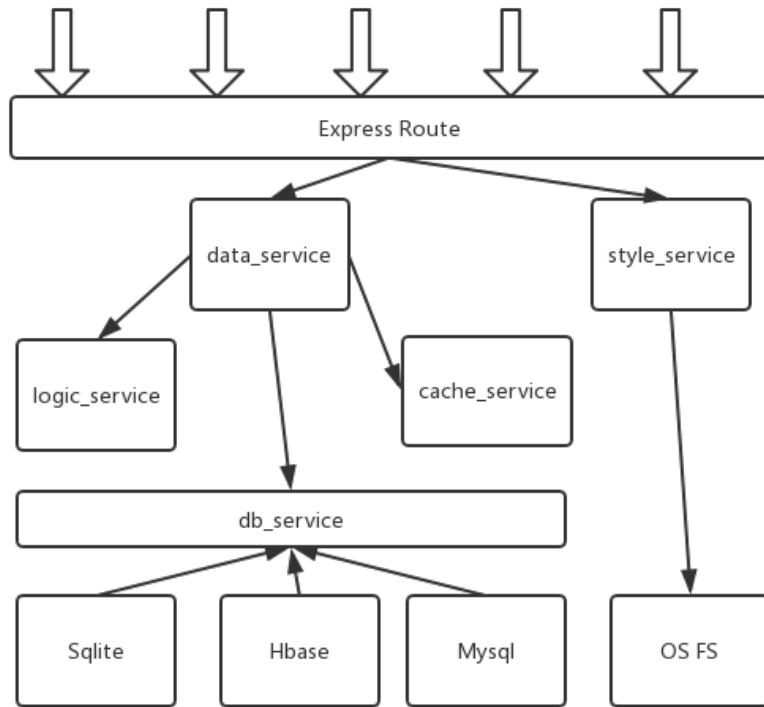


图 5.1: tile-server整体结构图

能，介绍详细设计和实现。局部更新功能也是其他开源地图瓦片服务所不能提供，本文所独立设计和实现的功能。

## 5.2 地图局部更新功能概述

所谓地图局部更新，指的是一张完整地图，只更新全部地图的一小部分。比如中国地图中，只更新北京市地图。这一功能的价值很直观，在MTS的很多商用场景中都有使用价值。本文根据Web莫卡托投影法的计算原理，实现了这一功能。

由于用户在浏览器界面上框选出来的，一定是经纬度坐标的范围。因此，该功能的输入参数应该是一个north,south,west,east四个数值组成的矩形框和对应的zoom，也就是bounding-box。所以，要想更新对应的瓦片数据，就必须首先要根据像素精度和显示比例尺，将经纬度坐标转换为瓦片坐标，也就是对应级别的x,y，再将对应地图瓦片数据替换掉。这一过程中，除了坐标转换之外，为了避免歧义和出现地图更新空隙，还需要做边界、精度和一致性的检查。

