Астрономия в эпоху Big Data

О.С. Бартунов, С.В. Карпов

Государственный астрономический институт им. Штернберга при МГУ,
Специальная Астрофизическая Обсерватория РАН
obartunov@gmail.com, karpov.sv@gmail.com

Аннотация

В докладе рассматриваются проблемы, встающие перед современной астрономией в связи с лавинообразным ростом объёмов научной информации. Подчёркивается, что изменения в подходах к хранению и анализу этих данных становятся качественными, а не только количественными — астрономия сталкивается с известной в других областях проблемой больших данных (Big Data). Помимо разработки стандартов и методов хранения, поиска, обмена и анализа такой информации необходимо также и изменение парадигмы астрономического образования, которая должна ориентироваться в том числе и на обучение студентов основам астроинформатики.

1. Введение

Широко известный закон Мура гласит, что мощности вычислительных ресурсов компьютеров удваиваются каждые 18 месяцев. В то же время, объёмы данных, получаемых современными научными экспериментами, растут существенно быстрее, удваиваясь каждый год. На масштабе 10 лет разница становится огромной — объём данных вырастает на порядок больше доступных для их обработки ресурсов! Более того, пропускная способность каналов связи растёт ещё медленнее, возрастая за 10 лет лишь в разы. Такой быстрый рост объёмов данных становится уже не только количественной, но и качественной проблемой. С ними уже невозможно работать так, как было привычно астрономам двадцать и даже десять лет назад.

Причинами взрывного роста данных в последние десятилетия являются как эволюционное изменение используемых астрономами приёмников, так и появление существенно новых режимов работы. Размеры астрономических приёмников излучения — главным образом, ПЗС-матриц — растут существенно быстрее, чем размеры и даже количество самих телескопов. А рост числа пикселей в ПЗС-приёмнике — это пропорциональное увеличение объёма получаемой с него информации. Повышение качества сопутствующей электроники, выражающееся в понижении шумов считывания и повышении доступного динамического диапазона, также ведёт к увеличению размеров каждого кадра и, кроме того, позволяет получать изображения научного качества за меньшее время — то есть делать за ночь больше снимков неба. Появляются и научные задачи, требующие получения огромного количества последовательных снимков одного и того же объекта с малыми экспозициями — это и поиск быстрой переменности пульсаров, рентгеновских двойных систем, так и задачи достижения дифракционного качества изображений на наземных инструментах (спекл-интерферометрия или Lucky Imaging).

Другим источником всё возрастающего потока научной информации становятся многочисленные систематические обзоры небесной сферы, благодаря которым астрономия вступает в совершенно новую эру. Нельзя сказать, что такие обзоры не проводились ранее — достаточно вспомнить знаменитые оптические атласы северного и южного неба, снимавшиеся ещё в доцифровую эпоху на фотопластинки в Паломарской и Англо-Австралийской обсерваториях. В 90-е годы эти фотопластинки были оцифрованы с использованием специализированнных сканеров-микроденситометров и преДОСтавлены в пользование всем заинтересованным астрономам как Цифровой Обзор Неба — Digitized Sky Survey. Электронный атлас представлял собой 102 компакт-диска и был также доступен через интернет.

Одного такого обзора, естественно, недостаточно для полноценного понимания того, как устроена вселенная. Начиная с 90-х годов проводится всё больше и больше экспериментов, включающих в себя обзоры небесной сферы в различных диапазонах длин волн и с различными задачами. Так, целью запуска спутника Hipparcos, работавшего с 1989 по 2003 г, было построение предельно точного астрометрического каталога наиболее ярких звёзд, содержащего точную информацию об их положениях, собственных движениях и параллаксах — то есть о полной

Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012), Санкт-Петербург, Россия, 2012.

трёхмерной картине наших звёздных окрестностей. Обзор Two Micron All-Sky Survey (2MASS), проведённый на рубеже 21-го века, стал первым крупным инфракрасным атласом неба, проводящийся с 2003 года спутниковый эксперимент GALEX — первым и самым крупным ультрафиолетовым. Вот уже более 10 лет одним из самых результативных астрономических экспериментов, и одним из наиболе цитируемых (а цитирование в науке — показатель успешности и востребованности результатов) экспериментов в физике вообще является до сих пор проводимый Слоановский обзор неба (Sloan Digital Sky Survey, SDSS) — проект по построению многоцветного атласа неба в спектральных фильтрах, оптимизированных для упрощения разделения различных классов астрономических объектов по их цветам для последующего отбора кандидатов для получения их спектров.

