基于互联网社交媒体的地理信息大数据分析与应用平台项目工作验收报告

**文档审查记录**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **日期** | **版本** | **修改说明** | **编写** | **校对** | **审核** | **批准** |
| 2015/10/19 | 1.0 | 首次撰写 | 杨超，熊艳 | 杨超 | 莫文婷 |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**目录**

[一、概述 4](#_Toc399162487)

[1、项目来源 4](#_Toc399162488)

[2、项目的技术和市场条件要求 4](#_Toc399162489)

[3、 研究依据 6](#_Toc399162490)

[二、研究工作完成情况 7](#_Toc399162491)

[1、研究工作的主要过程 7](#_Toc399162492)

[2、研究内容完成情况 8](#_Toc399162493)

[3、关键技术解决情况 13](#_Toc399162494)

[4、平台核心模块设计成果 15](#_Toc399162495)

[5、技术指标实现情况 26](#_Toc399162496)

[6、知识产权情况 28](#_Toc399162497)

[三、产品维护方案 28](#_Toc399162498)

[1、产品修改验证 28](#_Toc399162499)

[2、出错及纠正方法 28](#_Toc399162500)

[3、专用维护技术要求 29](#_Toc399162501)

[四、研制过程组织与管理 29](#_Toc399162502)

[1、研制计划执行情况 29](#_Toc399162503)

[2、技术管理情况 29](#_Toc399162504)

[3、质量管理情况 30](#_Toc399162505)

[五、成果及其应用 30](#_Toc399162506)

[1、推动成果转化的措施、方案 30](#_Toc399162507)

[2、前景分析 31](#_Toc399162508)

**项目工作总结报告**

## 一、概述

### 1、项目来源

从社交媒体信息中获取地理信息是一个新研究方向。随着社交网络的兴起，全球社交媒体服务的注册用户超过20亿。社交媒体（Kaplan and Haenlein 2010）信息，作为一种高活跃度的信息，在各个层面提供了海量的数据。能够在人和社会的交互层面，提供大量事件的信息。以新浪微博为例，每天产生的数据量高达数十个PB，规模庞大(Volume)，这些数据种类繁多(Variety)，有文本信息、签到信息、GPS数据、图片信息与视频信息，数据输入和处理速度要求高(Velocity)，用户在一些高峰时刻或者突发事件下，数据的输入和访问量非常惊人；价值密度低(Veracity)，社交媒体数据存在大量非相关数据。因此，社交媒体信息是一种符合大数据(Big Data) 4V 特征(李国杰和程学旗2012)的典型大数据。在社交媒体信息中，蕴含着大量的、潜在的地理信息。社交媒体信息的主体信息部分是文字信息，针对特定事件的表达关键词，通过知识挖掘和数据挖掘技术，可以从这些社交媒体信息中获取所需的地理信息。在大数据背景下，通过分析社交媒体中的地理信息，进行分析整个用户的地理信息行为，从而为城市的决策提供重要的信息。

在对地观测领域，我们所指的数据多是具有规范格式矢量数据和影像数据，在社交媒体信息中，这些文本、照片和视频信息等非结构化信息都不是规范表达的信息，如何对社交媒体信息进行提取、处理，组织，进而对海量社交媒体用户所发布的信息进行知识挖掘，获取用户发布地理信息行为，将是我们研究的重点。此外，由于各个社交媒体服务厂商各自定义的数据访问API对各不相同，目前并没有一个明确地针对某个特定社交媒体信息挖掘的基础服务框架。因此，针对海量数据的社交媒体服务（如新浪微博）构建相应的社交媒体信息挖掘服务框架，进而标准化的获取社交媒体信息，是将社交媒体信息纳入到对地观测网络中的关键。

### 2、项目的技术条件要求

根据项目的研发过程，我们从项目调研和项目的设计及开发中问题的解决等方面，对本项目中的技术条件的要求有了深刻的分析和讨论。

#### 2.1项目对技术上的要求

本项目首先要应用的操作系统为Cent OS 6.4版本。

目前国内外具备良好移植能力的软件也有很多，从设计的角度来看，往往要涉及到开发的语言的选择、开发框架的选择和开发方法的选择。另外，这种该具备移植能力的软件的设计思想一般都源于Linux软件的设计思想，从使用编译开关、通用开发语言、规范化API接口和通用应用设计模式等角度来实现系统的设计和移植。

其次是对开发语言的要求，本项目需要保证代码能够支持多个系统平台，因此，本项目采用Java语言作为开发语言。

第三，从空间数据的管理角度来看，大数据量的空间存储和快速访问技术一直是个技术的关键点，如何能够解决大数据量的空间存储，除了在空间索引上的优化设计外，高效的数据物理存储和读取模块的设计不可或缺。

第四，空间算法性能上的优化是平台的核心，设计简洁高效的算法除了从空间数据的分析入手外，在数据的缓存管理、数据的算法实现上都要做大量工作。

#### 2.2项目对市场的发展分析与要求

本项目针对社交媒体地理信息大数据的获取和分析为切入点，发展社交媒体信息服务方法，旨在1）提出开放式的社交媒体数据提取模型；2）构建通用的社交媒体大数据服务框架；3）对社交媒体大数据进行分析，结合城市事件和突发事件，揭示普通大众在城市突发应急事件中的活动和规律，从而为城市应急决策提供辅助决策支持。本项目的研究成果将拓展传统对地观测的研究内容，实现社交媒体信息在辅助决策系统研究中理论方面的创新，从而为智慧城市的智能决策提供新的理论和方法，对城市的发展研究具有重要意义。

#### 2.3同类产品的比较

目前，广大研究机构针对社交媒体数据的分析，尤其是中文社交媒体信息的研究，主要集中在有地理位置信息的媒体中的位置信息的抽取或优化位置的精度,或者利用这些数据对不含有地理位置信息的媒体或者用户进行位置的预测和推断。总的说来，和本系统相比，其存在具备如下不足：

首先针对社交媒体信息的地理信息的提取缺乏系统性，社交媒体服务作为一个庞大的用户系统，针对社交媒体信息中地理信息的获取应该是一个长期的过程，仅仅获取片段时间和片段区域的数据，无法对整个地理过程进行分析，典型的是深圳大学空间信息智能感知与服务深圳市重点实验室针对2014年新浪微博的签到数据进行城市功能分区研究。