## 5.3 地图局部更新功能详细设计与实现

### 地图局部更新功能流程图

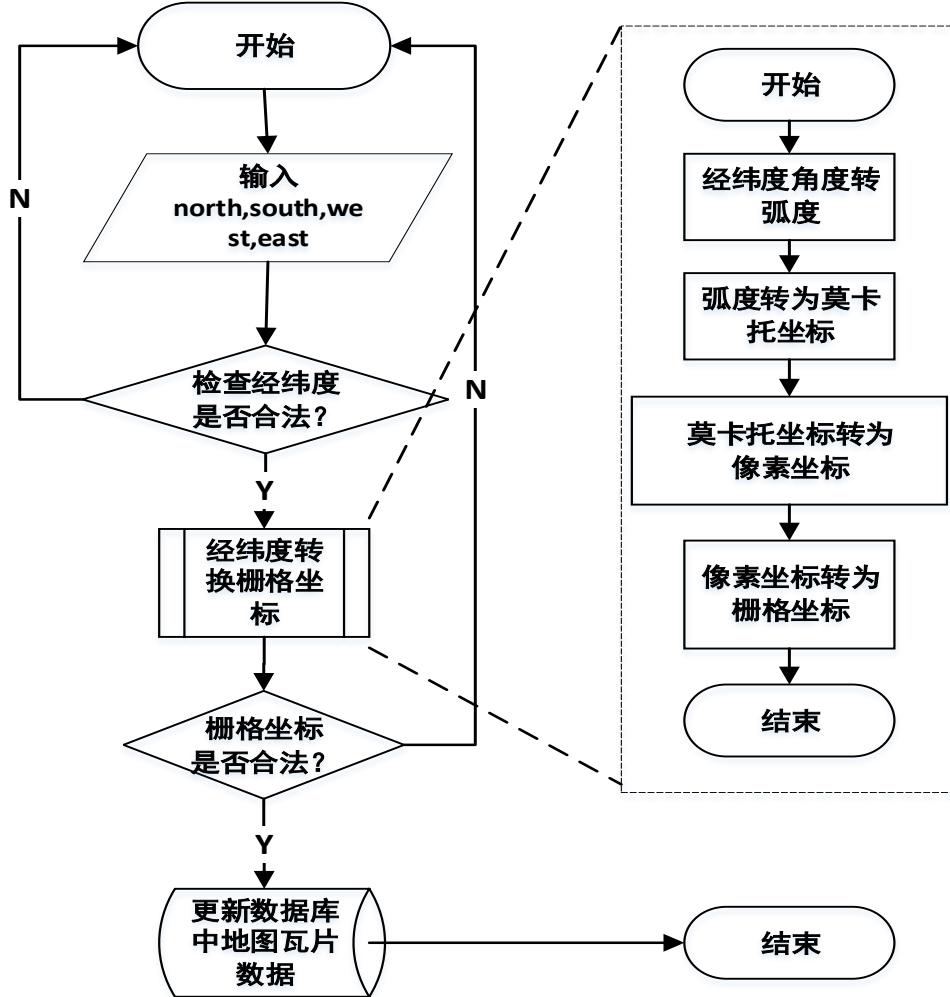


图 5.2: 地图局部更新功能流程图

如图??所示，这里值得注意的是，在流程中第二次检查，是对栅格坐标结果的检查。因为对于不同zoom，zoom值越大，地图的精度越高，更新涉及的完整瓦片越多，反之则越少。那么对于那些zoom值较低的更新，其所计算出来的栅格坐标值可能不是完整的的瓦片，而只是某一瓦片中一个位置，由于本文对瓦片更新的粒度不涉及到单个瓦片内部的结构，而只是单个瓦片的整体更新。因此对于这种情况，要根据用户对精度的要求进行取舍，而取舍的前提要保证更新行为的一致性。具体可见后文。

**地图局部更新功能代码实现：**如代码??中所示的degree2xy方法，是局部更新功能的核心逻辑，其作用是在一个由north,south,west,east四个经纬度值划定的矩形框内，计算某一个经纬度点的瓦片坐标。如代码中注释可知，该degree2xy方法的主要逻辑是对Web莫卡托坐标转换原理的实现。另外，

```

function convert2radians(deg){
    return deg*Math.PI/180;
}
//角度转换为弧度的函数
function mercator2y(lat) {
    y=Math.log(Math.tan(lat)+(1.0)/Math.cos(lat));
    return y;
}
//纬度的莫卡托转换的函数

function degree2xy(lat,lon,north,south,west,east,size,zoom)
{
    lat=convert2radians(lat);
    lon=convert2radians(lon); //将经纬度转为弧度

    north=convert2radians(north);
    south=convert2radians(south);
    west=convert2radians(west);
    east=convert2radians(east); //将经纬度转为弧度

    let yMin=mercator2y(south);
    let yMax=mercator2y(north); //纬度边界做莫卡托转换

    let y=mercator2y(lat); //目标纬度做莫卡托转换

    let xfactor=size/(east-west);
    let yfactor=size/(yMax-yMin); //计算单位经纬度的像素值

    let x=(lon-west)*xfactor;
    y=(yMax-y)*yfactor; //计算目标坐标的像素值

    x=x/256;
    y=y/256; //得出栅格坐标值

    return {"x":x,"y":y}
}

```

图 5.3: 地图局部更新功能核心代码实现

mercator2y方法是将纬度值转换为莫卡托纬度，convert2radians是将经纬度角度值转为弧度值，这些是地理坐标转换的数学必备部分。

## 5.4 局部更新一致性方法

所谓的局部更新一致性，指的是确保整体更新和各个部分分别更新的结果是一致的。之所以可能出现不一致的现象，主要是因为单纯的瓦片坐标模糊取整很可能遗漏部分瓦片，导致某些瓦片由于取整被舍弃，而没有被任何一个部分更新，进而导致更新空隙的问题。对于这个问题的解决办法，基于以下原则进行处理。

1.首先判断用户bbox更新中设置的经纬度的差值在其所设定的zoom下，是否大于等于一个Tile的边长。如果用户划定的范围太小，要求的精度太高，则直

接告知用户，在当前zoom下不能更新这个bbox。如图??所示，用户选定的更新范围s1的宽度小于Tile1的边长，而更新范围s2的长和宽都小于Tile2的边长。这两种情况下，都不对Tile进行更新。