Объёмы данных, получаемых в этих обзорах, показательно растут со временем — от нескольких десятков гигабайт в DSS до десятка терабайт в 2MASS и нескольких десятков — в GALEX и SDSS.

Обзоры, подобных DSS, 2MASS, SDSS — однократные, они пригодны лишь для изучения объектов, не меняющих свои свойства и не движущихся по небу. Проводятся и готовятся и другие обзоры, и они ещё более объёмны в плане данных — мониторинговые обзоры неба, направленные на поиск подобных объектов — как далёких космических транзиентов, гамма-всплесков и сверхновых звёзд, так и быстродвижущихся объектов в Солнечной системе, которые могут оказаться опасными для Земли астероидами.

Уже запущена первая серия четырёхтелескопного эксперимента PanSTARRS, который предполагает многократный регулярный обзор всего неба с проницанием, сравнимым с SDSS — то же самое, но не за десять лет, а раз в несколько дней! И все получаемые данные должны немедленно обрабатываться, чтоб обнаруживать потенциально опасные астероиды. Естественно, и используемая для этого аппаратура, и получаемые объёмы данных, и методики их анализа очень сильно поменялись за прошедшие годы. Так, изображения на PanSTARRS получают с помощью ПЗС-матрицы, имеющей более миллиарда пикселей — больше одного гигапикселя (а обычные фотоаппараты, как и приёмники "обычных" телескопов, имеют размеры всего лишь в несколько, до пары десятков, мегапикселей!). А суммарный объём информации, который будет получен в этом обзоре, составит до 40 петабайт — действительно астрономическое количество по сравнению с предшествующими обзорами.

Ещё более амбициозный проект, Большой Синоптический Обзорный Телескоп (Large Synoptic Survey Telescope), который заработает через несколько лет, обещает дать до 100 петабайт научных данных, благодаря приёмнику размером в 3.2 гигапикселя, который будет давать изображения каждые пару десятков секунд в течение многих лет!

Немаловажным является и то, что всю получаемую в таких обзорах информацию надо сразу же обрабатывать — это нельзя отложить на потом, ведь новые кадры с телескопов получаются непрерывно, и откладывать их "на потом" некуда.

Естественно, такой взрывной рост объёмов научной информации требует и изменения подходов для её хранения, как и для работы с ней.

Ситуацию, когда объёмы и темпы поступления данных становятся настолько большими, что приходится менять устоявшиеся практики работы с ними, называют проблемой больших данных (Big Data).

2. Big Data в астрономии — проблемы и подходы

Рост объёмов научной информации в астрономии связан как с количественными — возрастанием числа телескопов и размеров приёмников на них, так и с качественными причинами — новыми режимами наблюдений, проведением крупномасштабных и многократных обзоров небесной сферы в различных диапазонах, проведением крупномассштабных численных экспериментов по моделированию эволюции вселенной в целом, итд.

Первой из встающих перед астрономами проблем является сохранение всей этой информации в виде, пригодном для дальнейшего использования. Проблема эта общая для многих отраслей как науки, так и бизнеса, и пути её решения достаточно хорошо разработаны — это современные системы управления базами данных. Современные традиционные СУБД плохо приспособлены для Big Data, так как они разрабатывались для старой архитектуры вычислительных машин, в них используются алгоритмы работы с данными, ориентированные на медленные хранилища, небольшую оперативную память, одноядерный процессор, в то время как современные системы — это кластеры многоядерных серверов с большой оперативной памятью и практически неограниченными дисковыми подсистемами. Попытки "приспособить" эти СУБД к новым условиям успехов не приносят, требуется разработка СУБД "с нуля".

Обычно, из-за низкой производительности современных систем, исходные данные научных наблюдений хранятся вне каких-либо СУБД, и только метаданные индексируются в базе данных. Для доступа и обработки исходных данных научным коллективам приходится разрабатывать свои программные системы под каждую конкретную задачу. При таком подходе очень трудно поддерживать версионность данных, историю их изменений, получение научных

результатов из "сырых" данных, что нарушает один из основных принципов науки — повторяемость научных результатов.