其次是缺少大数据技术的支持，社交媒体地理信息数据相对于传统的地理信息数据，其大数据特征更多体现在庞大的数据总量和高效的处理要求，但目前的针对社交媒体地理信息的研究并没有在大数据平台下进行，对于社交媒体地理信息数据的采集、编辑、空间索引的灵活定制及数据的更新服务等特殊的要求基本没有成熟的解决方案，总体过于僵化。

最后在目前可看到的社交媒体信息处理系统中，大部分的研究缺乏整合的研究平台，即缺少从数据的获取、处理、分析、服务的整合平台。

基于上述的原因，可以看出，和本系统虽有相类似的研究，但就本项目的需求仍需要从多方面进行系统的规划和设计，要能满足既具备良好架构，又要具备灵活扩展特征，同时能够完成高性能的分析服务的平台产品。

### 3、 研究依据

#### 3.1 研究目标

本项目的研究目标是构建一个基于大数据技术的社交媒体地理信息信息获取和服务平台，针对互联网环境下社交媒体海量数据的获取、处理和分析需求，采用分布式计算技术实现设计媒体数据的高效获取、存储和处理功能，构建相应的社交媒体数据聚类服务，通过适用性评价与分析，建立可靠性强、效果优的分析算法模型，实现对用户的地理信息行为的分析，从而为辅助决策支持提供可靠的决策信息，从而为地理国情检测，实时应急服务等应用场景提供动态、实时的社交媒体信息和相应的决策信息。

为了能够很好地验证本平台的研发成果，还需要设计完成一套应用服务系统，选取2015年5月北京沙尘暴，开展互联网社交媒体的地理信息大数据采集、分析与应用技术研究实验，研究适用于国情普查数据更新、应急服务的可用性。

#### 3.2 研究内容

项目的研究内容主要包括以下几个方面：

（1）基于社交媒体大数据的地理信息服务框架

利用分布式大数据处理平台（Hadoop）构建社交媒体大数据获取平台，突破社交媒体服务商对于数据访问的限制，从而社交媒体信息的实时获取方法；将社交媒体地理信息数据按照空间进行索引、存储、分析，从而实现将社交媒体数据的获取和处理过程纳入到标准对地观测服务体系中。

（2）构建社交媒体大数据的事件提取和分析模型

基于数据挖掘和短文本信息提取技术，发展社交媒体大数据提取模型，实现精细内容的提取方法。实现事件态势的实时分析与跟踪，提高突然事件时空数据感知的准确性和时效性，从而缩小突发事件态势的“感知鸿沟”； 针对社交媒体信息中所蕴含的时空信息和文本信息，采用 DBSCAN聚类的分析方法对社交媒体信息进行数据挖掘，发现社交媒体信息所反映的用户空间聚合度和用户的地理信息行为，从而揭示社交媒体大数据和事件之间的关联关系。

## 二、研究工作完成情况

### 1、研究工作的主要过程

#### 1.1 结合自身特色，制定研究计划

首先，分析本课题所需要解决的核心难点，并进行逐级分解,将系统分解为新浪微博采集模块、Hbase数据库数据存储模块、Hadoop数据处理模块、Web服务模块。并利用分布式大数据处理平台（Hadoop）构建社交媒体大数据获取平台，突破社交媒体服务商对于数据访问的限制，从而实现社交媒体信息的实时获取；通过采用OGC制定的Geospatial Web服务规范，将社交媒体数据获取和处理平台按照Web服务的方式进行请求和处理，从而实现将社交媒体数据的获取和处理过程纳入到标准对地观测服务体系中，为开展数据分析工作奠定基础。

其次，结合社交媒体大数据的特点，利用自然语言处理（NLP）中常用的基于关键词的技术或者机器学习的方法，对社交媒体信息中事件信息的描述进行进一步的过滤处理，通过构建语料库或者训练数据的方法，对社交媒体大数据信息进行准确的分析。

最后，基于数据挖掘和短文本信息提取技术，发展社交媒体大数据提取模型，实现精细内容的提取方法。实现事件态势的实时分析与跟踪，提高突然事件时空数据感知的准确性和时效性，从而缩小突发事件态势的“感知鸿沟”； 针对社交媒体信息中所蕴含的时空信息和文本信息，采用DBSCAN和K means聚类的分析方法对社交媒体信息进行数据挖掘，发现社交媒体信息所反映的用户空间聚合度和用户的地理信息行为，从而揭示社交媒体大数据和事件之间的关联关系，从而为（城市）事件的实时感知提供技术基础。

#### 1.2 详细梳理，确定研究路线

基于上述因素，在合同要求的研究内容上，重点将产品定位为如下功能：

第一，新浪微博数据采集平台：本平台将持续地将所有新浪微博数据并存入MongoDB数据库中，进行数据去重后按照行政区划导出为文本文件，将文本文件导入至Hbase数据库中；

第二，社交媒体大数据分析平台：基于Hadoop技术，将社交媒体数据进行高效的存储、分析、处理和服务。

#### 1.3 具体研发计划与实施过程

本课题的开发共分了三个阶段来实施，其中第一个阶段主要开始进行新浪微博API接口分析，将采集的数据存入MongoDB数据库；第二个阶段主要进行数据分类并保存至MySQL数据库，实现与Web模块的交互显示；第三个阶段主要是利用Hadoop进行数据处理，完成聚类和时间序列分析。

### 2、研究内容完成情况

#### 2.1 研究内容的分解

除了初期对新浪微博采集模块进行分析外，重点研究了Hadoop数据处理模块、Hbase数据库数据存储模块，分析了Hbase数据表中的微博数据表，城市用户表和微博总用户表的字段。Hadoop数据处理模块提供相应的计算接口，从Hbase中读取数据计算并返回至数据库中。并针对应急事件或地理区域事件，开展相应的应用研究。主要研究内容如下

**（1）基于通用对地观测服务框架的社交媒体信息服务方法**

基于OGC定义的对地观测传感网服务标准定义社交媒体信息服务流程，挑选国内主流的社交媒体服务作为数据来源，针对互联网环境下社交媒体海量数据的获取、处理和分析需求，采用分布式计算技术实现设计媒体数据的高效获取、存储和服务功能，从而实现动态、实时的社交媒体信息感知服务。