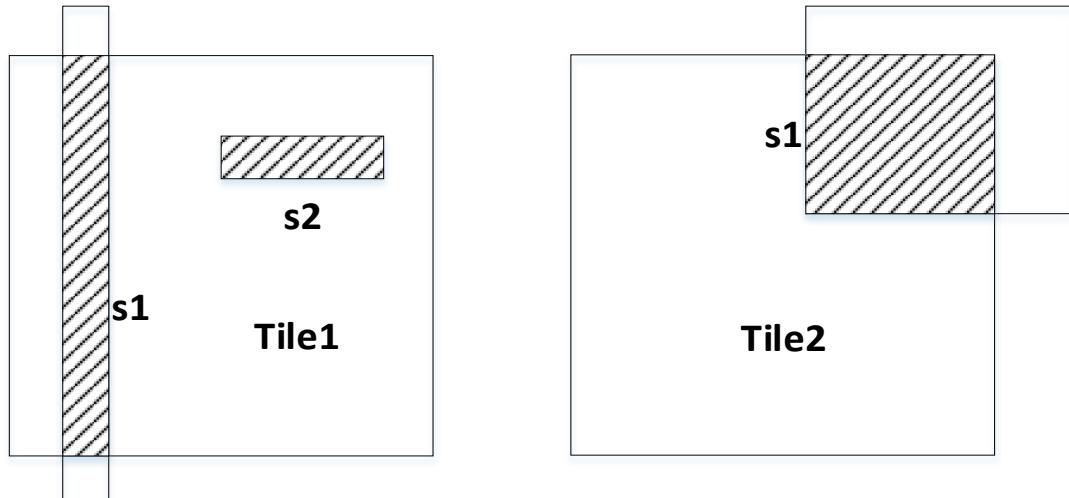


图 5.4: 瓦片更新前提情况

2.如果bbox的长度和宽度都大于等于Tile的边长。那么在这种情况下，对于一个Tile的bbox划分。如果划分本身是严密的，没有缝隙的。那么，1个Tile能且只能被划分为4个bbox.而1个Tile的4个bbox划分中，至少有一个，其面积超过了1/4。以此作为是否更新的标准，能够保证每一个Tlle至少被更新了一次，从而避免了局部更新漏掉某些Tile产生更新缝隙的问题。

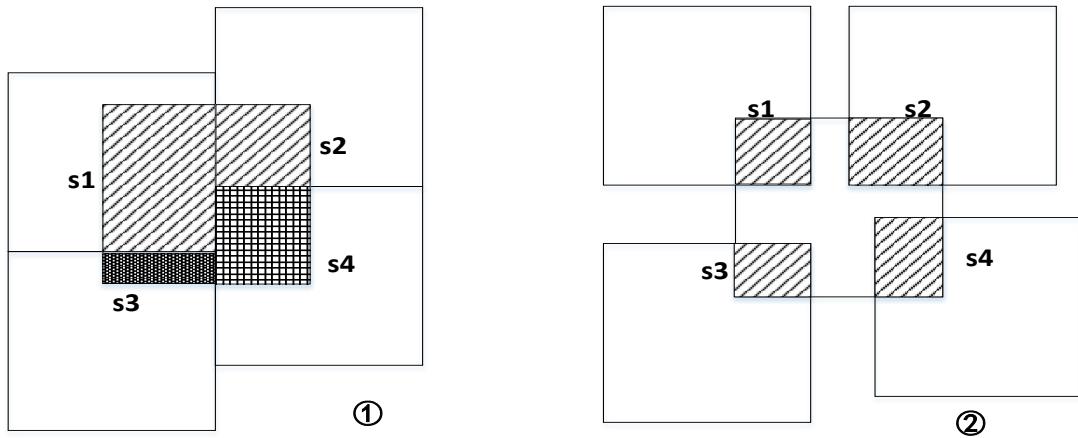


图 5.5: 瓦片更新条件

如图??所示，图①所示，s1-s4，4块划分中，s1的面积显然大于总面积的1/4, 此时触发瓦片更新。而图②所示，s1-s4四块划分的面积都不足1/4，造成

这种局面的原因是用户选定的bbox范围本身就存在间隙，没有填满整个Tile，这种情况下空隙是用户逻辑不完备造成的，瓦片数据服务不予更新。

综上所述，本文实现的瓦片数据服务更新一致性的功能是瓦片级别的弱一致性，其依据的是用户操作本身bbox划分的逻辑一致性。

```

function whether_to_cover(accurate_xy_pixel,center_xy_pixel,){
    if(accurate_xy_pixel.x<center_xy_pixel.x&&
        accurate_xy_pixel.y<center_xy_pixel.y){
        /.../中心点在右下角，精确坐标向右下取整
    }
    else if(accurate_xy_pixel.x<center_xy_pixel.x&&
        accurate_xy_pixel.y>=center_xy_pixel.y){
        /...../中心点在右上角，精确坐标向右上取整
    }
    else if(accurate_xy_pixel.x>=center_xy_pixel.x&&
        accurate_xy_pixel.y<center_xy_pixel.y){
        /...../中心点在左下角，精确坐标向左下取整
    }
    else{
        /...../中心点在左上角，精确坐标向左上取整
    }

    let round_area=Math.abs(round_x-accurate_xy_pixel.x)*
        Math.abs(round_y-accurate_xy_pixel.y);
    //计算模糊区域的面积，单位为像素平方

    let tile_area=65536; //65536是固定的瓦片面积像素平方数

    if(area>=(tile_area/4)) return true;
    return false;
}

