

Московский государственный университет имени М.В. Ломоносова

На правах рукописи

ПОЙДА Алексей Анатольевич

ПОИСК СЦЕНАРИЕВ СОБЫТИЙ НА ГРИДЕ ИСТОЧНИКОВ
ДАННЫХ ОБ ОКРУЖАЮЩЕЙ СРЕДЕ

Специальность 05.13.11 – математическое и программное
обеспечение вычислительных машин, комплексов и компьютерных
сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2009

Работа выполнена на кафедре системного программирования факультета вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова.

Научные руководители:

кандидат физико-математических наук
Жижин Михаил Николаевич;

кандидат физико-математических наук
доцент
Гуляев Анатолий Викторович

Официальные оппоненты:

доктор технических наук
профессор
Модяев Алексей Дмитриевич;

доктор физико-математических наук
Родкин Михаил Владимирович

Ведущая организация:

***Институт системного
программирования Российской
академии наук.***

Защита диссертации состоится 20 февраля 2009 г. в 11 часов на заседании диссертационного совета Д 501.001.44 в Московском государственном университете им. М.В. Ломоносова по адресу: 119991, ГСП-1, г. Москва, Ленинские горы, МГУ им. М.В. Ломоносова, 2-ой учебный корпус, факультет ВМиК, ауд. 685.

С диссертацией можно ознакомиться в библиотеке факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова.

С текстом автореферата можно ознакомиться на портале ВМиК МГУ им. М.В.Ломоносова <http://cs.msu.su> в разделе «Наука» – «Работа диссертационных советов» – «Д 501.001.44».

Автореферат разослан « 19 » января 2009 г.

Ученый секретарь
диссертационного совета
профессор

Трифонов Н.П.

Общая характеристика работы

Объект исследования и актуальность темы

Интенсивное развитие технологий распределенных вычислений и инфраструктуры Грид требуют новых методов параллельной обработки и поиска данных в распределенных системах управления базами данных и базах знаний, а также создания новых инструментальных средств разработки интеллектуальных систем для поддержки принятия решений и обучения. Актуальность этих проблем в настоящее время становится особенно значимой: при интерактивном анализе данных в распределенных сверхбольших базах данных по окружающей среде; для выбора реалистичных сценариев погоды в обучающих программах и виртуальных тренажерах; для оценки степени воздействия окружающей среды на техногенные системы (транспорт, коммуникации) в системах поддержки принятия решений.

В настоящее время наблюдаются быстрые глобальные изменения в окружающей среде, во многом определяя характер деятельности человека. Поэтому большое внимание уделяется сбору, анализу и прогнозу данных об окружающей среде. Полвека насчитывает история глобальных цифровых сетей наблюдений изменений среды и космического зондирования: сейсмология, метеорология, магнитосфера и ионосфера, космическая погода, архивы цифровых изображений и телеметрия со спутников. Выводы Межправительственной группы экспертов ООН по изменению климата, МГЭИК (Intergovernmental Panel on Climate Change, IPCC)¹ во многом делаются на основании изучения цифровых моделей, использующие сверхбольшие базы данных по метеорологии, представляющие из себя временные ряды с географической привязкой по станциям или координатам. Оценки МГЭИК публикуются с 1991 г. каждые четыре года, пятый доклад намечен на 2013 г. На совещании² центров данных МГЭИК по подготовке пятого доклада в Сиэтле в сентябре 2008 г. предполагается объем научных баз данных для пятого доклада 3 петабайта с зеркальным хранением в трех центрах данных в США и Западной Европе и с оперативным доступом для 23 ведущих исследовательских центров по метеорологии, экологии, биохимии и численному моделированию. Аналогичные показатели по объемам и топологии имеют центры данных по дистанционному зондированию административных районов США по океанам и атмосфере (NOAA)³ и исследованию космического пространства (NASA)⁴. В сентябре 2008 г. первый

¹ <http://www.ipcc.ch/languages/russian.htm>

² <http://go-essp.gfdl.noaa.gov/2008/agenda.html>

³ <http://www.osd.noaa.gov/class/>

⁴ <http://nssdc.gsfc.nasa.gov/>

параллельный кластер для научных баз данных объемом более 1 Пбайт был запущен в университете Джона Хопкинса, США. На нем предполагается хранить многомерные массивы данных по астрофизике и по сенсорным сетям для био-экологического мониторинга.

Нарастает процесс глобальной информатизации научных исследований на основе Интернета, веб-сервисов и динамических веб-приложений (порталов). На национальном уровне научные сети сегодня имеют емкости до терабит в секунду (Национальный центр добычи данных США в Чикаго)⁵. Сетевые емкости международных научных сетей (GLORIAD⁶, Geant⁷) сегодня находятся на уровне 10-100 Гбит/с с возможностью динамической перестройки топологии и выделения емкости по запросу от приложений. Это создает предпосылки для создания сервисов распределенной параллельной обработки и анализа сверхбольших баз данных (cloud computing). При этом предполагается доставка вычислительных задач на процессоры рядом с данными, а не передача больших объемов данных для обработки на удаленный суперкомпьютер.