**（2）基于社交媒体信息的地理行为分析方法**

对社交媒体数据进行分析，进一步采用诸如DBSCAN, K Means 聚类、时间序列分析等大数据分析理论与方法，通过适用性评价与分析，建立可靠性强、效果优的分析算法模型，实现对用户的地理信息行为的分析，从而为辅助决策支持提供可靠的决策信息。

**（3）系统原型验证**

选取2015年1-2个应急事件或地理区域，开展互联网社交媒体的地理信息大数据采集、分析与应用技术研究实验，与常规方式数据获取方式进行对比，进一步对算法进行检核、验证与优化，对数据价值密度和可用性进行评价，研究适用于国情普查数据更新、应急服务的可用性。

#### 2.2 研究路线与方法的选择

根据课题的功能要求和性能指标，我们在完成了微博数据抓取的开发方法学习后，迅速开展了整体平台的重新设计与功能模块划分，以完整抓取微博数据的高效存储和管理为重点，把数据存入Mongodb数据库，然后将微博抓取信息修改及数据导出，在一天的数据抓取工作完成以后，将当前微博抓取的日期写入MySQL数据库中，并对该天的数据进行去重，将数据按照行政区划导出为文本文件。在一天的数据导出后，将导出的数据按要求导入至Hbase数据库中，导入文本数据至Hbase数据库的同时，将微博数量信息写入MySQL数据库的城市每天微博数量统计表中，修改省级信息表与市级城市信息表中微博总数。合理科学的利用这三个数据库，完成了整个课题要求的功能指标和性能指标，从而达到项目的数据存储与管理要求。

1、研究的技术路线

项目的总体技术路线如下图所示，在研究了总体系统框架设计、数据获取模块，Hbase数据存储模块，MySQL数据模块，Hadoop数据处理模块，web模块等内容的基础上，分别实现空间数据的获取，存储及管理等关键技术的同时，还要完成相关的平台数据存储规范、平台模块设计准则和应用开发及接口规范等规范，为构建可扩展态势提供基本保障。

完成系统的总体设计和对应规范后，重点进行新浪微博数据的抓取和存储、并研究了Hadoop数据处理模块、Hbase数据库数据存储模块，分析了Hbase数据表中的微博数据表，城市用户表和微博总用户表的字段，保证平台的正确性和高效性。

最后，基于数据挖掘和短文本信息提取技术，发展社交媒体大数据提取模型，实现精细内容的提取方法。通过web模块关联数据库，对数据进行形象化展示，并且支持用户对各种社会事件进行主题词查询，从而揭示社交媒体大数据和事件之间的关联关系，从而为（城市）事件的实时感知提供技术基础。

项目总体技术路线图如下所示：

//添加图片

图2.1 项目总体技术路线

1. 相关设计的分析

（1）数据模块设计

数据模块设计是整个平台是否具备良好扩展性的基础。数据模块包括Hbase数据存储模块，MySQL数据模块，Hadoop数据处理模块。整个平台、包括基于平台的各个应用，都是建立在数据模块的基础之上，并以模块为功能单位来实施，因此这些模块的设计和划分需要严格遵循以下的要求：

1. 先决定模块类型，再选择相应技术开发；
2. 严格遵守模块定义的接口；
3. 模块应设计为可重用设计；
4. 命名模块时应以最大化易用性为目标；
5. 模块应具有精心选择的粒度；
6. 模块功能应是内聚而完整的；
7. 模块应对实现细节进行封装；
8. 模块适应多种调用模式；
9. 模块具有无状态接口；
10. 模块应使用状态事务建模；
11. 模块的功能定义应表示业务动作。
12. 数据库的选择

由于获取的微博数据数据是需要强大的数据库作为支持，对于不同的模块功能，采取不同的软件平台。在微博数据存储模块中，利用十分强大的数据库Hbase，它能够融合key/value存储模式带来实时查询的能力，以及通过MapReduce进行离线处理或者批处理的能力，并可以从中获得综合分析报告。在Hadoop数据处理模块中采用MapReduce，因其是一个软件框架，用以轻松编写处理海量（TB级）数据的并行应用程序，以可靠和容错的方式连接大型集群中上万个节点。

1. web模块设计

web模块包括数据展示模块，用户模块，管理员模块。数据展示模块包含以下功能：展示当前Hbase数据库中的数据量；定义example展示，展示需求示例。web模块能够很好地反应平台模块设计的合理性、功能的正确性和算法的高效性。我们根据微博数据的特征，对数据进行整理分类，对微博数据和城市微博数据利用Echart产品进行数据可视化展示饼图(环形图)、柱状图(条状图)、折线图(区域图)、地图、散点图(气泡图)、K线图、力导向布局图等各种12类图表，同时可以在这些图表上添加标题、图例、工具箱、详情、数据区域、值域、时间轴等组件。同时，还在web系统上设计了主题词管理功能，可以让用户进行主题词数据查询和主题词与地图数据的可视化交互操作，能够更加全面的理解微博数据。

3、技术路线评述

本项目实施过程中遵循的技术路线有如下特征：

（1）充分利用现有规范，注重平台构建和分析

本项目研究与系统研发中，充分利用了国内外对社交平台的最新研究理论和服务应用领域的案例，将重点放在将社交媒体数据视为人类观测的结果，采用传感网服务标准，针对社交媒体数据的获取、存储、和服务进行操作的封装，可以使用传感网服务的操作接口对社交媒体信息的获取过程进行操作。结合社交媒体信息的现势性，利用统计分析和空间分析方法，探寻突发事件的时间势和空间分布，为应急响应提供决策支持。

（2）重视服务接口标准化和规范化，强调开发的规范性、重视集成

在项目实施过程中，采用OGC制定的SWE服务规范，将社交媒体数据获取和处理平台按照Web服务的方式进行请求和处理，从而实现将社交媒体数据的获取和处理过程纳入到标准对地观测服务体系中。用户通过OGC Web服务接口实现对社交媒体数据的操作请求，用户的操作请求在社交媒体信息提取和分析处理框架（Social Media Information extracting, analyzing, and processing Framework）中进行解析，该框架将动态的分析用户的请求并调用相应的功能模块完成对社交媒体数据的操作。