```

图 5.6: 判断是否覆盖当前瓦片的代码

如代码??所示，计算重合部分面积，首先要确定实际精确坐标点与部分重合瓦片中心坐标点的相对位置关系。根据左上，右下，左下，右上四个不同方向结果，对精确坐标采用相同方向的取整操作。然后根据精确坐标与模糊坐标的横纵差值计算重合部分的面积。最后比较重合坐标面积是否超过了65536像素平方的四分之一，也就是一个瓦片面积的四分之一，来判定是否要覆盖该瓦片。

## 5.5 地图瓦片数据子服务Rest调用示例

如图??所示为地图瓦片局部更新的HTTP curl命令Rest接口调用示例。localhost:9202是本文测试所用的IP端口，图中①可见POST请求的参数包括

```
curl localhost:9202/update_bbox_tile -X POST -d "north=36.39 & south=37.61 & west=115.6 & east=117.2 & file=/tmp/osm-2017-07-03-v3.6.1-asia.china.mbtiles" //①

curl localhost:9202/upload_tile_file -F "file=@/tmp/osm-2017-07-03-v3.6.1-asia.china.mbtiles" //②

curl localhost:9202/data/china.json //③
```

图 5.7: 地图瓦片数据子服务Rest调用示例

了north,south,west,east 四个bbox的范围参数, file参数指定的是服务端文件路径, 这就要求首先要通过②命令上传地图瓦片数据文件。图中③为读取单个瓦片数据集合的Rest接口, china.json是数据集合的json描述文件, 该文件符合Mapbox标准。

地图瓦片服务的其他Rest接口类似, 此处不再赘述。

## 5.6 本章小结

本章描述了地图瓦片数据服务的详细设计和实现, 主要介绍了地图瓦片数据服务的整体结构, 并针对地图局部更新功能进行了详细设计说明。通过流程图和代码实现来说明设计和实现的思路, 最后对局部更新可能出现的不一致问题给出了解决办法和原理以及实现代码。

## 第六章 系统测试

### 6.1 测试环境准备

表 6.1: 系统测试环境准备

设备名称	运行程序	程序版本信息
用户测试计算机(Windows 10)	Firefox浏览器	Firefox 66.02(32位)
DELL服务器1 (CentOS 7.3)	Nodejs 6	Nodejs 6.4.1
DELL服务器2 (CentOS 7.3)	Nodejs 6,JVM 8	Nodejs 6.4.1,JVM 1.8.0_151
DELL服务器3 (CentOS 7.3)	JVM 8,Hbase	JVM 1.8.0_151,HBase 2.0.5
DELL服务器4 (CentOS 7.3)	JVM 8,ElasticSearch	JVM 1.8.0_151,ElasticSearch 6.5

本小节将根据概要设计中所描述的物理视图，介绍服务的测试环境。如表格 ??所示，4台DELL服务器分别运行了可视化子服务，地图瓦片子服务，地图瓦片数据库和轨迹数据子服务。这里需要说明的是，本文测试时使用了Hbase存储瓦片数据，所以要在服务器3上安装HBase，并在服务器2上安装JVM以运行Hbase Client。此外，本文对Lucene索引结构的扩展最终被集成到ElasticSearch 6.5中运行，由于整个服务的设计和实现都没有涉及到ElasticSearch，所以前文予以忽略。

表格 ??中的DELL服务器的硬件规格均为32核，128G内存，3\*1T SATA硬盘。安装了JVM的机器，JVM堆内存大小为32G。

### 6.2 测试数据准备

本文用于测试的地图瓦片数据选择了OpenMapTiles的开源数据osm-2017-07-03-v3.6.1-asia.china.mbtiles，size约等于3.2G。

本文用于测试的轨迹数据选择了从12306爬取的8000余条真实火车轨迹，并对每一条火车轨迹进行了站台随机模糊，衍生出约4千万条类似轨迹，示例数据如表格 ??所示。这里需要说明的是，对于轨迹数据测试，暂时只考虑了LineString类型，Polygon类型的数据暂时缺失。而前文所谓的站台随机模糊，指的是将站台列表中的任意站台，替换为10公里范围内的其他任意随机站台，从而产生了大量的相似路线，这就是站台随机模糊。