Для виртуализации доступа и оптимизации нагрузки на вычислительную и сетевые ресурсы разрабатывается промежуточное программное обеспечение (ППО) и развертывается инфраструктура под собирательным именем Грид. Первоначально Грид охватывал лишь распределенные вычисления, например, распределенную статистическую обработку экспериментальных данных на Большом адронном коллайдере в ЦЕРНе (инфраструктура LCG, позднее EGEE). К настоящему времени существует три основных пакета ППО и основанных на них инфраструктуры для грид-вычислений: gLite⁸ и созданный на его основе EGEE⁹, UNICORE¹⁰ и инфраструктура DEISA¹¹, Globus Toolkit 4¹² и TeraGrid¹³. Естественно, сложные вычислительные задачи моделирования глобальной циркуляции атмосферы, океана, магнитного поля и потоков частиц в околоземном пространстве, обработки данных дистанционного зондирования находят реализацию на всех доступных грид-инфраструктурах. При этом, в связи с высокой стоимостью вычислительных экспериментов (для климатических моделей нередко вычисления на самых мощных суперкомпьютерах в течение нескольких лет), необходимо не только гарантировать качество больших объемов

⁵ <http://www.ncdm.uic.edu/>

⁶ <http://www.gloriad.org/>

⁷ <http://www.geant2.net/>

⁸ <http://glite.web.cern.ch/glite/>

⁹ <http://www.eu-egee.org/>

¹⁰ <http://www.unicore.eu/>

¹¹ <http://www.deisa.eu/>

¹² <http://www.globus.org/>

¹³ <http://www.teragrid.org/>

входных данных, но и долговременно хранить результаты дорогостоящих расчетов с возможностью поиска и анализа данных.

Информационная поддержка сельскохозяйственной деятельности и землепользования, транспорт, туризм уже не могут обойтись без оперативных потоков данных о погоде, космоснимков и электронных карт. Виртуальные игровые миры и профессиональные тренажеры требуют детальных и реалистичных моделей окружающего мира. Это ускоряет доступ и интеграцию данных между различными областями науки и коммерческих приложений.

Чтобы поддерживать такие темпы информатизации, нужны новые технологии: распределенные Грид-вычисления, параллельные ГИС, веб-сервисы электронных карт на основе VirtualEarth и GoogleMaps, семантический веб, сверхбольшие базы данных и хранилища метаданных.

В итоге, сложилось два полюса в сфере ИТ: огромный объем исходных данных с одной стороны, и приложения, работающие с этой информацией с другой. При этом многие терабайты информации не позволяют использовать их напрямую, простым скачиванием и перебором. От приложений требуется более “интеллектуальный” подход к определению области своих интересов. В результате, приходим к необходимости анализа данных не только в научных расчетах, но и в “бытовых” приложениях. Кроме того, ИТ стремятся к эмулированию “естественного языка” и человеческого мышления, используя экспертные системы и искусственный интеллект. Не последнюю роль в этом играет тот факт, что объем данных растет в экспоненциальной прогрессии, в то время как число экспертов, их обрабатывающих, - в линейной. В связи с этим, условия выбора и анализ данных становятся все более сложными, что затрудняет их математическую формализацию. Поэтому все больше возникает потребность в коммуникаторе, который будет предоставлять достижения информационной технологии в пользование рядовых потребителей, нужен универсальный механизм поиска, интеллектуального анализа и добычи научных данных (data mining).

К настоящему времени довольно широкое развитие получил текстовый поиск и текстовое индексирование. Также существуют системы для поиска метаданных, такие например, как Global Change Master Directory (GCMD), разработанная в НАСА, или Master Environmental Library (MEL), созданная Отделом моделирования Министерства обороны США. Однако, поиск внутри самих цифровых архивах данных пока недостаточно развит. Весьма показательна ситуация с развитием поисковых систем корпораций Google и Microsoft:

- 1) Первый уровень: поиск по вебу (Google, Live Search, Yandex и др.)

- 2) Второй уровень: специализированный поиск (например, поиск внутри документов и изображений)
- 3) Третий уровень: поиск и визуализация данных с географической привязкой (адреса и карты Google Earth, Microsoft VirtualEarth)

На четвертом уровне должен стоять поиск и анализ внутри многомерных массивов научных данных, в первую очередь временных рядов, например по истории климата или по биржевым сделкам.

Анализ временных рядов можно разделить на два основных вида: анализ трендов и анализ экстремальных выбросов. Первый можно представить как анализ движения во времени центра масс пространства данных, в то время как второй - анализ тенденции движения границ. Анализ трендов имеет значение при изучении долгосрочных изменений климата. Анализ выбросов также важен, поскольку анализирует граничные (предельные) состояния среды, зачастую являющиеся критическими для любой деятельности и инфраструктуры, которые обычно рассчитываются исходя из среднестатистических показателей.