#### 2.3 研究内容完成细节

本本项目结合合同要求，对相关研究的完成情况如下：

* 五个模块数据

1、新浪微博采集模块

2、Hbase数据库数据存储模块、

3、Hadoop数据处理模块、

4、MySQL数据处理模块

5、Web模块

* 技术成果

1、把海量的微博数据存储在分布式计算机节点，并且提供接口然这些数据能够被访问。

2、搭建面向多主题的自动数据分析系统框架。

3、构建行政区域划分的索引结构和不同领域的多层级主题词。

4、获取微博数据的统计分析和微博数据在相关主题词上的时空分析，提供用户提交层级主题词的公共服务接口。

* 文档列表

《2015-08-15-初次撰写-V1.0》

《2015-08-20-添加模块设计-V2.0》

《2015-08-31-修改添加MySQL模块-V3.0》

《2015-09-20-修改部分数据库及命名-V4.0》

### 3、关键技术解决情况

#### 3.1 TB级数据管理

由于抓取的所有微博数据都是以JSON格式存储在文件中，且数据都是无规律的，而且数据量非常大，都是TB级的。并且这样的数据集无法用传统数据库工具对其内容进行抓取、管理和处理。然而，Hadoop是一个开源的可运行于大规模集群上的分布式文件系统和运行处理基础框架。通过集群上构建Hapdoop环境进行海量数据(结构化与非结构化)的存储与离线处理。我们结合Hadoop中的HDFS和MapReduce功能，利用HDFS(Hadoop Distributed File System)在大数据集方面的应用，实现了对数据读写的高吞吐率。我们也利用HDFS以流式数据访问模式存储超大文件，在硬件的集群上运行。采用数据处理的编程模型MapReduce，处理大型数据集。

#### 3.2 数据库字段的查询与增添

大范围新浪微博数据的高效管理是本系统的核心内容，也是本平台开发中的重要技术要求，由于地理空间数据的来源广泛、格式多、标准杂、数据量庞大，且随着web模块功能的实现，用户不断地往数据库增添新的字段，这就会无形中增加数据库存储与查询数据的压力，这时需要为数据库非结构化的数据提供透明的水平扩张能力。我们利用Hbase的面向列存储的分布式存储系统，实现高性能的并发读写操作，同时利用Hbase对数据进行透明的切分性能，这样就使得存储本身具有了水平伸缩性。在我们设计的数据库中，列的可以动态增加，并且列为空就不存储数据,节省存储空间。

#### 3.3 构建基于社交媒体大数据的事件提取和分析模型

基于数据挖掘和短文本信息提取技术，发展社交媒体大数据提取模型，实现精细内容的提取方法。实现事件态势的实时分析与跟踪，提高突然事件时空数据感知的准确性和时效性，从而缩小突发事件态势的“感知鸿沟”； 针对社交媒体信息中所蕴含的时空信息和文本信息，采用K means聚类的分析方法对社交媒体信息进行数据挖掘，发现社交媒体信息所反映的用户空间聚合度和用户的地理信息行为，从而揭示社交媒体大数据和事件之间的关联关系。

结合社交媒体大数据的特点，利用自然语言处理（NLP）中常用的基于关键词的技术或者机器学习的方法，对社交媒体信息中事件信息的描述进行进一步的过滤处理，如图2所示，通过构建语料库或者训练数据的方法，对社交媒体大数据信息进行准确的分析，从而为（城市）事件的实时感知提供技术基础。

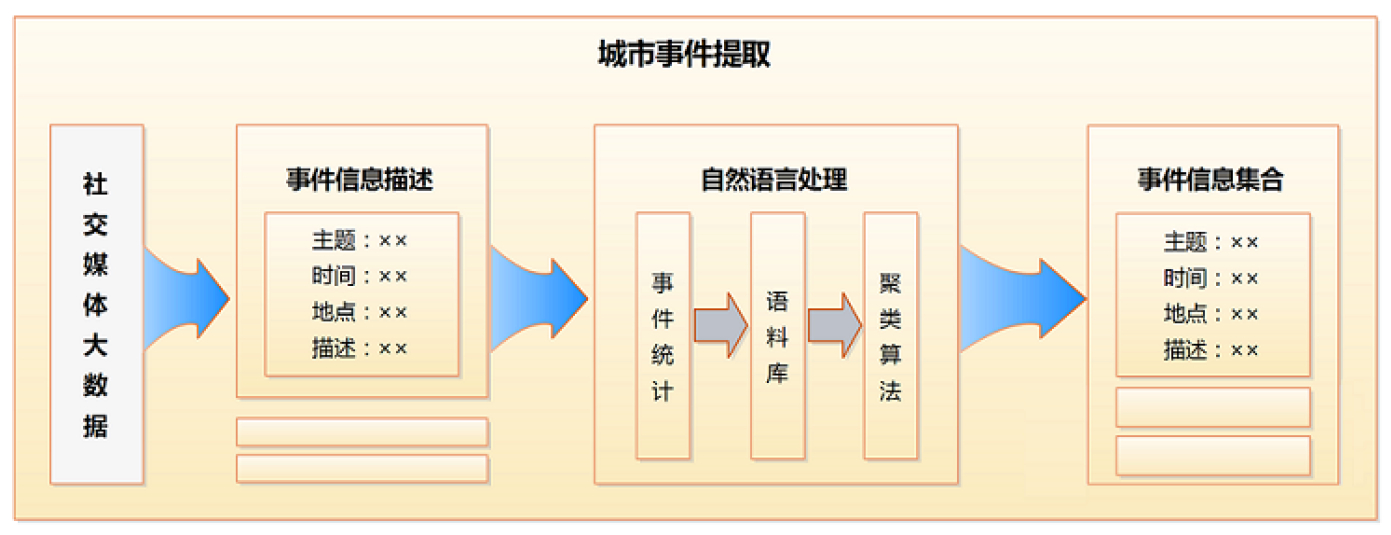


图2 基于社交媒体大数据的（城市）事件提取分析方法

#### 4、平台核心模块设计成果

#### 4.1 平台的层次化设计成果

参照设计，整个系统是从底层到应用层分别进行划分所构成，从底层的抽象层到上层的应用层都实现了完全分离，由底向上可以将态势核心引擎依次划分成资源层、数据层、服务层及应用及展示层。