表 6.2: 测试用轨迹数据示例

日期(date)	车次(name)	类型(type)	始发站(from_station)	终点站(to_station)	路线(path)	站台(stations)
2019-02-02	T297	T	北京	牡丹江	LineString	[北京,唐山北.....,海林,牡丹江]

### 6.3 单元测试

单元测试指的是对软件最小可测单元进行的功能检查和验证。本系统中，单元测试使用ElasticSearch自带的ESTtestCase模板针对比较重要的Lucene轨迹索引类进行测试。

表格 ??为相似轨迹检索接口测试表，本文采用列表比对的方式确认轨迹结果集合与答案集合的吻合，测试结果符合预期。

表 6.3: 相似轨迹检索接口测试

测试用例编号	TC-01
测试目标	验证相似轨迹检索结果的正确性
接口函数	VpTree中的nearestNeighbourQuery方法
前置条件	轨迹检索子服务接收到目标轨迹的ID和相似数量K
输入	目标轨迹的ID和相似数量K
预期输出	与目标轨迹最相似的K条轨迹的ID列表

表格 ??为轨迹索引初始化接口测试表，在该测试中通过广度遍历结果优先点树节点的方式，确认生成的轨迹索引符合预期。

表 6.4: 轨迹索引初始化接口测试

测试用例编号	TC-02
测试目标	验证所建立的轨迹索引的完整性和逻辑正确性
接口函数	VpTree中的createNode方法
前置条件	轨迹检索子服务接收到批量轨迹数据
输入	批量轨迹数据
预期输出	逻辑正确的轨迹索引

表格 ??为插入新轨迹接口测试表，该测试同样采用广度遍历索引节点的方式进行检查，测试结果符合预期。

表格 ??为地理坐标转换接口测试，本文采用结果坐标集合数字比对的方式进行检查，测试结果符合预期。

表 6.5: 插入新轨迹接口测试

测试用例编号	TC-03
测试目标	验证插入新轨迹数据之后，轨迹索引的正确性
接口函数	VpTree中的insert方法
前置条件	轨迹检索子服务接收到新的轨迹数据
输入	新的轨迹数据
预期输出	更新后的，逻辑正确的轨迹索引

表 6.6: 地理坐标转换接口测试

测试用例编号	TC-04
测试目标	验证地理坐标转换与目标区域的实际坐标相一致
接口函数	bounding_util的degree2xy方法
前置条件	地图瓦片子服务接收到目标经纬度坐标范围
输入	经纬度坐标范围
预期输出	正确的瓦片坐标范围

## 6.4 功能测试

本文针对地理轨迹数据服务进行了功能测试。注意，由于本文所设计和实现的子服务内部高度内聚，而子服务之间采用松耦合REST风格进行接口调用，因此功能测试与集成测试的方法和结果类似，就不再赘述集成测试。

表 6.7: 相似轨迹检索功能测试

测试用例编号	TC-RS-01
测试目标	验证相似轨迹检索功能结果和展示效果的正确性
前置条件	用户进入操作界面
测试步骤	1. 用户输入要检索的目标ID。(ID不存在，报错) 2. 用户选定要检索的相似轨迹的数量。 3. 用户点击确认。
预期结果	检索结果正确返回并显示在界面上。

本文对核心功能相似轨迹检索进行了功能测试，如表格 ??所示，在界面操作中，前端拦截了诸如错误文件格式，错误ID格式，目标数量过大等问题。而针对目标ID不存在，是由服务器验证并返回警告用户的。最终测试结果和预期相符。

## 6.5 性能测试

本文测试了单个节点下，相似轨迹检索的时间消耗和轨迹索引常驻内存大小。

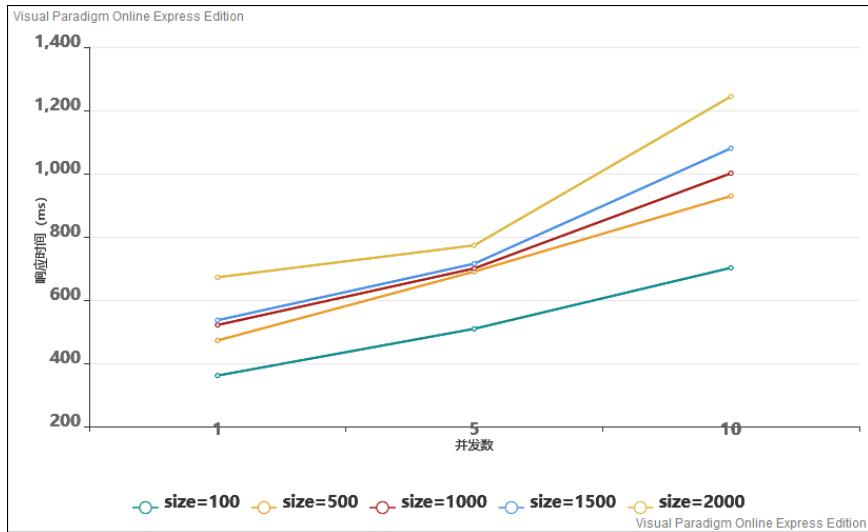


图 6.1: 相似轨迹查询性能测试折线图

相似轨迹检索在四千万测试集合下，不同并发数的响应时间如图 ?? 所示。横轴表示并发数，纵轴表示相似查询的响应时间(单位是ms)。不同折线表示size 取不同值的结果。