Современная наука во многом междисциплинарна, например: анализ влияния климата на биосферу, анализ причинных связей в метеорологии и космической погоде. Это накладывает дополнительные требования на систему добычи и анализа данных: во-первых, процесс анализа должен производиться в различных средах совместно, во-вторых, требуется унифицированный интерфейс доступа к данным в разных источниках. Назрела потребность в «глобализации» научных данных.

Актуальность данной работы обусловлена тем, что она посвящена поиску, междисциплинарному анализу и выборке данных в распределенных сверхбольших архивах данных по окружающей среде. При этом объектом исследования является модель научных данных в виде набора многомерных массивов с семантической нагрузкой, язык запросов и обработки данных для класса источников, отвечающих этой модели, определение и алгоритмы поиска областей значений внутри модели данных (сценариев событий), оптимизация параллельных алгоритмов для использования в распределенной Грид-среде.

Цели и задачи работы

Целью работы являлось разработка системы поиска, анализа и распределенной выборки данных в сверхбольших архивах данных по окружающей среде.

Для достижения указанной цели необходимо решить следующие задачи:

- 1) Разработать систему алгоритмов для распределенного поиска в сверхбольших архивах научных данных. Предоставить пользователю возможность формулировать запросы к архивам с помощью естественных языковых конструкций.
- 2) Разработать программное обеспечение (ПО), которое бы реализовывало алгоритмы из п. 1, а также поддерживала распределенные выборки и добычи данных в сверхбольших архивах данных по окружающей среде.
- 3) Встроить данное ПО в инфраструктуру Грид.
- 4) Провести апробацию и оценить эффективность предложенных алгоритмов на различных прикладных задачах.

Научная новизна

Для прогноза глобальных изменений климата строится множество теорий. Каждая из них выводит определенные закономерности и сценарии развития событий. Результаты данной работы могут применяться для проверки и отслеживания предполагаемых климатических сценариев. В работе формулируется определение сценария событий окружающей среды в терминах нечеткой логики, которое включает не только значение параметров, но и каузальные связи - изменение значений параметров во времени. В работе также формулируется алгоритм поиска таких сценариев на классе источников, подчиняющихся единой модели данных.

В работе предложен новый подход к системам поиска и анализа данных по окружающей среде. В отличие от аналогичных систем, в которых поиск осуществляется по метаданным и текстовой информации, в данной работе представлен подход к поиску по значениям временных рядов параметров среды, который оптимизирован для работы со сверхбольшими и распределенными массивами данных в интерактивном режиме. Разработана общая модель представления данных для подобной системы, а также методы оценки качества и сравнения результатов поиска.

Инфраструктура системы совместима с Грид-технологией, что дает возможности пользоваться всеми возможностями Грид: системой авторизации, Грид-инфраструктурой (ресурсы, мощности), промежуточным программным обеспечением. В работе также реализованы потоковые алгоритмы обработки данных, оптимизированные под работу в сети с большими объемами данных - пользователю не требуется закачивать данные и промежуточные результаты расчетов на свою машину. Это дает возможность использовать результаты работы для расчетов по числовым моделям в Грид-среде.

Практическая значимость работы

В настоящей работе предложен абстрактный уровень веб-сервисов данных для виртуализации баз данных, с помощью которых выбираются временные ряды с параметрами окружающей среды для поиска событий. Интерфейс источника данных реализован в виде плагинов в контейнере Грид-сервиса данных OGSA-DAI с простыми XML-схемами ввода и вывода. Следование спецификации OGSA-DAI и использование языка программирования Java/J# позволяет нам погрузить виртуальные веб-сервисы данных в большинство из существующих сегодня контейнеров веб- и Грид-сервисов, включая Microsoft ASP.NET, Apache Tomcat/Axis, WSRF Globus Toolkit 4, OMII, и EGEE gLite.

Представим, например, что пользователю нужны не только все данные о погоде в Московской области за последние 50 лет, но и примеры атмосферных фронтов вблизи Москвы. Более того, представим, что пользователю нужны изображения со спутников в момент прохождения атмосферного фронта и оценка, как часто наблюдались эти события вблизи Москвы за последние 10 лет. Ответы на подобные вопросы можно получить с помощью разработанного в рамках настоящей работы ПО. Данная работа может быть использована также в исследовании глобального изменения климата, в агроклиматологии.

Настоящая работа выполнялась в Геофизическом центре (ГЦ) и в Институте космических исследований (ИКИ) РАН, а также на факультете вычислительной математики и кибернетики МГУ при поддержке грантов «Система поиска погодных сценариев – СППС» (Environmental Scenario Search Engine – ESSE) от Исследовательской лаборатории Майкрософт (MSR) в Кембридже, Великобритания, и «Интерактивный ресурс данных по солнечно-земной физике» (Space Physics Interactive Data Resource – SPIDR) от Национального управления по исследованию атмосферы и океана (NOAA) США.

