WEB – ТЕХНОЛОГИЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ПРОСТРАНСТВЕННО-ВРЕМЕННЫХ ТЕМАТИЧЕСКИХ ДАННЫХ[1]

В.В. Парамонов, Р.К. Федоров, Г.М. Ружников, П.В. Белых
ФГБУН Институт динамики систем и теории управления Сибирского отделения РАН,
г. Иркутск
e-mail: {slv, fedorov, rugnikov, polina}@icc.ru

Аннотация

В работе предлагается Web-технология для статистического анализа распределенных пространственно-временных данных, которая позволяет сократить временные и трудовые затраты обработку, анализ данных и публикацию результатов.

Введение

Развитие информационных технологий позволяет перевести исследовательскую работу на новый уровень более тесной интеграции данных и методов обработки и анализа. Реализация программной интеграции, базирующейся Web - технологий, дает возможность многократного применения различными пользователями как данных, так и методов обработки. Одним из часто используемых методов обработки данных является статистический анализ. Зачастую анализируемая статистическими методами информация связана с какими либо территориями, данные о которых включаются в базовые пространственные данные (БПД). Соответственно использование единых БПД могут позволить сопоставлять различные данные по территориям. Такого рода данные используется в статистических отчетах, например, документах, представляемых Федеральной службой государственной статистики, внутренних базах данных различных ведомств (Росгидромет, Министерство здравоохранения и социального развития и т.п). Поэтому является актуальным создание технологии статистического анализа, позволяющей использовать БПД для ввода данных и проводить статистический анализ с учетом БПД. Пространственная информация предоставляет возможность наблюдения и анализа данных во времени и с учетом их положения, что открывает более широкие возможности для исследования и дальнейшей визуализации информации, например, в виде тематических карт.

Ввиду того, что данные, как правило, являются распределенными, то наиболее эффективным способом взаимодействия между ними представляется взаимодействие через протоколы сетей Интернет/Интранет. Сами же технологические решения по вводу, обработке данных и выводу результатов актуально реализовывать в виде сервисов (SOA, Service oriented architecture) Web-среды. Такой способ реализации позволяет обеспечить достаточно высокий уровень интероперабельности, предоставляет возможность более гибкого использования ресурса на различных платформах [1].

Технологические решения

В работе рассматривается работа с реляционными данными. Их сбор и обработка осуществляется посредством возможностей Геопортала ИДСТУ СО РАН [2]. Это позволило

[1] Работа частично поддержана РАН (ФНМ-48), РФФИ (грант 14-47-04125), Советом по грантам Президента Российской Федерации для государственной поддержки ведущих научных школ (НШ-5007.2014.9).

реализовать систему как распределенную, а взаимодействие с пользователем организовать посредством стандартного WEB-браузера.

Технология статистического анализа данных состоит из следующих этапов:

- сбор данных и их загрузка;
- нормализация;
- пространственно-статистический анализ.

Рассмотрим реализацию этапов данной технологии подробнее.

Сбор, загрузка данных

Анализируемая информация может экстрагироваться из различных источников. Информация может быть представлена в виде реляционных таблиц или серии GRID файлов формата GeoTIFF [4]. Для реляционных данных в геопортале реализован сервис ввода и редактирования реляционных данных, содержащих пространственные атрибуты и обеспечивающие возможность многопользовательской работы. Пользователь может создать и вводить данные в собственную таблицу с необходимой структурой, где в качестве атрибутов можно указать ссылки на БПД. Сервис позволяет отображать вводимую информацию в виде таблицы и карты. Для создания GRID файлов в рамках геопортала разработаны ряд сервисов, например апроксимация точечных данных, расчет плотности точечных и линейных объектов и т.д.

Нормализация данных

При импорте данных из других источников часто требуется их нормализация, т.е. приведение к простым типам данных, представленных в полях таблицы, а также привязка значений к иерархическим справочникам (если это определено структурой таблицы). Для решения поставленной задачи авторами на платформе Node JS разработана программная библиотека. Библиотека обеспечивает полуавтоматическую нормализацию данных. При этом владелец данных проводит их структурное описание – для каждого столбца обрабатываемой таблицы указывается является ли поле справочным значением (адрес, название растения, насекомого, код услуги и т.п.), единицей измерения (дюйм, градус, м3 и т.п.), текстом и т.п. Итоговая нормализированная таблица содержит значения идентификаторов справочников, в том числе БПД. Это позволяет проводить различного рода аналитические операции над данными с учетом их пространственных характеристик.

Статистический анализ пространственно-временных данных

Первичная статистическая обработка количественных признаков, а также оценка их значимости являются основой для построения математических моделей, описывающих взаимосвязь характеристик в точке. Данный подход является одним из основополагающих, например, при обработке медицинской информации [5].

Для реализации статистического анализа разработаны WPS-сервисы (стандарт OGC [6]), обеспечивающие проведение анализа через сеть Интернет/интранет. На текущий момент производится первичная статистическая обработка - поиск коэффициентов регрессии и корреляции. Это позволяет сделать первичные оценки о зависимости различных данных в пространственно-временном разрезе.

Результаты проведенного статистического анализа могут быть представлены как в виде чисел, так и в виде тематических карт. Генерация тематических карт позволяет визуализировать и сделать более наглядными результаты исследования.

На вход WPS-сервисам подается таблица с двумя атрибутами. Значения атрибутов формируются запросом из разных таблиц на основе соответствия пространственно-временных характеристик, либо две серии GeoTIFF файлов, где соответствие значений определяется положением ячеек.

В случае работы с таблицами реляционных данных пользователь имеет возможность сузить объем необходимой информации через использование различных фильтров.

Выводы

По результатам работы предложена Web-технология проведения статистического анализа пространственно-временных тематических данных. Практическое применение технологического решения позволяет проводить анализ зависимости различных тематических данных, имеющих пространственную и временную привязку. При работе с ресурсом, реализующим технологию статистической обработки данных, имеющих пространственную привязку, у пользователя исключаются потребности в поиске специализированного программного обеспечения, его установки, настройки и изучения; загрузке базовых пространственных данных, различных справочников (классификаторы животного мира, почв, национальностей, профессий и т.п.); разработки механизмов ввода данных и их отображения; способов операций над данными и отображения результатов. Следующим шагом в развитии технологии является разработка математических моделей, обеспечивающих анализ данных исходя из множества параметров.

ЛИТЕРАТУРА

1. Tom Narock et al. A provenance-based approach to semantic web service description and discovery [Электронный ресурс]. - 2014 г. Режим доступа- http://dx.doi.org/10.1016/j.dss.2014.04.007.
2. Бычков И.В. Создание инфраструктуры пространственных данных в управлении регионов/, И.В. Бычков, В.М. Плюснин, Г.М. Ружников [и др.] // География и природные ресурсы. - 2013. - № 2. - С. 145 - 150.
3. OGC OpenGIS Web Processing Service.[Электронный ресурс] - Режим доступа: http://www.opengeospatial.org/standards/wps. - Дата доступа: 08.09.2014.
4. N. Ritter & M. Ruth The GeoTiff data interchange standard for raster geographic images // International Journal of Remote Sensing -Volume 18, Issue 7, 1997, pp. 1637-1647.
5. Юнкеров В.В., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований. - СПб.: ВМедА, 2002. - 266 с.
6. Anthony M. Castronova, Jonathan L. Goodall, Mostafa M. Elag Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard // Environmental modelling & software. Vol. 41. 2003. pp. 72-83.