Еще одна особенность современных научных экспериментов — это сочетание распределенного хранилища данных с необходимостью доступа к высокопроизводительным вычислительным комплексам для получения научных данных из результатов эксперимента и их анализа. Такие вычислительные комплексы в настоящее время в основном используются для решения расчетных задач, не требующих работы со сверхбольшими данными. Однако, и в задачах численного моделирования появились требования к возможности сохранения текущего состояния в СУБД, например, расчет космологической эволюции Вселенной требует сотни гигабайт для сохранения одного "слепка" Вселенной. Подобные расчеты ведутся на распределенных кластерах с тысячами процессоров и возможность работы с такими данными в СУБД позволяет проследить историю эволюции отдельных объектов Вселенной (частицы, звезды, галактики, скопления галактик…).

Сложившаяся ситуация в больших научных проектах была оценена ведущими учеными из разных наук,представителями коммерческих компаний и разработчиками в области СУБД (систем управления баз данных) на серии конференций XLDB 2007 и 2008 гг, в результате чего возник проект SciDB под руководством профессора MIT Майка Стоунбрейкера и его коллег из крупнейших университетов США. Основная цель проекта — разработка в кратчайшие сроки СУБД для нужд больших научных и промышленных проектов, в которых требуется анализ сверхбольших объемов данных (сотни и тысячи петабайт), масштабируемая на тысячи серверов. База данных для научных данных SciDB призвана обеспечить полный цикл работы с данными — хранение, обработка и анализ. Однако, в настоящее время пока не существует полноценной СУБД для астрономических данных, готовой для промышленного использования. В частности, это объясняется спецификой астрономических данных.

Важной особенностью большей части астрономических данных является сферическая система координат, в которой они задаются (звёзды наблюдаются в проекции на небесную сферу). Невозможность однозначной проекции с сохранением отношения близости точек единичной сферы ни на прямую, ни на плоскость существенно затрудняет построение поисковых индексов, оптимальных для основных задач, встающих перед астрономами — поиска объектов по неточно заданным координатам, что сводится к запросу в окрестности заданной точки, и поиска проявлений одного и того же объекта в различных диапазонах или в разные эпохи наблюдений, что сводится к задаче поиска "ближайших соседей" (k nearest neighbors, KNN) в многомерном пространстве. Возможными решениями являются либо отказ от сохранения отношения близости (что приводит к возрастанию вычислительной сложности и нагрузки на жёсткие диски из-за нелокальности хранения пространственно близких данных), либо переход к большей пространственной размерности — вложение единичной сферы в трёхмерное пространство, что в полтора раза увеличивает нужный для хранения позиционной информации объём памяти, усложняет расчёты расстояний и т.д., но позволяет использовать оптимизированные для поиска структуры данных вроде многомерных деревьев.

Астрономические данные (объём которых, к слову сказать, уже превысил петабайт) совершенно разнородны — они получались и получаются на различных инструментах по различным методикам, различные каталоги содержат совершенно разную информацию о каждом из объектов. Более того, данные разбросаны по различным институтам, обсерваториям и датацентрам — давно прошло то время, когда можно было скачать или получить по почте все нужные для работы каталоги, развернуть их на одном компьютере и работать с ними. Важной проблемой становится работа с разнородными и распределёнными данными — и даже сам по себе поиск информации об этих данных! Типичной непростой задачей учёного становится сбор сведений о том, кто, когда и на каком инструменте наблюдал интересный ему участок неба — зачастую не нужно подавать заявку на крупный телескоп, проходить конкурсный отбор, ждать подходящей фазы луны и хорошей погоды для наблюдений, ведь нужный объект уже наблюдался кем-то с другой, скорее всего, целью, и нужная информация давно хранится в архивах и публично доступна (по сложившейся на крупных обсерваториях практике данные являются собственностью автора лишь первые несколько лет, а затем становятся общедоступными) — нужно лишь знать о том, что она есть! Ну и, естественно, иметь возможность получить её в каком-то пригодном для анализа универсальном формате. Решением для таких ситуаций призвана стать всемирная инициатива по выработке универсальных стандартов аннотирования, поиска и доступа к астрономической информации — Виртуальная Обсерватория. В её рамках создаются централизованные репозитории хранения информации об архивах научных данных — реестры, позволяющие, к примеру, найти все сервисы, обеспечивающие доступ к каталогам или атласам неба по стандартным протоколам, а затем отобрать из них те, которые содержат информацию об интересующей учёного области неба.