图2.2系统层次结构图

#### 4.2 可移植底层设计成果

根据硬件抽象层（HAL）和操作系统抽象层（OSAL）的设计思想，我们在平台架构层设计中，抽象层的实现主要涉及的模块包括内存管理、文件、线程、时间等部分，同时还针对不同的操作系统进行了变量、字节序、通用数据结构等内容的封装设计。为了能够很好地实现这种不同平台的差异性处理，利用宏变量开关的方法来对整个代码的结构做了组织，这些与架构有关的代码都单独组织，统一放在了inc目录下的arch目录下，里面包含了平台的差异化定义，内存的操作，文件的访问等函数接口声明，使得所有与平台架构有关的部分都进行了剥离并不与平台的其他模块产生关联。



图2.3 平台架构层的抽象层设计成果

对于基础平台定义文件来说，它主要包括了与字节对齐和应用有关的一些数据结构，部分结构主要如图2-4：



图2.4 部分公用数据结构定义

对于具体的平台来说，其包含的内容都有所不同，这些内容在每个具体的系统都会有所差异，具体可以参考每个平台对应头文件中内容的描述。通过设定不同的操作系统下的编译开关，可以很容易地实现对同一套代码的一致性编译，并实现对多好总平台的支持。

#### 4.3 社交媒体大数据存储模型设计成果

1、空间数据处理过程分析

本系统的空间数据的存储和组织是本平台的关键技术。如何解决大数据量与有效设备资源之间矛盾的空间数据物理存储与组织方法具有重要的意义。对于数据的最终生成过程，其流程设计如下：



图2.5 地图数据处理流程

结合图2.5内容所示，需要对本系统中的空间数据存储从下面几个方面进行考虑和设计：

* 地图要素分层和LOD分层：通过将不同的地图要素类型进行分层保存，使得地图在显示时可以做到指定要素类型的显示，除去不需要显示的内容。同时在每一个地图要素层内部对数据进行LOD技术的分层，优化在多比例尺下的地图数据的显示。
* 地图数据的数据压缩：减少地图数据对外存的空间占用和在加载数据时的内存占用。同时处理大量的浮点数，尽量使用整形代替浮点，优化数据存储的位置和方式，在保证数据精度的前提下，一切以减少移动计算量为目的进行数据处理。
* 地图数据的分块：使用网格索引将地图数据分成固定大小的数据网格块，显示时只需计算屏幕显示涉及到的地图块，进行数据加载和显示，减少了地图在显示时读取的数据量，减少了内存占用，加快显示数据。

2、数据预处理标准

平台中为了兼容不同的数据格式，必须指定一个预处理标准。只有将多个厂商的数据都按着这个标准进行数据的预处理后，才能保证平台在进行数据转换时的正确。下面介绍点、线部分各自的处理标准：

1. 点

兴趣点表包含三个部分：ID、TYPE、NAME，按照TYPE类型从小到大排列，方便数据转换时按照类型划分图层。具体字段数据类型和说明如下表：

表2-1 POI点处理标准说明

|  |  |  |
| --- | --- | --- |
| **字段** | **类型** | **意义** |
| **ID** | INT | 点的序号 |
| **TYPE** | INT | 点的类型 |
| **NAME** | CHAR[40] | 点的名称 |

TYPE字段记录兴趣点的类别：一般地标、党政机关、餐饮住宿、观光旅游、文体休闲、购物指南、交通服务、科研教育、公共服务、商业机构等十大类。每一大类拥有独立的ID前缀，如001、002表示，大类下细分各个小类，如餐饮住宿大类003下分为五星级宾馆酒店ID：003100，四星级宾馆酒店003101，快餐店003200等，如果小类下数据量还是比较庞大或者为了提高数据查询的精确性，还可以在细分子类，例如连锁酒店003105小类下的如家快捷003105001，7天连锁003105002等，根据来源数据的详细情况，可不断细分。分类的数据需要正确放置，以保证图层归类时不出现错误。

道路结点表分为以下几个部分：ID、NAME、LINKROADID，按照结点ID排序，说明如下表：

表2-2 拓扑结点处理标准说明

|  |  |  |
| --- | --- | --- |
| **字段** | **类型** | **意义** |
| **ID** | INT | 结点ID |
| **NAME** | CHAR[40] | 结点名称 |
| **LINKROADID** | CHAR[40] | 结点连接的道路ID |

LINKROADID记录的是与当前结点想关联的道路的ID，可能为多条，使用“-”符号分隔。

1. 线

线表示的道路与点不同，即表示了真正的道路（除去铁路），也包含了拓扑关系中的边，需要以下几个字段：ID、LEVEL、NAME、LANES、DF、TYPE、LENGTH、SNODE、ENODE。按照LEVEL的值顺序排列，用于道路数据的分层使用，拓扑数据部分结合道路结点的表共同转换成平台使用的路网拓扑关系文件。线数据字段说明如下表：

表2-3 线数据处理标准说明

|  |  |  |
| --- | --- | --- |
| **字段** | **类型** | **意义** |
| **ID** | INT | 道路ID |
| **LEVEL** | INT | 道路管理等级 |
| **NAME** | CHAR[40] | 道路名称 |
| **LANES** | INT | 道路车道数 |
| **DF** | INT | 道路的允许行驶方向 |
| **TYPE** | INT | 道路形态及特征（主道、辅路、匝道等） |
| **LENGTH** | FLOAT | 道路长度 |
| **SNODE** | INT | 道路起始结点ID |
| **ENODE** | INT | 道路终止结点ID |

道路管理等级LEVEL分为高速公路、国道、快速路、省道、主要道路、次要道路、一般道路、出入目的地道路、细道路、步行道路，道路导航时使用7级以上道路进行道路计算。DF道路的允许行驶方向有四种情况：双向通行（默认值）、与数字化方向相同可通行、与数字化方向相反可通行、双向禁行。

#### 4.4 数据索引方案设计成果

地图数据中的空间地理信息都是带有地理坐标的，一般采用的球面坐标，也就是经纬度坐标表示。经纬度坐标通常使用双精度浮点数来保存才能保证在大比例尺情况下的显示精度，在拥有大量空间坐标数据的GIS平台中直接保存双精度浮点经纬度会占用非常多的外存空间。需要对地理数据进行数据的压缩，使得地图数据的数据体积尽可能的小，同时保证数据的精度在可接受的范围内。一种是将双精度浮点改为单精度浮点数进行存储，对于一般要求的地图数据，是可以接受的。另一种是将浮点数改为整数存储，可以进行直接的经纬度扩大进行经纬度的整数化存储，也可以通过投影计算将地理坐标转换成投影坐标，以米为单位等，这种方法比第一种的好处是在移动环境中，整数的计算往往比浮点计算要高效很多（很多的移动设备硬件不支持浮点运算），同时地图的很多运用都是使用米为单位，所以以米为单位的投影坐标可以进行直接的计算，加快计算速度。