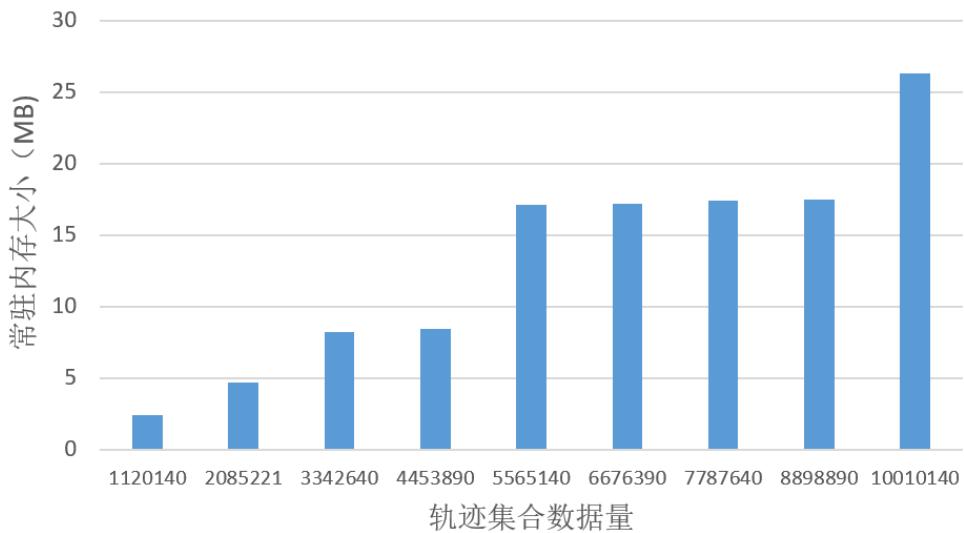


图 6.2: 轨迹索引常驻内存统计

如图 ?? 所示为轨迹相关索引常驻内存统计，横轴表示轨迹数据集合的数

量，纵轴表示常驻内存大小，单位为MB。

轨迹数据索引的建立以批量数据导入的方式进行，在四千万条数据的测试集合下，索引建立耗时约为30min。

## 6.6 本章小结

本章针对地理轨迹数据服务的重要功能进行了单元测试，功能测试和性能测试，描述了测试环境，测试数据，并使用测试用例说明测试步骤和方法，使用折线图和条状图直观地表明测试结果。

## 第七章 总结和展望

### 7.1 总结

随着移动互联网和地理定位技术的发展成熟，运动实体产生的地理空间轨迹数据被大量地记录下来。在海量的轨迹数据中，蕴含着丰富的人类社会活动规律。而在这些规律中，轨迹的相似性，是揭示实体运动关联性的重要指标。为了满足用户对大数据量地理轨迹的相似检索和可视化需求，本文设计和开发了地理轨迹数据服务。

本文的前两章节分别介绍了项目背景和所涉及到的技术标准以及每一项技术在系统实现中的作用，包括使用优先点树作为轨迹索引结构，以全文检索库lucene作为检索基础库，用地图瓦片作为可视化服务的基础单元，为了实现通用性而满足Mapbox 技术标准以及使用Nodejs Express 框架作为服务基础框架等。

本文第三章通过概述系统整体结构，对最重要的两个功能子模块，轨迹数据服务和地图瓦片数据服务，分别通过用例图和用例描述的方式进行需求分析，对需求的正常流程和可能出现问题的异常流程进行了详细说明。最后使用软件架构4+1模型，从逻辑蓝图，过程蓝图，开发蓝图，物理蓝图和场景脚本等多个角度对系统整体架构进行了设计。

第四章和第五章分别针对轨迹数据服务和地图瓦片数据服务，这两个主要功能子模块进行了详细设计与实现的阐述。使用类图，算法流程图和原理示意图说明设计思路和问题解决办法。针对每一个设计焦点进行了详细讲解，用具体的代码展示和说明来阐述实现细节。并在这一过程中，对每一重点设计决策的意义和局限性进行了分析，提出了更好的解决方案。除此之外，还为重要的功能展示了服务运行界面效果和Rest接口调用示例。