Такой подход к астрономическим данным уже является новым — он не предполагает скачивания всех нужных каталогов и локальной работы с ними для выбора отдельной площадки. Скачать

современные каталоги зачастую невозможно — они занимают десятки терабайт, не хватит ни пропускной способности каналов связи, ни доступного обычному учёному объёма дискового хранилища. Достаточно скачать лишь их небольшой кусок с информацией о малой области неба. Однако, порой научные исследования требуют работы и со всем большим каталогом разом — к примеру, если нужно сравнить оптический каталог с каталогом в рентгеновском диапазоне, выбрав все объекты, обладающие одновременно заданными свойствами в обоих диапазонах (и, естественно, видимые в обоих диапазонах одновременно). Скачать оба каталога для локальной работы невозможно — но можно написать программу (скажем, SQL-запрос) и запустить её в большом дата-центре, хранящем оба этих каталога одновременно! Это ещё более высокоуровневый подход к работе с астрономическими данными, и многие современные архивы и эксперименты дают пользователям возможность загружать и исполнять произвольные SQL-запросы в своих СУБД через CAS-интерфейсы (Catalogue Access System).

Характерной чертой нынешней ситуации в астрономии является то, что данных становится настолько много, и они настолько сложны, что сами по себе они перестают представлять ценность, их нет смысла закрывать. Единственно важным становится знание, как именно анализировать эти данные для получения научно значимого результата — более того, одни и те же данные можно анализировать совершенно по-разному, искать в них ответы на совсем непохожие научные вопросы, большая часть которых, скорее всего, и в голову не приходила авторам исходного эксперимента, в рамках которого они были получены. Таким образом, возникла новая модель научных публикаций, когда данные становятся доступными раньше, чем астрономы опубликуют научные статьи, содержащие анализ этих данных.

Одной из основных задач, встающих перед учёными при анализе данных современных обзорных экспериментов, становится выделение объектов, интересных настолько, что для их понимания нужны дополнительные наблюдения. В случае мониторинговых обзоров это могут быть транзиентные — вспыхивающие — объекты, которые появляются на небе ненадолго, такие, как космические гамма-всплески или взрывы сверхновые звёзды. Важны и движущиеся объекты — астероиды, которые могут оказаться потенциально опасными для жизни на Земле, кометы. Выделять такие объекты нужно как можно быстрее — на то они и транзиентные, появляющиеся на небе лишь на короткое время или быстро меняющие своё положение — и, с другой стороны, выделять нужно лишь самые интересные или потенциально опасные из них. Действительно, современные обзоры, подобные PanSTARRS, благодаря широким полям зрения и огромным размерам детекторов наблюдают сотни тысяч и миллионы объектов каждую ночь, тогда как проведение сколь либо детальных исследований на "обычных" телескопах возможно в лучшем случае для десятков из них.

Современные методики поиска новой информации в базах данных — датамайнинга — призваны решать в том числе и такие задачи. Кластеризация, классификация, многомерные регрессии, поиск "аномальных" объектов — все эти операции на огромных объёмах данных — а современные обзорные каталоги включают в себя миллиарды записей, каждая из которых может иметь сотни полей с информацией! — становятся непростой вычислительной задачей

3. Big Data и астрономическое образование

Столь бурное развитие технологий научного эксперимента, которое привело к экспоненциальному росту данных, получаемых от инструментов, компьютерных симуляций и сенсорных сетей, привело к необходимости создания инструментов и технологий, с помощью которых ученые могли бы проводить свои исследования.

В дополнение к экспериментальному, теоретическому и вычислительному методам познания мира буквально в последнее десятилетие в науку пришел новый метод — поиск закономерностей в огромных массивах данных. Jim Gray считается отцом-основателем этого направления. Поиском закономерностей астрономы занимались и раньше, новое здесь "огромные массивы данных"! Речь идет о многих петабайтах данных, которые уже доступны сегодня, и сотнях петабайтах ожидающих нас завтра.