Результаты работы нашли прямое применение в совместных междисциплинарных исследованиях ИКИ РАН и MSR для анализа региональных трендов и зависимостей между изменениями в климате и в растительности (CLIVT - Climate Induced Vegetation Change Analysis Tool)¹⁴. Созданные в работе сервисы поиска и обработки сверхбольших баз данных по истории климата используются совместно с разработанными в ИКИ методами дистанционного зондирования растительности с использованием данных спутниковых наблюдений. Компания Майкрософт осуществляет техническую экспертизу, а также предоставляет необходимые вычислительные ресурсы, включая кластер для параллельной обработки и хранения данных.

Созданные в результате работы Грид-сервисы OGSA-DAI для поиска и обработки данных установлены на ресурсном центре СКИФ-Грид в ГЦ РАН в Москве, а также в

¹⁴ <http://www.tcirussia.com/clivt/default.aspx>

Национальном геофизическом центре NOAA в Болдере, штат Колорадо, и регулярно используются для доступа к архивам данных по климату и космической погоде в системах доступа к распределенным архивам данных NOAA Comprehensive Large Array Stewardship System (CLASS)¹⁵ и Space Physics Interactive Data Resource (SPIDR)¹⁶.

Апробация работы и публикации

Результаты диссертации опубликованы в работах [1]-[5] и обсуждались на следующих конференциях и семинарах:

- 1) Zhizhin M, E Kihn, R Redmon, A Poyda, D Mishin, D Medvedev, V Lyutsarev, *Integrating and mining distributed environmental archives on Grids*, VLDB DMG Workshop, Seul, September 2006
- 2) Алексей Пойда, Михаил Жижин, Дмитрий Мишин, Дмитрий Медведев, Сергей Березин, Дмитрий Войцеховский, Василий Люцарев, Эрик Кин. *Система поиска погодных сценариев*. Пятая Юбилейная Открытая Всероссийская конференция “Современные проблемы дистанционного зондирования Земли из космоса” Москва, ИКИ РАН, 12-16 ноября 2007г.
- 3) Zhizhin M, D Medvedev, A Poyda, V Lyutsarev, *Grid data mining with CDM active storage*, 3rd GRID e-collaboration Workshop for Earth Science and Space 16 - 17 January 2008, ESRIN, Frascati (Rome), Italy
- 4) Пойда А.А. *Поиск данных в Грид: соотношение производительности сетей, вычислительных кластеров, хранилищ данных*. “Современные информационные технологии для научных исследований” Магадан, 20-24 апреля 2008

Объем и структура работы

Диссертация состоит из введения, четырех глав, заключения, списка литературы (46 наименования) и приложения. Общий объем работы составляет 113 страниц, работа содержит 31 иллюстрацию и 5 таблиц.

Краткое содержание работы

Во **введении** излагаются: объект исследования и актуальность темы, цели и задачи работы, научная новизна, практическая значимость работы; перечисляются публикации и выступления на семинарах.

¹⁵ <http://www.osd.noaa.gov/class/>

¹⁶ <http://spidr.ngdc.noaa.gov/spidr/index.jsp>

В **первой главе** приводится обзор технологий поиска научных данных и технологий, необходимых для понимания дальнейших глав диссертации. Здесь дается обзор алгоритмов, используемых в поисковых системах в среде Интернет, а также обзор проектов, связанных с хранением, выборкой, поиском, обработкой и визуализацией больших объемов научных данных. Здесь также описываются структура поисковой системы данных и проблемы, которые необходимо решать разработчикам при ее создании. Конец главы посвящен распределенной выборке и обработке данных, в Грид-среде.

Вторая глава посвящена алгоритмам распределенного поиска сценариев событий в окружающей среде и состоит из пяти разделов.

В **первом разделе** рассматривается проблема абстрактной модели данных для представления научных данных. Под абстрактной моделью мы понимаем контракт на структуру связей между элементами контейнера данных, который используется для обмена между компонентами распределенной системы. При работе с массивами данных в этом контейнере должны содержаться многомерные массивы различных типов данных, а также массивы значений их координатных осей, метаданные о единицах измерений и, возможно, о происхождении и истории обработки данных.

В разделе вводится абстрактная модель данных ESSE CDM (рис. 1) разработанная в рамках данной работы, которая сравнивается с уже существующими моделями.