最后，本文的第六章介绍了系统的测试情况，包括测试数据的来源和测试的软硬件环境。然后以测试用例描述表的方式，阐述了单元测试和功能测试的测试过程和结果。最后使用折线图和条形图给出了系统性能测试结果。

### 7.2 进一步工作展望

地理轨迹数据服务还有进一步的发展空间。从索引结构上看，目前实现的版本是单个优先点的多路树，这种结构的剪枝性能在数学上被证明不如另一

种多优先点多路树好，未来可以考虑更改VpTree为Mvp-Tree，从而进一步提升检索性能。

当前版本对于轨迹相似检索考虑的标准是几何距离，而没有考虑轨迹数据点的次序和方向性，这使得轨迹本身的属性有一定的损失，造成检索精度的下降。未来可以考虑加入对方向和次序的考量，从而提高检索精度。

当前实现中，轨迹索引初始建立时间较长，与优先点选取算法过于繁琐，计算量太大有关。这也是可以优化的地方。除此之外，轨迹索引常驻内存也较大，这与距离缓存过多而缺乏过期策略有关，未来可以通过更好的数据缓存管理来减少常驻内存的使用。

对于地图瓦片数据服务，当前版本的瓦片更新的粒度最小也是一个瓦片，这个粒度太粗，未来应该提升到对瓦片内部结构的更新功能。

综上所述，地理轨迹数据服务还有诸多可以改进的地方，这些也是未来优化和发展的方向。

## 参考文献

- [1] BaiduYingyan, <http://lbsyun.baidu.com/trace> (2018).
- [2] TileServerGL, tileserver System, <http://tileserver.org>.
- [3] D. Mateos-García, J. García-Gutiérrez, J. C. R. Santos, On the evolutionary weighting of neighbours and features in the k-nearest neighbour rule, Neurocomputing 326-327 (2019) 54–60.  
URL <https://doi.org/10.1016/j.neucom.2016.08.159>
- [4] P. N. Yianilos, Data structures and algorithms for nearest neighbor search in general metric spaces (1993) 311–321.  
URL <http://dl.acm.org/citation.cfm?id=313559.313789>
- [5] K. Fukunaga, P. M. Narendra, A branch and bound algorithms for computing k-nearest neighbors, IEEE Trans. Computers 24 (7) (1975) 750–753.  
URL <https://doi.org/10.1109/T-C.1975.224297>
- [6] D. A. Forsyth, P. H. S. Torr, A. Zisserman (Eds.), Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II, Vol. 5303 of Lecture Notes in Computer Science, Springer, 2008.  
URL <https://doi.org/10.1007/978-3-540-88688-4>
- [7] L. Hirsch, T. Brunsdon, A comparison of lucene search queries evolved as text classifiers, Applied Artificial Intelligence 32 (7-8) (2018) 768–784.  
URL <https://doi.org/10.1080/08839514.2018.1506972>
- [8] P. Yang, H. Fang, J. Lin, Anserini: Reproducible ranking baselines using lucene, J. Data and Information Quality 10 (4) (2018) 16:1–16:20.  
URL <https://doi.org/10.1145/3239571>
- [9] M. McCandless, E. Hatcher, O. Gospodnetic, Lucene in Action, Second Edition: Covers Apache Lucene 3.0, Manning Publications, 2010.  
URL <http://www.springer.com/computer/swe/book/978-3-642-12577-5?changeHeader>

- [10] V. Antoniou, J. Morley, M. M. Haklay, Tiled vectors: A method for vector transmission over the web, in: Web and Wireless Geographical Information Systems, 9th International Symposium, W2GIS 2009, Maynooth, Ireland, December 7-8, 2009. Proceedings, 2009, pp. 56–71.  
URL [https://doi.org/10.1007/978-3-642-10601-9\\_5](https://doi.org/10.1007/978-3-642-10601-9_5)
- [11] W. Osborn, A  $k$ -nearest neighbour query processing strategy using the mqr-tree, in: Advances in Network-Based Information Systems, The 20th International Conference on Network-Based Information Systems, NBiS 2017, Ryerson University, Toronto, ON, Canada, August 24-26, 2017., 2017, pp. 566–577.  
URL [https://doi.org/10.1007/978-3-319-65521-5\\_49](https://doi.org/10.1007/978-3-319-65521-5_49)
- [12] A. W. Fu, P. M. Chan, Y. Cheung, Y. S. Moon, Dynamic vp-tree indexing for  $n$ -nearest neighbor search given pair-wise distances, VLDB J. 9 (2) (2000) 154–173.  
URL <https://doi.org/10.1007/PL00010672>
- [13] H. Proen  a, J. C. Neves, Fusing vantage point trees and linear discriminants for fast feature classification, J. Classification 34 (1) (2017) 85–107.  
URL <https://doi.org/10.1007/s00357-017-9223-0>
- [14] G. Zeng, Q. Li, H. Jia, X. Li, Y. Cai, R. Mao, An inclusion rule for vantage point tree range query processing, in: Human Centered Computing - First International Conference, HCC 2014, Phnom Penh, Cambodia, November 27-29, 2014, Revised Selected Papers, 2014, pp. 777–783.  
URL [https://doi.org/10.1007/978-3-319-15554-8\\_68](https://doi.org/10.1007/978-3-319-15554-8_68)
- [15] S. G. Liu, Y. W. Wei, Fast nearest neighbor searching based on improved vp-tree, Pattern Recognition Letters 60-61 (C) (2015) 8–15.
- [16] 高强, 张凤荔, 王瑞锦, 周帆, 轨迹大数据:数据处理关键技术研究综述, 软件学报28 (4) (2017) 959–992.
- [17] 吉根林, 赵斌, 时空轨迹大数据模式挖掘研究进展, 数据采集与处理30 (1) (2015) 47–58.
- [18] 杨伟, 艾廷华, 基于车辆轨迹大数据的道路网更新方法研究, 计算机研究与发展53 (12) (2016) 2681–2693.