Каким образом увеличение количества информации привело к появлению новой парадигмы? Открытия на кончике пера (на экране компьютера) случались в астрономии неоднократно и раньше, а количество данных влияет только на время получения результатов. Изменяется информационная архитектура науки, компьютерный мир становится все более распределенным, параллельным и многоядерным, чтобы уметь работать с такими объемами данных, но так ли многое изменилось, чтобы говорить о новой парадигме? На наш взгляд, основное что изменилось — это требование машинного доступа к данным!

Современные интерфейсы рассчитаны на интерактивную работу, когда ученый должен сам определить источник данных, способ доступа к этим данным (списаться с владельцем и получить имя пользователя и пароль, например). Однако, попробуйте ввести через веб-форму данные о тысяче звезд и нажать тысячу раз кнопку! А если вам нужно получить данные о миллиардах звезд? До появления Виртуальной Обсерватории, которая в сущности является реализацией Сервисно-

Ориентированной архитектуры в астрономии, астрономы-умельцы писали скрипты, которые могли выделять из веб-страниц нужную информацию и автоматически запрашивать базы данных. Понятно, что все это работало очень ненадежно и неудобно. Сейчас уже можно писать программы, которые могут по унифицированным официальным интерфейсам Виртуальной Обсерватории запрашивать данные. Большинство крупнейших астрономических центров так или иначе поддерживают подмножество запросов ВО. Однако, унифицированный API к ресурсам — это только самая простая проблема! Прежде надо решить задачу нахождения этих ресурсов и тут уже вовсю стает задача семантического описания ресурсов, задания семантических запросов и принятия решений. Астроном тем или иным способом в конце концов сможет разобраться в данных, например, через переписку с владельцем данных, но сколько времени это может занять! А как это сделать вашему программному агенту (программе), которому была поставлена задача поиска данных, и на каком языке эту задачу поставить? Каким образом этот агент сможет отыскать ресурсы и решить, что они подходят для вашей задачи? Пример научной задачи: найти все Сверхновые звезды, которые вспыхнули не в галактиках. Эту задачу сейчас можно решить обычным путем, однако, это будет не масштабируемое решение, т.е. для нескольких тысяч Сверхновых можно все сделать вручную, но через несколько лет, когда Сверхновые будут регистрироваться раз в пол-минуты, это будет уже нереально.

Вот так простое требование машинного доступа к данным приводит к появлению новой парадигмы научного познания. Здесь слово "научное" очень важно, оно означает соблюдение принципа науки — принципа воспроизведения научных результатов, который в наше переходное время не соблюдается.

Центром новой парадигмы является семантика межмашинного взаимодействия, все остальное тоже важно, но не так принципиально. Однако, существует большая проблема современной астрономии — до сих пор нет более-менее полной онтологии астрономии. Это очень сложная задача, которая требует координации усилий многих астрономов, и требуется новое поколение информационно-подкованных астрономов, которые смогут работать на стыке дисциплин (астрономии и информационных технологий). Чтобы новое поколение астрономов появилось, необходимо внести изменения в программу обучения — добавить курсы, ориентированные на работу с данными: описание данных, хранение данных, обработка данных, организация доступа к данным, визуализация данных, статистические методы анализа данных, и т.д. Так же как в биологии существует отдельная дисциплина — биоинформатика, так и в астрономии необходим курс астроинформатики, который и будет объединять астрономию с прикладной вычислительной математикой, учить студентов методологическим основам работы с данными в новых условиях всемирной сети астрономических архивов, организованных в единых стандартах, которые предоставляют доступ ко всем наблюдениям, практически неограниченные компьютерные ресурсы для обработки и анализа данных, и только фантазия ученого и его навыки работы с данными являются единственными факторами, который влияют на эффективность научного исследования.

Astronomy in the Big Data era

O.Bartunov, S.Karpov

The talk briefly reviews the problems arising in modern astronomy due to exponential growth of the volume of scientific data. We stress that these problems lead to the qualitative changes in methods of both the storage and analysis of these data — astronomy meets the well-known in enterprise world Big Data problem. To solve it, the changes of the paradigm of astronomical education, which should pay more attention to astroinformatics, are needed, alongside with development of standards and methods of keeping, searching, exchanging and analysing the information itself.