1、滤点压缩

在地图数据中，一条线或者一个面是由大量的地理坐标点的串组成的。而在地图进行显示的时候，根据LOD技术在不同的比例尺下，需要的显示精度也是不一样的。在小比例尺的情况下进行显示时，由于移动设备屏幕大小的限制，地图要素的不同坐标点转换为屏幕坐标点有会有大量的点重合在一起，可以通过提前将这些预计会重合的点去除，以达到减少数据量却又不影响显示效果的目的，这种方法叫做滤点压缩法。

通过对步长法、线段过滤法和Douglas-Peucker算法的分析比较，我们最后才用了Douglas-Peucker算法来进行基本的数据压缩，减少点的数量，来优化存储。

2、属性信息压缩

对于存储于文本数据库的属性内容，为减少其存储空间，往往也会对其进行无损的信息压缩。在导航数据的拓扑数据中记录了大量的关于道路属性的信息，如道路的等级，道路的单双行状态和收费情况等。但这些属性的状态一般只有有限的几种，如道路的通行情况有：0双向通行（默认），1数字顺方向通行，2数字逆方向通行，3双向禁行4种状态，在数据存储的时候如果按着位（bit）的方式进行存储的话，只需要2位即可。同理道路的等级，长度等信息都可以压缩到位单位的信息存储，如图2.6所示。



图2.6 属性位压缩示意图

通常采用一个整型，即具有4个字节，32位的一个数据类型进行属性信息的存储，将所有需要存储的属性按照一定的位置和位数放入这32位中，然后按照一个整型的类型进行数据的存储。在数据获取时，读取这一个整型数据，按照不同的位偏移，就可以将各个属性数据获取并赋值给相应的变量。

3、算法压缩

在地图数据进过地图分层，数据简化、属性压缩、数据分块后，整理出各个数据块内容，有序的排列，即将进行数据存储的时候，针对仍然比较大的数据块内容，可以借助zlib压缩库进行数据的再次压缩，而后对数据块索引信息头内的压缩位进行压缩标识的修改，明确此数据块进行过数据压缩，解压是需要解压才能真正获取数据。此种方式的压缩并不影响平台在启动时的数据加载速度，只有在需要获取真实地图要素数据时，才需要进行数据的解压获取，但由于数据按照数据块的形式进行存储，每一块的数据量并不是非常的大，同时高效的zlib压缩模式，使得数据在获取时的效率基本不受解压缩过程的影响。

#### 4.5 平台后台管理设计成果

导航应用中的道路网是整个系统应用的核心，道路网络数据是由节点、边线的拓扑信息和相关的属性构成。数据提供商在提供道路网数据时，通常提供三个部分的数据：

1）道路数据： 道路的形状节点坐标；

道路的拓扑连接信息：PLID、SNODE(始结点ID)、TNODE(终结点ID)，

道路的属性信息：TL(红绿灯)、LLvl(主辅道识别码)、Dist(道路实际长度)、TOLL(是否收费)、HSL(最高限速)、LSL(最低限速)、FRegionDes(政区代码)、RLvl(道路类别)、LC(车道数及内部路标识码)、FC(道路等级)、Num(道路编码)、OW(单行道)、DL(专用道路识别码)、AveS(平均速度)

2）道路结点数据：结点的形状坐标；

结点的拓扑连接信息：NodeID、NodeNum、JNID1…JNID8、JA1…JA8

3）转向表数据： 描述了一条边通过某结点到另一条边的转向信息；

主要包括：FromID、JUNTIONID、ToID、时间限制表、车辆种类限制表等，

在对这三部分的数据分析基础上，我们将道路的网络拓扑描述如下：

* 在分省引导道路文件中，每条道路的PLID都是唯一的；
* 在分省引导道路结点文件中，每个结点的NodeID都是唯一的；
* 每条道路的属性字段中详细的记录了道路信息，包括道路矢量化方向、道路管理等级、道路类别等信息；
* 每条道路的属性字段中，都记录了该道路两端连接的结点NodeID，通过道路，可以快速的找到对应的结点及附属信息；
* 每个结点的属性字段中，都记录了该结点连接的道路PLID，通过该结点，可以快速的找到结点和道路、道路和道路之间的连接关系；
* 在结点转向表中，每条记录通过道路的PLID，描述了道路之间的转向关系，是否存在车辆种类限制、通行时间限制等信息

其结构如下图所示：



图2.7 引导道路网络拓扑组织

在这些拓扑数据的抽象下，将边线的拓扑信息和属性信息作为一个整体构成边的拓扑数据，使用元数据格式存储，节点拓扑信息和属性信息作为一个整体构成点的拓扑数据，并使用元数据格式存储。

由于网络是存于不同的层中，且根据不同的需要，调用不同等级的拓扑数据来进行分析，因此，在网络中需要设计全局的网络ID来保证网络分析中的ID唯一性。

#### 4.6 社交大数据统计展示设计成果

地图数据是一种空间数据，所以地图数据的索引需要采用空间索引算法。地图数据按照图层和LOD技术进行了两次分层，空间索引算法需要针对最小的LOD分层进行索引的建立。空间索引策略有很多种，包括BSP树、K-D树、K-D-B树、网格索引、四叉树索引、R树、R变种树等，本系统需求主要是对地图数据的查询，很少设计到数据的修改、删除等功能，所以采用具有较高查询效率的空间索引策略。

网格索引的主要思想就是将一整幅的地图按照行列的形式分成等大的网格，地图数据存储在对应的网格数据块内。网格索引的最大外框是根据当前需要切割的图层的所有要素所占的最大范围决定，同时将这个矩形分成M×N块的网格。假设地图最大矩形的长宽分别为L、W，则每一块网格的长宽为L/M、W/N。为得出真正的M和N的值，需要确定单个网格期望包含的平均要素数量S，根据图层总要素数计算出应该需要有多少个网格才能包含下所有要素。同时为保证网格的形状尽量接近正方形，M和N的参数选择比例需要基本接近L和W，即。最终的计算公式为，。