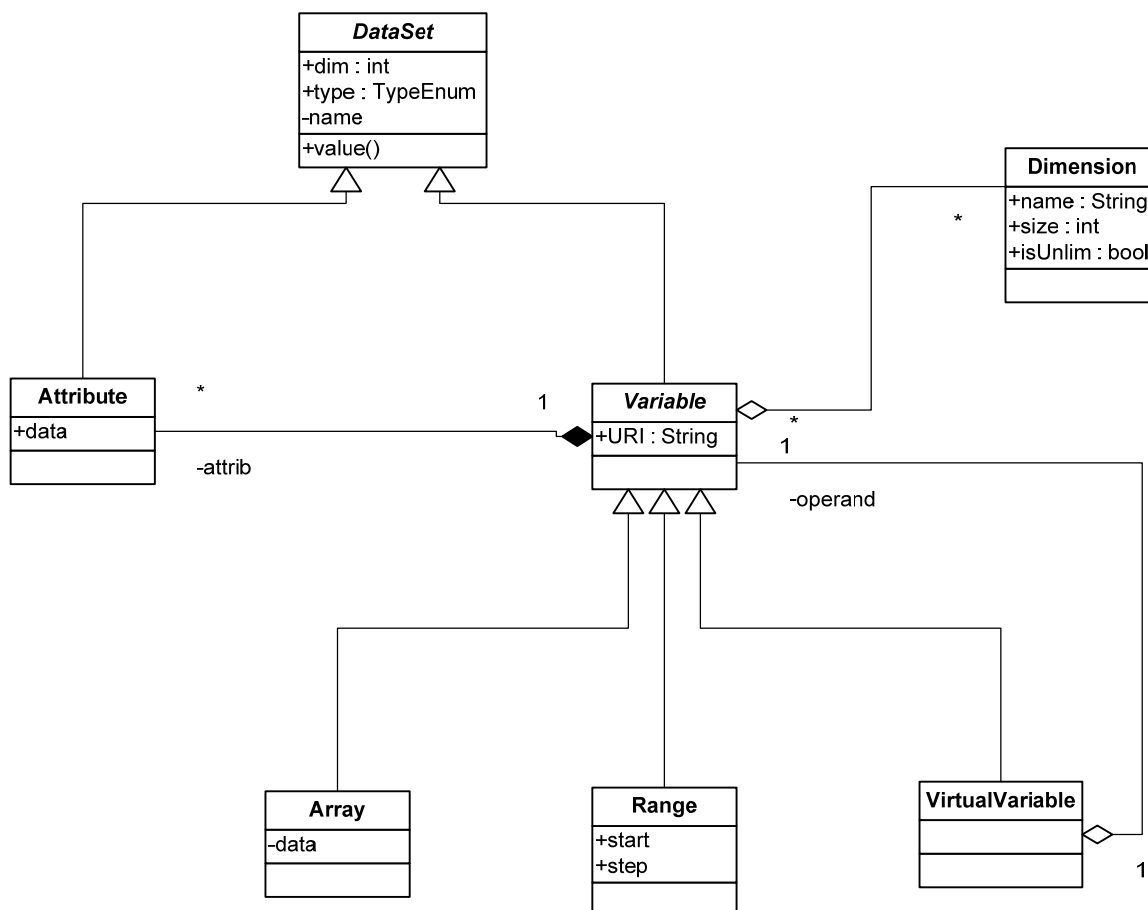


Рисунок 1 - UML диаграмма классов абстрактной модели данных ESSE CDM

Также приводятся наиболее часто используемые в настоящее время схемы организации научных данных (траектории, сеточное покрытие, несвязный набор точек, станции), и анализируется возможность их реализации в рамках ESSE CDM модели.

Во **втором разделе** определяются границы того, что мы ищем, а именно – вводится понятие события. Каждое событие представляет собой совокупность состояний параметров окружающей среды, например “(‘температура около нуля’) и (‘низкое атмосферное давление’)”. Под сценарием события мы понимаем упорядоченную последовательность состояний, отстоящих друг от друга во времени. В этом же разделе вводятся базовые алгоритмы поиска сценария событий с применением аппарата нечеткой логики. Каждое состояние выражается набором функций принадлежности нечеткой логики, связанных между собой операторами нечеткого логического “И”/”ИЛИ”. В свою очередь, состояния связываются между собой в единый сценарий с помощью нечеткого оператора временного сдвига, который не входит в стандартное множество операторов нечеткой логики и впервые вводится в данной работе. Это позволяет задавать сценарий события в виде последовательности состояний, связанных во времени с помощью нечетких категорий, таких как “одновременно”, “зadолго до”, “вскоре”.

Центральными функциями в этом алгоритме являются операторы объединения и пересечения нечетких множеств. Можно использовать разные библиотеки операторов, например минимаксные, но мы предлагаем использовать гладкие функции без сингулярностей, что повышает вероятность получения нетривиальных результатов поиска в многопараметрическом пространстве. Например, можно использовать функции Ягера, которые определяются формулами

$$T_Y(\mu_A(x), \mu_B(x), q) = 1 - \min \left\{ 1, \left[(1 - \mu_A(x))^q + (1 - \mu_B(x))^q \right]^{\frac{1}{q}} \right\}$$

$$S_Y(\mu_A(x), \mu_B(x), q) = \min \left\{ 1, \left[(\mu_A(x))^q + (\mu_B(x))^q \right]^{\frac{1}{q}} \right\}$$

В эти формулы входят степенные функции. Это приводит к существенным временным затратам, учитывая, что речь идет о работе с большим объемом данных. Проблемам оптимизации вычисления сложных функций посвящен **третий раздел**, в котором эти функции аппроксимируются дискретными аналогами и табличным вычислением. Так, например, если значение нечеткой функции принадлежности разделить на N равных частей и ввести дискретизацию, то нам понадобится всего две таблицы: $[N]^q$ и $[N]^{\frac{1}{q}}$. Чем больше значение N – тем точнее результат, однако тем медленнее работает система. Это позволяет существенно повысить эффективность алгоритмов нечеткого поиска, на порядок сокращая время обработки запроса.

