СЕКЦИЯ 5
НАУКА О ДАННЫХ

БОЛЬШИЕ ДАННЫЕ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ: ПРОБЛЕМЫ И ВОЗМОЖНОСТИ

С.Б. Попов

Институт систем обработки изображений РАН, Самара, Россия

Обсуждаются характерные особенности данных дистанционного зондирования Земли в контексте больших данных и возникающие при этом новые возможности, проблемы и направления исследований. Рассматриваются преимущества использования методологии больших данных при создании систем обработки данных дистанционного зондирования Земли. Отмечается, что это решение обеспечивает прозрачное наращивание функциональности подобных систем и улучшение их качества, формирование новых интеллектуальных свойств.

Ключевые слова: большие данные, дистанционное зондирование, распределённая обработка данных.

Данные, получаемые при дистанционном зондировании Земли (ДЗЗ), в настоящее время характеризуются существенным увеличением объема информации, а также наращиванием скорости её поступления. Только дюжина спутников формирует более 2 терабайтов данных в день или пол петабайта в год, причём два из них обеспечивают почти половину этого объёма. Сомнений в том, что это большие данные здесь, конечно, не возникает.

Обсуждая третью характеристику больших данных дистанционного зондирования, разнообразие данных (Veriety), следует выделить два аспекта: разнообразие источников и разнообразие форматов их представления.

Среди источников данных можно отметить:

- спутниковые данные (в перспективе – группировки микро-спутников);
- данные аэронаблюдений (в перспективе – группировки БПЛА);
- геоинформационные системы;
- научные, правительственные и коммерческие базы данных и архивы;
- данные, формируемые независимыми наблюдателями (Citizens Science);
- данные социальных сетей.

Имеющееся разнообразие форматов изображений, получаемых в процессе наблюдений, дополняется разнообразием типов данных, которые необходимо совместно использовать при решении задач получения нового знания с использованием данных ДЗЗ:

- данные высокого разрешения;
- гиперспектральные данные;
- видеоданные космического наблюдения и аэросъёмок;
- векторные данные;
- данные в виде графов;
- различные виды структурированной информации;

- неструктурированная информация произвольного вида.

Получение качественно новых знаний о подстилающей поверхности и расширение круга решаемых прикладных задач диктует необходимость привлечения большого объема дополнительной, часто неструктурированной информации для обработки, дешифрирования и анализа. При этом процесс формирования процедур обработки/анализа носит исследовательский характер, при решении конкретных задач обработки данных ДЗЗ и связанной с ними информации зачастую необходимо реализовывать многовариантный подход, при котором используются различные методики формирования результатов с их последующим анализом.

В докладе выделены и обсуждаются несколько направлений исследований в данной области:

- Многоуровневый анализ: одновременное использование всех данных (данные с различным пространственным и спектральным (гиперспектральные) разрешением, радиолокационные данные, наземные наблюдения).
- Мультимодальная обработка (совместная многоэтапная обработка, использующая различные подходы): преимущества взаимодополняемости разнородных методов коллаборативной обработки.
- Мультитемпоральный анализ: возможность новых видов анализа, учитывающих как долговременные (урбанизация), так и циклические (сельское хозяйство) изменения, причём с возможным наличием нерегулярности наблюдений.
- Возможности получения изображений с малым периодом обновления: новые инкрементальные методы обработки и анализа, позволяющие адаптировать классификации или индексы, извлечённые из изображений, обновлять тематические модели (дрейф понятий) без необходимости повторной обработки всего объёма информации при получении её новой порции.
- Семантический разрыв с базовыми знаниями: отсутствие соответствия между информацией низкого уровня (т.е. автоматически извлекаемой из изображений) и информации высокого уровня (знаний о подстилающей поверхности, объектах на ней) преодолевается использованием кластеризации с последующей интеграцией человека-эксперта в процесс обучения классификаторов.
- Качество данных и робастность алгоритмов: разработка методов оценки и корректировки ошибок или неточностей в данных, разработка алгоритмов, способных учитывать ошибки, неточности в данных, несоответствия и несогласованность в базах знаний.
- Разнородные ресурсы: объединение всей информации, доступной на обследованной территории, независимо от источника данных.
- Масштабируемость: необходимость переосмысления алгоритмов и принципов создания программного обеспечения при переходе к распределённой или параллельной обработке.