网格索引在数据的划分时比较的简单，只需要遍历所有的地图数据，判断出该要素的外包矩形属于哪一个网格内即可。

当需要获取地图数据时，计算出需要获取的地图数据的范围，在根据网格的大小和各自的长宽，计算出范围覆盖的网格是哪几个，即可通过建立的网格索引，快速定位到数据的所在偏移，进行数据读取。

#### 4.8 地图展示设计成果

### 5、技术指标实现情况

本系统时使用Java语言进行开发，最后平台能够在Centos系统上编译通过，并正确执行演示系统，均满足本课题合同要求，。

#### 5.1 功能指标实现情况

1、带有地理位置信息的微博数据的获取

本产品实现了对微博数据的持续抓取，数据抓取平台可以持续的每日对新浪微博的数据进行抓取并进行存储。我们通过Hadoop技术，保证海量数据的快速空间存储，以及对数据的快速获取和分析能力。

2、地图数据的访问

本产品能够对地图数据进行访问读取，特别是对道路网数据，道路路段与节点间拓扑关系的访问。对于如下接口，包括：

1. 打开指定路径和文件名的地图；
2. 关闭指定路径和文件名的地图；
3. 获取当前打开地图的存储路径和文件名称；
4. 获取当前打开地图的边界地理范围；
5. 获取当前打开地图的精度（比例尺）；
6. 获取当前打开地图的图层数量；
7. 获取指定图层名称的地图图层；
8. 获取地图中道路路段和道路节点的总数；
9. 获取指定ID的道路路段和道路节点；
10. 获取指定道路路段或道路节点的ID值；
11. 获取与指定ID的道路节点相连接的道路路段；
12. 获取与指定ID的道路路段相连接的道路节点；
13. 获取指定道路路段的行驶方向；
14. 获取指定道路路段的长度信息。

都已经完全实现，并能够进行测试实验。

3、地图数据的基础计算分析

本产品已经实现的基础计算分析功能包括距离量算、坐标转换、缓冲区分析等功能。

1. 距离量算：计算指定经纬度的两个坐标点之间的距离长度（单位：米）；
2. 坐标转换：经纬度坐标与屏幕坐标的相互转换；
3. 缓冲区分析：已知当前定位点：double dX, double dY；搜索半径：double dR。求以当前定位点为圆心，以dR为半径的圆形候选区域内所有路段（包括与圆部分接触）的ID的集合。

4、应用演示系统开发

我们以武汉市的导航路网数据为基础搭建了一个演示系统，用于所有功能和性能的演示和测试，我们根据自己录制的GPS轨迹，用我们的算法验证了匹配过程中的系列算法，本算法在Windows系统上可视化运行，在Linux和VxWorks系统上，是利用输出文件来验证结果的。该系统能够在Windows、Linux和Vxworks三个操作系统上应用，并验证了上述的2和3中包括的功能，并能够配合甲方完成用于导航定位校正的原型系统。

#### 5.2 性能指标实现情况

1、实时的信息处理能力

该平台具备高效的信息处理能力，在实时信息处理过程中，要求本平台能够对每秒接收到的信息包进行匹配运算，从接收到数据到单次匹配校正结束的总时间开销不超过500毫秒。根据我们的真实测试，拿武汉市20公里长的GPS轨迹进行实际匹配测试，根据生成的匹配分析结果，这些处理的结果都小于500毫秒，能够达到实时的信息处理要求。

2、高效的空间数据存储能力

我们使用了高德提供的武汉市导航路网数据进行了测试，地域范围大于40\*40公里，精度为10米，最后生成的数据也小于50兆，完全满足本课题要求。

3、快速的功能扩展能力

平台的开发都是基于相对独立的模块化算法集合实现，模块间的和模块内的函数可以根据需要进行扩展增强，对于总的架构和模块关系都不会产生影响，有很好的扩展性。

4、优秀的跨平台能力

本软件产品目前已经支持Windows、Linux和Vxworks三种操作系统，其源码需具备优秀的移植性能。

#### 5.3 数据输入

本产品能够直接导入MapInfo（\*.tab）地理数据格式，并进行格式转换，生成本平台所需要的地图数据格式，数据中包括了导航电子地图标准道路网数据及其拓扑信息。目前已经能够对高德提供的标准导航路网数据进行处理，通过对武汉市的整个城市的导航路网数据做了测试，能够处理道路中心线精度<=5米的导航电子地图，平台可以处理全国范围内任何位置的导航路网数据。

本软件产品在VxWorks操作系统中，以DKM方式提供相应开发接口，在Windows操作系统中，以静态库方式提供开发接口，Linux中则是提供完整的源代码和Makefile。

### 6、知识产权情况

本次开发成果的知识产权归双方共同所有，相关的关键技术和文档及专利也是如此。

## 三、产品维护方案

### 1、产品修改验证

在开发的过程中，由于整个的开发需要兼顾多个操作系统的支持，因此在开发工具和验证过程中严格遵循统一的软件工具和开发方法，我们选用的开发工具主要有如下几种，Windows平台上的Visual Studio 2005，Windows平台上的VxWorks仿真开发工具WorkBench，Linux平台上的GNU系列工具（4.0以上版本），由于这几个工具都是目前主流的开发方法，采用的开发语言也是最为标准的C，在工具开发中采用了标准的C++开发，为了减少调试的难度，我们尽量避免了模板类、STL的使用，整个开发过程中由于多采用已有的软件开发成果和技术积累，因此没有存在大的产品修改过程，并提前完成了产品开发。

### 2、出错及纠正方法

开发中的出错根据目前的总结主要集中在以下几个部分，第一类是转换工具的错误，其原因主要是导航路网数据不规范带来的错误引入，这类错误的纠正多是从数据的导出工具的规范化处理上进行实现，第二类是空间数据存储时的属性管理上有部分信息遗漏，这个在利用工具中的路网数据测试时就可以发现就能集中解决，主要是对导航路网属性信息的检查和验证；第三类是空间分析算法的模块开发中，算法的正确性问题，此类问题也是通过实际的空间模型和算法分析比对结果来检查验证，由于属于平台级的开发，其实大量的函数和功能验证在转换工具和模拟测试平台中就已经能够系统地验证，因此可以说很多问题是在开发过程中解决的，能够较好实现软件质量的控制。