Четвертый раздел описывает усовершенствование поисковых алгоритмов, приведенных в предыдущих разделах, путем добавления в сценарии нечетких временных промежутков между происходящими событиями. Например, мы можем задавать сценарий типа ”событие А произошло незадолго до события В”, где временной промежуток между событиями А и В неоднозначен. Чтобы обеспечить подобную функциональность, здесь также используются функции принадлежности нечеткой логики.

Пятый раздел описывает распределенный вариант алгоритма поиска. Недостаточно определить и уметь вычислять сценарий событий. Надо учитывать, что мы работаем с распределенными источниками в интерактивном режиме. Перемещение всех данных в одно место с целью последующей обработки требует больших временных затрат. Поэтому обеспечена возможность вычисления подвыражений на разных машинах, до тех пор, пока не требуется их слияния для дальнейшего расчета. Когда же требуется совместить частичные результаты поиска, выбирается одна из нескольких задействованных машин, и на нее пересылаются промежуточные результаты других подвыражений. Промежуточными результатами для выражений нечеткой логики являются списки текущих коэффициентов

релевантности. Чтобы как можно больше снизить трафик, обеспечена возможность передавать только ненулевые значения (при обработке больших объемов данных, большая часть значений релевантности близка к нулю). Можно добиться такого планирования ресурсов, чтобы сократить число пересылок до минимума.

Третья глава посвящена программной реализации сервисов распределенной обработки и поиска данных. Глава состоит из пяти разделов. **В первом разделе** выстраивается архитектура ПО (рис.2). Выделяется три центральных сервиса: сервис выборки данных, сервис обработки (преобразования) данных и поисковый сервис. При этом система позволяет последовательно объединять эти сервисы в единый рабочий поток. Например, выбрать данные из разных источников, преобразовать эту выборку (вычислить виртуальный параметр), провести нечеткий поиск, обработать результаты поиска (сгладить) и передать их пользователю.