- [19] 袁冠, 移动对象轨迹数据挖掘方法研究, Ph.D. thesis, 中国矿业大学(2012).
- [20] 王凯, 陈能成, 陈泽强, 基于mongodb的轨迹大数据时空索引构建方法, 计算机系统应用26 (6) (2017) 227–231.
- [21] 龚俊, 柯胜男, 朱庆, 张叶廷, 一种集成r树、哈希表和b?树的高效轨迹数据索引方法, 测绘学报44 (5) (2015) 570–577.
- [22] 陈建军, 于志强, 朱昀, 数据可视化技术及其应用, 红外与激光工程30 (5) (2001) 339–342.
- [23] 吴秀君, 网络环境下地图数据可视化方法的研究及其应用, Ph.D. thesis, 苏州大学(2008).
- [24] 朱孔强, 丁林花, 朱立顺, 聂国豪, 席永科, 基于百度鹰眼的校园巴士app, 数字技术与应用v.36; No.334 (04) (2018) 52–53+55.
- [25] 罗智勇, 黎小东, 基于数据库存储方案的高性能瓦片地图服务研究, 地理与地理信息科学29 (3) (2013) 52–55+112.
- [26] 霍亮, 杨耀东, 刘小勇, 乔文昊, 朱王璋, 瓦片金字塔模型技术的研究与实践, 测绘科学37 (6) (2012) 146–148.
- [27] 路东林, 智广玉, 地图发布平台下瓦片金字塔技术研究, 数字技术与应用 (3) (2013) 99–99.
- [28] M • Kumar, 朱慧, 1984世界大地测量系统(wgs84), 测绘技术装备 (2) (1994) 40–42.
- [29] 李长春, 蔡伯根, 上官伟, 王剑, 基于web墨卡托投影的地图算法研究与实现, 计算机应用研究29 (12) (2012) 4793–4796.
- [30] 胡水平, 岳淑英, 张求喜, 谷歌地图卫星影像数据获取关键技术研究, 测绘与空间地理信息41 (10) (2018) 79–81.

## 简历与科研成果

基本情况 韩淳，男，汉族，1994年8月出生，吉林省松原市人。

### 教育背景

**2017.9~2019.6** 南京大学软件学院 硕士

**2013.9~2017.6** 南京大学软件学院 本科

这里是读研期间的成果（实例为受理的专利）

1. 刘海涛, 韩淳, “基于优先点树的地理轨迹数据服务的设计和实现”, 申请号: 20xx1018xywz.a, 已受理。

## 致 谢

衷心感谢我的指导老师，感谢刘海涛老师在这近一年的论文写作过程当中给予我的悉心指导和帮助。刘老师平易近人的生活作风、严谨求实的治学态度以及一丝不苟、勤勤恳恳的工作精神将永远是我学习的榜样。感谢星环科技分布式存储部门给我提供一个良好的学习和实践的机会，一个能将理论知识运用于实践的理想场所。在这里，有幸参与到地理轨迹数据相关工作，让我接触到了很多业界最新的技术和观念，获得了许多书本上学不到的知识和技能，这将对我以后的生活和工作产生深远的影响。感谢项目组的同事们在项目的开发过程中给我的无私帮助和指导，从他们身上我学到了很多工作的方法和做人的道理，在星环实习的日子是我永远难忘的记忆。能够成为南京大学软件学院的硕士学员，能够在这里结识这位老师和朋友，我感到非常骄傲！衷心祝愿南京大学软件学院能够越来越灿烂和辉煌。最后，向答辩委员会的各位老师致以深深的谢意，感谢你们所付出的辛勤的工作！