Прямое использование программных платформ Big Data для организации обработки и хранения данных дистанционного зондирования наталкивается на значительное число проблем, связанных с тем, что технологии Big Data в первую очередь ориентированы на обработку текстовой информации. Отдельные проекты использования платформы Hadoop для хранения и обработки набора изображений только подтверждают это.

Однако, общие принципы и методологию распределённого хранения и параллельной обработки данных полезно использовать при разработке систем обработки и хранения данных ДЗЗ.

В основе технологий «больших данных» лежат относительно простые принципы [1]:

- Разделяй и властвуй («Divide and Conquer!») – тотальное распараллеливание данных и вычислений с обеспечением отказоустойчивости как на уровне хранения, так и при обработке.
- Обрабатываем там, где храним («Move Code to Data!») – данные в процессе обработки не перемещаются, наоборот, процедуры обработки доставляются к данным и запускаются на вычислительных ресурсах распределённых систем хранения.
- Данные навсегда («Data Are Forever!», No UPDATE operations) – данные в процессе обработки не изменяются и не удаляются, результаты просто сохраняются здесь же в виде нового набора.

При творческом развитии и адаптации этих принципов применительно к задачам разработки распределённых систем обработки и хранения данных ДЗЗ необходимо учитывать особенности этих данных, представленных преимущественно в виде изображений, т.е. пространственно зависимых данных. Простое разбиение крупноформатного изображения на отдельные фрагменты, которое реализуется в системах на базе платформы Hadoop, в большинстве случаев не применимо, оно должно быть программно-контролируемым, особенно при использовании в процессе обработки локальных операций на основе скользящей окрестности. Решением в данном случае является использование концепции Data-as-a-Service.

Основной структурной единицей хранения в системе является распределенное изображение, которое доступно пользователю как набор взаимодействующих сервисов хранения, т.е. каждый фрагмент распределенного изображения представляется в виде отдельного сервиса (фрейм-сервиса) со своим уникальным именем в иерархии изображений-сервисов. Декомпозиция данных при формировании федерации фрейм-сервисов одного изображения выполняется на основе принципов, изложенных в работе [2]. Реализация этих принципов обеспечивает отказоустойчивость при распределенном хранении данных ДЗЗ. В соответствии с ранее изложенными принципами фрейм-сервис, помимо интерфейса доступа к данным, реализует интерфейс, обеспечивающий получение и выполнение задания на обработку контролируемых данных. Причём, в процессе такой обработки данные не изменяются, а создаётся новое распределённое изображение, фрагменты которого могут быть не согласованы между собой в части обеспечения последующей сбалансированности распределенной обработки. Тем не менее, наличие "интеллекта" у фрейм-сервиса и связей с соседями позволяет выполнить все необходимые для такого согласования действия автономно в фоновом режиме, предоставляя пользователю необходимые ему функции (чаще всего это визуализация) и данные непосредственно сразу после создания. Таким образом, реализуется ориентированный на данные подход к организации вычислений, при котором фрагменты данных заранее распределены по узлам хранения распределенной системы, а в процессе обработки вместо данных пересылаются процедуры их обработки.

Предлагаемая архитектура позволяет организовать технологию обработки данных ДЗЗ с использованием метапрограммирования и мультимодального подхода [3], при котором

пользователь может указывать не конкретные операции обработки, а некоторые обобщающие этапы, формулировать требуемые цели обработки и намечать пути их реализации. Сервис-ориентированный подход дает возможность реализовать эффективные решения в ранее рассмотренных направлениях исследований.

Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект № 15-29-07077).

Литература

1. Popov S.B. The Big Data methodology in computer vision systems. Proceedings of Information Technology and Nanotechnology (ITNT-2015), CEUR Workshop Proceedings, 2015, vol. 1490, pp. 420-425. doi: 10.18287/1613- 0073-2015-1490-420-425.
2. Kazanskiy N.L., Popov S.B. Distributed storage and parallel processing for large-size optical images. Proc. SPIE 8410, 2012, pp. 84100I-1-84100I-11. doi:10.1117/12.928441.
3. Kazanskiy, N.L., Popov S.B. Integrated Design Technology for Computer Vision Systems in Railway Transportation. Pattern Recognition and Image Analysis, 2015, 25(2), pp. 215-219. doi: 10.1134/S1054661815020133.