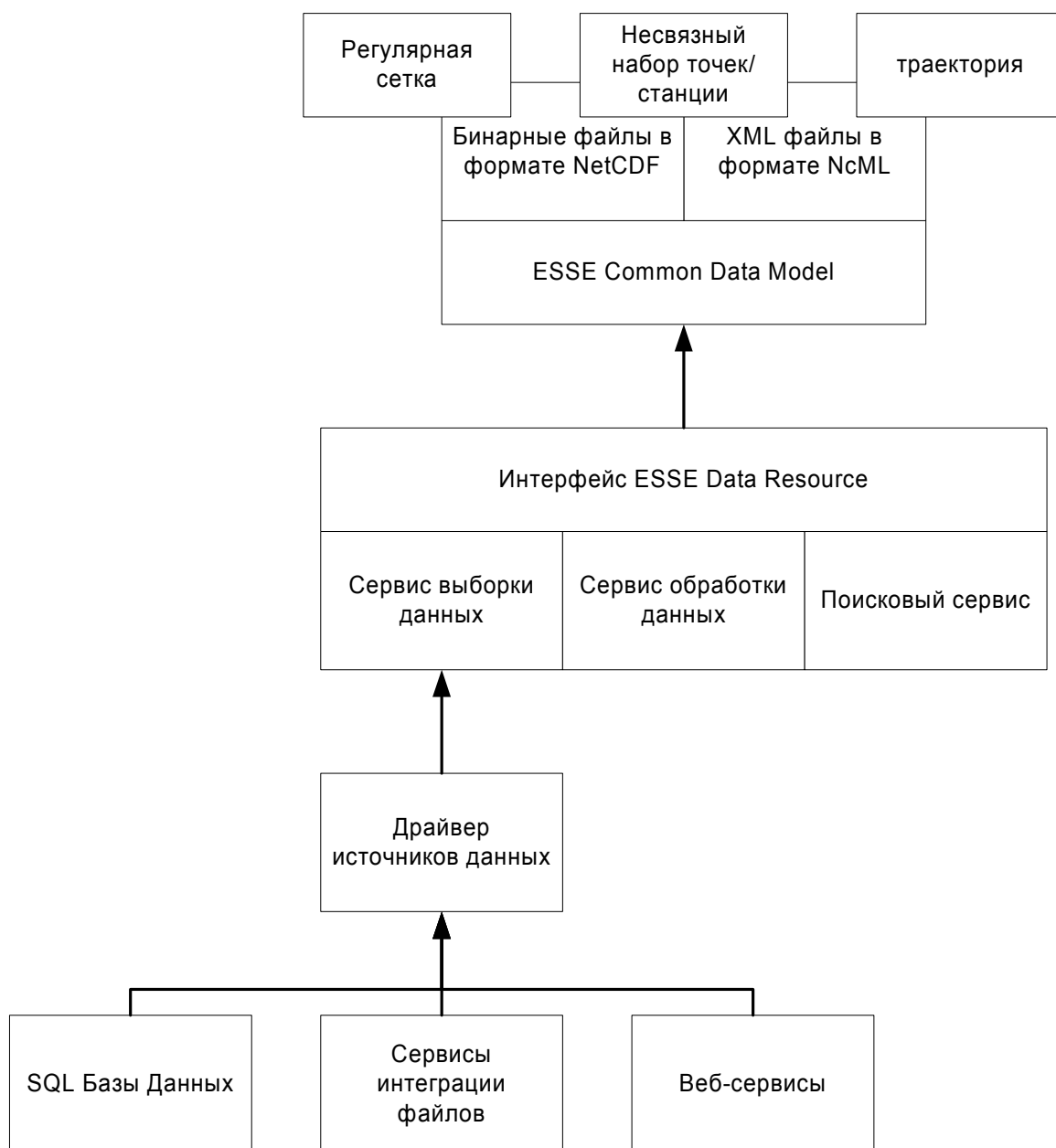


Рисунок 2. Архитектура системы

Сервис выборки данных, детально рассматриваемый во **втором разделе**, реализует интерфейс доступа к источникам данных (например, это может быть база данных или удаленный веб-сервис или просто файл и т.п.). Его задача – предоставлять унифицированный интерфейс для всех источников. Доступ к данным осуществляется на основе метаданных, которые делятся на два уровня. В метаданных нижнего уровня (инвентаризация) записана информация о наличии/отсутствии параметров за определенный промежуток времени, дана информация о месте расположения и типе хранилища параметра. Метаданные верхнего уровня (каталог) описывают свойства параметра, не привязанные к низкоуровневым хранилищам. К таковым относятся: факт наличия параметра за все время, единицы измерения, название параметра, его описание и т.п.

Поисковый сервис, детально рассматриваемый в **третьем разделе**, реализует алгоритмы поиска сценариев событий в терминах нечеткой логики, описанные в предыдущих разделах. В разделе приводится описание языка запроса к сервису и логика его работы.

Сервис обработки данных рассматривается в **четвертом разделе**. Сервис необходим для получения производных параметров, например, среднесуточных значений температуры, суммарных месячных осадков, эффективных годовых температур и т. п. В разделе описывается язык запроса к сервису и механизм его работы.

В **пятом разделе** показано, как встроен поисковый сервис в инфраструктуру Грид, используя существующий контейнер для Грид-сервисов данных OGSA-DAI. Описываются интерфейсы доступа и механизмы взаимодействия компонентов системы. В OGSA-DAI все операции с любым источником данных осуществляется через менеджера, реализующего стандартный интерфейс (Resource). Этот менеджер является виртуальным хранилищем, агрегирующим несколько реальных источников данных (базы данных, файлы и т.д.) и выполняющим атомарные “приказы” (Activity), являющиеся, по сути, реализациями некоторых функций обработки данных. Результаты работы одной функции можно перенаправить на вход другой, организуя, таким образом, потоковую обработку данных. Диалог между Activity и пользователем, как и между самими Activity осуществляется посредством XML-документов с применением механизма веб-сервисов. Таким образом, для внедрения наших алгоритмов в OGSA-DAI нам пришлось создать класс, реализующий интерфейс Resource и написать ряд Activity, представленных в Таблице 1, для доступа, обработки и поиска данных.

Таблица 1 - Расширение платформы OGSA-DAI

Метод	Описание
GetMetadataActivity	Возвращает по запросу метаданные о требуемом источнике
GetCdmDataActivity	Производит выборку данных с требуемыми параметрами
FuzzySearchActivity	Принимает на вход один или более временных рядов и возвращает набор коэффициентов релевантности

DataProcessActivity	Принимает на вход один или более временных рядов и возвращает новый временной ряд
---------------------	---

В **четвертой главе** собраны примеры использования системы. Глава состоит из трех разделов.

В **первом разделе** описываются источники данных (см. Таблицу 2), уже адаптированных в систему.

