**时空大数据地理计算与分析实验平台**

**（一）项目研究的主要内容，解决的主要技术难点和问题等**

在地理信息相关课程的教学和实验环节，实现时空大数据的处理和分析案例对于学生充分理解和应用地理信息大数据具有重要的研究和实践意义。随着当今互联网、物联网、传感网等技术的发展，地理信息的应用进入了大数据时代。通过对时空大数据进行存储、管理、处理和分析，可以揭示时空大数据与现实世界对象、行为、事件间的对应关系，目前，基于时空大数据的地理计算已经应用在交通、气象、救灾、应急等研究和业务领域。

然而，由于时空大数据的海量数据特征，时空大数据的存储、管理和处理存在性能瓶颈，需要在高性能计算集群的环境支持下才能完成，学生普遍缺乏实验环境；此外，由于时空大数据计算环境配置的复杂程度，也超出了学生能够独立完成的能力范围。

针对在时空大数据的地理计算与分析的教学实验环节的困难和不足，本项目将研究构建可供学生操作的时空大数据地理计算与分析实验平台，结合已有的时空大数据平台硬件、软件和算法，为地理计算和分析实验教学提供一套基于Web环境的时空大数据计算和分析系统。本项目旨在为本院和全校师生提供一套基于时空大数据的地理计算和分析系统，能够方便、易用的操作和处理时空大数据，方便学生理解基于时空大数据的地理计算模式和分析方法，具有非常大的研究价值和应用价值。

项目将围绕学生在学习和实践时空大数据地理计算时需要学习和掌握的高性能数据存储、处理和分析过程，从数据存储模型、地理计算模型、时空分析引擎框架以及软件研发四个方面出发，采用分布式数据库存储模型与检索理论，开展并行高性能计算环境下时空大数据存储模型、时空大数据处理模型、时空数据可视化分析以及时空大数据管理引擎软件研发等方面的研究。

以本项目在实验教学中的定位，主要包含以下的研究内容：

1:典型时空大数据内容的组织和存储

以中国地质大学（武汉）广受开设的地理信息系统课程以及相关空间分析课程中涉及的知识和应用场景，研究采集和组织包括遥感影像、出租车轨迹数据、微博签到数据等内容，为地理计算与分析构建相应的时空大数据沙盘场景；研究时空大数据的存储环境，为时空大数据的分析提供技术支持。

2:时空大数据的地理计算与处理模型

深入研究分布式环境下时空大数据的地理计算与处理模型方法，在硬件平台部分，研究目前主流的HBase + Spark计算方案，提出分布式环境下时空大数据高效计算框架，能够支持地理计算和处理的并行计算环境；另一方面，我们将研究主流地理计算方法在并行计算平台下的实现，如K-Means, DBSCAN等通用地理计算模型和方法适配在本计算平台下。

3:在线操作与可视化平台

针对学生的教学与实验需求，研究方便学生使用的在线数据操作方法和可视化分析Web平台，实现网络环境下时空大数据处理过程的操作与可视化。主要包括在Web环境下实现时空大数据共享、数据服务、专题制图、模型分析、和实验分析结果展示等功能。

主要技术难点：

本项目将研究针对中国地质大学（武汉）本科教学中地理信息系统及相关空间分析课程中涉及的时空大数据场景，研究典型的时空大数据存储和组织，以及分布式环境下的数据处理、计算和分析可视化。主要技术难点如下：

1:时空大数据的存储和组织

时空大数据的采集包括遥感影像、出租车轨迹数据、微博签到数据等，这些数据存储格式和方式存在差异性，如何在分布式存储环境下对这些数据进行高效组织和管理，是本项目的一个主要技术难点。

2: 时空大数据地理计算与分析的在线处理和交互方法

本项目还将研究基于Web模式的时空大数据的地理计算与交互分析方法。研究动态准实时多维可视化方法，突破数据立方体快速建模、转换、渲染与渐进式优化传输等关键技术，使得网络环境下的Web端可视化的加载和交互更加平滑高效。

**（一）系统的结构组成及主要功能（需要给出系统的整体功能描述，并按不同的功能把一个大系统划分成相关联的多个小系统，描述每个小系统的主要功能。如果项目是小系统，则直接描述各模块功能）**

系统的结构组成及主要功能

1:时空大数据存储模块

本模块基于HBase的分布式数据库构建，其能够支持海量时空数据的存储与管理工作。该模块的主要功能是先通过建立分布式数据库下的时空数据模型，然后在此基础上构建时空索引，最后进行时空数据的存储与处理分析工作。该模块主要对数据库索引进行全生命周期的管理，包括索引的创建、更新、删除。索引的创建过程针对不同的应用需求、数据类型特点建立不同的索引方式，能够针对全文、空间、时间、属性等多个维度，建立图索引、B树索引、R树索引、Hash索引等多种索引方式。索引的更新机制使得索引能够在数据被增、删、改等操作之后进行同步更新，保证多种索引的现势性。