### 3、专用维护技术要求

本系统是一套平台级产品，且马上要开展第二阶段的模块升级与扩展，无特别的专用维护的要求。

## 四、研制过程组织与管理

### 1、研制计划执行情况

整个系统的研制过程总体顺利，总的说来，需求方与己方在项目的需求和技术交流探讨上有了很深的共识，保证了整个项目的顺利实施。在开发过程中，双方积极交流和提出建议，积极寻求问题解决方案，极大的推动的项目研发的进行。整个项目在开发方法上的研究、平台框架的搭建、功能的开发、服务的开发等各方面都能如期完成并进行严格测试后交付使用。需求方在开发过程中也能够主参与和承担测试与方案的可行性分析，整个项目如期交付，并在质量上有很好的保证。

### 2、技术管理情况

系统开发中的核心技术仍然由中国地质大学（武汉）这个团队所掌握，开发过程中，采用了版本管理，能够对问题进行追溯回放。同时，在项目开发过程中对客户的实际需求、应用场景有了深入理解，双方对项目的核心技术方案都有了共同的认识，在方案设定、核心技术解决、关键问题解决方案上，有了可遵循的技术文档做参考。

#### 2.1 技术协调总体情况

双方在研发的关键问题和关键技术上都有可参考的分析和解决方案，彼此对于项目的理解非常深刻，并且能够很好地将问题给予解决，相关的文档已经作为本项目的验收文档之一进行管理。

#### 2.2 软件版本控制情况

软件开发中建立局域网内的代码管理机制，研发人员按照模块进行任务划分，并由核心人员进行封装集成，保证代码的一致性和完整性，并且代码机器都能够进行一定程度的隔离，在移交后，客户重新对代码进行整理封装，再进行使用。整个过程都可进行控制监管。

#### 2.3 软件测试情况

功能测试主要是针对系统进行压力测试，通过测试软件对系统进行并发访问测试，检测系统的抗压性，观察系统在多并发访问情况下能否快速处理海量的微博数据，提供正确的请求结果。在测试过程中，我们模拟了多个用户同时发出http请求，系统在多并发访问情况下，能做到准确响应用户请求，响应时间短，数据返回准确，符合设计文档的要求。

### 3、质量管理情况

#### 3.1 质量目标实现情况

本项目按照质量目标的要求，已经基本完成了各项要求，系统测试效果良好，对于产品的质量、功能和性能测试反馈良好，已经实现了文档中约定的目标。

#### 3.2 质量岗位责任落实情况

开发过程中，研发小组每个成员都能很好履行自己的职责，开发中各成员工作效率都有一定保证，对分配给自己的任务一丝不苟的完成，成员间工作对接有条不紊。无论是产品的安全、质量还是个人的工作效率都能很好落实。

#### 3.3 质量保证措施落实情况

研发过程严格遵守文档约束，认真对待每一个细节，不马虎，不放松。在遇到困难时及时提出小组讨论并找出最优解决方案，很好的执行了各项质量保证措施，保证了产品的质量。

## 五、成果及其应用

### 1、推动成果转化的措施、方案

本项目主要围绕社交媒体数据的发现和分析方法开展研究，并针对应急事件或地理区域事件，开展相应的应用研究，分析当前社交媒体大数据所蕴含的用户发布地理信息行为，采用数据科学和大数据分析理论与技术，设计地理信息数据实时抓取、过滤、处理、分析、存储和规范化的软件原型系统，完成为国情普查信息更新、应急提供在线服务的可行性分析。截止目前，项目研究已取得一定成果，为了能够将本次研究成果能够有更好的拓展和推广，可采用以下措施：

#### 1.1 紧密配合，优化平台性能，改进平台质量

本项目的研究是在己方和需求方的共同协作下完成的，从项目开始的可行性研究分析阶段到项目正式开始，再到现在项目研究取得一定成果，每一个过程都是双方互相学习和改进的过程。通过双方不断的交流沟通，需求一步步明确，这些都为我们展开研究和提升技术要求、提升软件性能提供了良好的基础。在后期的服务过程中，我们应该在工作上紧密配合，积极主动地提出和解决问题，根据需求设计出最优的解决方案，改进平台质量，提升平台性能，让平台在稳定运行的同时提供更好的服务。

#### 1.2 进一步深入研发，将平台的适应性进一步加强

本项目的成果是一个基于社交媒体的海量数据服务框架，框架的关键数据和核心算法都已完成，并且能提供及时有效的数据信息和服务。随着社交媒体的快速普及和发展，平台所需处理数据量也会逐渐增加，为了能让平台在日益增加的数据前保持稳定运行，提供更加可靠的服务，建议在目前平台的开发基础上更加深入研究，优化数据挖掘、数据处理等算法，从根本提升平台的抗压性和适应性。

### 2、前景分析

随着互联网、无线通讯技术的发展，社交媒体服务蓬勃发展。目前，社交媒体用户已经覆盖了中国一半以上的人口。社交媒体信息中蕴含着大量潜在的地理信息，作为一种高度和“人”相关的数据，社交媒体信息可以在多种层面上反应人类的地理信息行为。

基于互联网社交媒体的海量空间数据挖掘与分析技术研究若成功实施，将解决社交媒体信息应用在对地观测领域中所面临的一些重要问题：如何使用通用的对地观测信息服务接口来按照用户的需求发现、获取和处理社交媒体信息？如何发现社交媒体信息中所蕴含的用户地理行为？例如在紧急事件中，社交媒体信息可以为决策部门提供什么样的信息？以地震救灾为例，这些信息包括：在灾害区域有多少人受灾，有多少人需要援助，这些人在哪里？社交媒体信息作为一种和个人高度相关的、实时信息，能够反应受灾（个体）群众的状态信息，通过获取、处理、和分析这些社交媒体信息，挖掘出有效的地理信息，最终为决策部门提供数据。

本项目将针对2015年中国城市事件或者突发事件进行研究，应用感知平台构建相应的事件感知服务，用于研究人群活动和城市事件之间的关联关系。本项目将推动社交媒体信息在对地观测研究中的应用。同时，希望通过本项目的研究，推动用户参与到地理信息数据的贡献中，并作为现有对地观测手段的补充。因此，本项目的研究兼具研究价值和应用价值。