Таблица 2 - Источники данных, используемые в проекте

Источник данных	Тип	Примеры параметров	Временные рамки	Зона наблюдения	Размер	URL
NCEP/NCAR	Метео	Температура, скорость ветра, облачность	1949 - 2007	Глобальная, с разрешением 2.5 градуса.	1 Tb	http://dss.ucar.edu/pub/reanalysis/
SWR	Космос	Электрический потенциал, проводимость, магнитное поле	1991 - 2002	Глобальная, с переменным разрешением	3 Tb	http://swr.ngdc.noaa.gov
SPIDR	Космос	Геомагнитные индексы, магнитные вариации, ионосфера	1933 - 2008	Глобальная, данные с обсерваторий	500 Gb	http://spidr.ngdc.noaa.gov

Во **втором разделе** приводится пример расчета статистики экстремальных событий и анализа климатических трендов. Например, используя поисковый сервис, мы ищем десять самых жарких дней в каждой географической точке и анализируем их распределение по трем последним десятилетиям. Это позволяет оценить тенденцию роста числа экстремальных событий. Также, с помощью сервиса обработки данных, мы производим расчет среднегодовых значений температуры в каждой географической точке и вычисление линейной регрессии (трендов). Подобные оценки статистики экстремальных событий и трендов по различным климатическим параметрам использовался в проекте CLIVT (Climate

Induced Vegetation Change Analysis Tool)¹⁴, в котором участвуют Microsoft Research и Институт космических исследований Российской академии наук. Полученные результаты сравниваются с аналогичными результатами, рассчитанными по другим данным Межправительственной группой экспертов ООН по изменению климата (IPCC)¹.

В **третьем разделе** описано применение алгоритмов поиска для оценки воздействий окружающей среды на техногенные системы. Суть его в следующем: есть набор систем, работоспособность которых зависит от условий окружающей среды. Кроме того, есть набор правил, определяющих неблагоприятные условия окружающей среды для каждой системы. Правила транслируются в сценарии событий на некоторых параметрах окружающей среды. Каждая система связана с одним или более правилами. Наличие связи показывает, что на работоспособность данной системы влияют условия окружающей среды, определяемые сценарием события, описываемым данным правилом. Связь может быть двух видов: “тревога” и “предупреждение”. Первый вариант говорит, что факт срабатывания этого правила влечет неработоспособность системы. Второй – что система находится в опасности. Программа имеет следующее назначение: с одной стороны она должна оценивать работоспособность систем в зависимости от климатологии, с другой она должна выдавать варианты причин выхода систем из строя. Например, при планировании спасательных операций средство доставки выбирают в зависимости от погодных условий (вертолет, машины и т.п.). Погодные условия могут влиять на выбор маршрутов при оптимизации транспортных потоков.

В **заключении** подводятся итоги, формулируются результаты, делаются выводы.

Основные результаты работы

1. Разработан новый подход к организации поиска научных данных по их значениям, а не по метаданным и описательной текстовой информации. Предложены методы оценки качества и сравнения результатов такого поиска.
2. Для систем поиска, реализующих предложенный подход, разработана общая модель представления и интеграции данных, оптимизированная для работы с большими распределенными массивами данных по окружающей среде в интерактивном режиме.
3. На основе оригинального определения сценария события в окружающей среде в терминах нечеткой логики, учитывающего свойства потока в фазовом пространстве динамической системы параметров, созданы инструментальные программные средства для разработки виртуальных тренажеров и интеллектуальных систем для оценки степени воздействия окружающей среды в системах поддержки принятия решений.

4. Разработанный инструментарий совместим с Грид-инфраструктурой. Все компоненты системы реализованы в виде Грид-сервисов, что дает возможность использовать результаты работы для расчетов по числовым моделям в Грид-среде. Разработанные Грид-сервисы вошли в качестве составной компоненты в пакет промежуточного программного обеспечения Российско-Беларусской суперкомпьютерной инфраструктуры “СКИФ-Грид”.

Работы автора по теме диссертации

- [1] M. Zhizhin, A. Poyda, D. Mishin, D. Medvedev, E. Kihn, V. Lyutsarev. Scenario Search on the Grid of Environmental Data Sources. MSR Technical Report, July 2006, 17 p. [электронный ресурс], url: <http://research.microsoft.com/apps/pubs/default.aspx?id=68030>
- [2] Zhizhin, M., E. Kihn, R. Redmon, A. Poyda, D. Mishin, D. Medvedev and V. Lyutsarev. *Integrating and mining distributed environmental archives on grids*, Concurrency and Computation: Practice and Experience, vol. 19, 2007, pp. 2157 – 2170
- [3] Zhizhin, M, E Kihn, V Lyutsarev, S Berezin, A Poyda, D Mishin, D Medvedev and D Voitsekhovsky, *Environmental scenario search and visualization*, Proc. 15th ACM symposium on Advances in geographic information systems, 2007, ISBN:978-1-59593-914-2, 10 p.
- [4] Zhizhin M, A Poyda, D Mishin, D Medvedev, E Kihn, V Lyutsarev, *Grid Data Mining with Environmental Scenario Search Engine (ESSE)*, Chapter 13 in Data Mining Techniques in Grid Computing Environments, Ed. Werner Dubitzky, Wiley, 2008, pp. 281-306,
- [5] М.Н. Жижин, А.А. Пойда, Д.Ю. Мишин, А.П. Платонов, А.А. Солдатов, В.Е. Велихов, М.Н. Боярский, Р.Р. Назиров. Поиск данных в Грид: соотношение производительности сетей, вычислительных кластеров, хранилищ данных. Открытое образование, № 4, 2008, стр. 29-39