2:时空大数据处理与分析模块

该模块为时空大数据处理与分析功能提供具体功能实体，包括多维数据挖掘、数据聚类处理、时空变化趋势分析等操作。时空大数据的内容包含位置信息（GPS定位信息）和时间、属性等信息，位置数据主要是以点的方式存在，为了能够支持不同的计算和分析需求，每个空间点的特征将参考GeoHash的方式进行存储管理，其目标是将二维的坐标信息以一维的方式组织起来，并进行Hash排序，这样能够根据地理范围将相近空间属性的数据点集中管理起来，便于检索和查找，进而便于计算和分析处理。位置信息的挖掘包括区域动态模式，个体移动模式，位置场景内容等。区域动态模式的挖掘主要是挖掘位置区域内的地理活动指标,如区域活跃程度、活动类型、交通潮汐变化等内容；针对个体移动模式(mobility pattern,简称MP)的挖掘是以单个移动对象为观察目标,包括其在一段时间内的移动独一性、随机性、周期性、转移性、动静间歇性和移动期望性等方面。

   多维数据挖掘接口研究泛在时空数据的地理空间信息特征，实现快速的收集、筛选、分析、归纳、展现决策者所需要的信息，将数据的各个属性值加以组合，并以多维数据的形式通过图表、三维等方式用以展现数据之间的关联信息。面向位置关联的泛在大数据，分析多维数据的空间与专题特征，对多维数据进行取样、清洗、模式构建等预处理操作，并提取数据的位置与主题特征实现数据间的准确映射关联。结合空间统计方法，根据具体需求采用聚类算法、支持向量机算法、遗传算法、决策树、链接分析等数据挖掘算法，利用相关分析、方差分析、线性回归等统计方法以及深度学习方法，从不同维度、不同粒度对相关领域数据进行挖掘。

    在时空变化分析层面上，提供基于关键词、主题和时空查询分析模式。面向关键词的查询分析服务是以大数据中接收信息的内容特征为线索，不受已知信息的限制，用户可以自由的表达信息需求，专指度高，信息集中，简单易用，能够满足各层次用户的查询分析需求；由于关键词本身的多义性，因此不同文本内容往往表达的是相同的语义，把这些有着相同语义或有相关含义的词可以归纳为一个类型，可以命名为主题，整个数据的处理同样是在数据接收后，进行语义内容的分解和处理，将这些同样主题的数据信息放到相同的数据内容中进行索引和存储管理，用来支撑业务与主题相关的数据查询分析服务。基于空间和时间等上的索引信息，方便完成系列不同数据集合上的数据的快速查询和分析服务。

3: 时空大数据协同计算动态Web可视化模块

本可视化模块研究基于Web的位置网大数据时空分布及挖掘结果数据的动态准实时多维可视化方法，突破数据立方体快速建模、转换、渲染与渐进式优化传输等关键技术，使得网络环境下的Web端可视化的加载和交互更加平滑高效。研究不同形式数据的多维度交互式可视化方法，使得数据的不同属性维度得以不同的形式直观展示，并建立不同视图不同维度展示之间的联动和关联机制，提升可视分析能力。针对时空大数据，研究时空动态演化模式中对象与视觉元素间的映射，研究多尺度、多语义时空过程动态可视化分析方法，构建基于人机探索式的可视化模型。采用不同形式的可视化方法如面状动态符号、动态线形符号、动态饼图、动态直方图、柱状图、散点图、热力图等来实现相应主题时空数据的过程再现、实时跟踪和运动模拟和时空动态演化，更直观丰富地为用户展示数据挖掘结果。模块可根据不同的应用场景以及不同用户需求，为用户提供指定空间范围、指定图层名称（对应不同数据源）、指定类型（栅格/矢量）、指定坐标系统、指定样式的动态可视化资源。

   基于分布式存储和计算技术，实现大规模社交媒体数据集的快速降维、分块、和并行查询处理的方法；针对不同可视化方法（如柱状图、散点图和热力图等），提供相应的用户交互可视化操作，包括提供充分的交互操作方便用户对数据从时间、空间和宏观及微观视角进行查看的功能。通过对数据进行整理分类，对位置数据利用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位置大数据。本模块的可视化操作可分为两种主要操作

（1）基于地理数据可变比例尺展示精度的查看模块需求  
空间数据和行业数据的分布情况应该是能够根据用户需要来对观察的范围进行调整变化的，用户可以在基础简化地图上能够通过缩放操作来查看数据不同显示比例下的分布情况，要求能够并支持这些基础数据的快速显示。用户可以点选单个数据，并查看其具体的属性特征。

（2）基于时间轴的时空特征交互模块需求  
为了便于查看数据在时空上的分布特征，需要提供能够对海量数据时空分布规律的宏观查看，即在指定时间范围内的数据集在空间上的分布特征，为了能够让这个展示更加直观，需要能够具备随着时间轴的动态数据展示。

   在这种自动展示过程中，可以将每个数据集都直接显示出来，适合用来表达较小的数据集合，便于小范围的数据查看。另外，对于更大的数据集合，为了数据的展示能够更加简洁，可以采用地理范围内的颜色填充来区分不同数据集合在不同时间内的分布特征。最后，如果数据密度过高，或者数据除了数量外，更要考虑的数据的重要性，可以采用热力图示的方式将这种特征进行表达。

**（二）主要技术参数指标（指标需要是定量描述，具有可考核性**

本项目开发完成的时空大数据地理计算与分析平台，可用于多个地理信息相关课程，该平台能满足以下主要技术参数和指标：

1:能够实现时空大数据的存储和管理功能

2:能够在Web端进行操作，可以在线分析如轨迹数据、遥感数据、社交媒体数据等典型时空大数据

3: 系统具有可靠性、易用性、可维护性和可扩展性

4:编写一套软件操作使用手册