вычисления и задачи управления (PACO'2010): Матер. V Междунар. конф. – М., 26–28 октября 2010. – М., 2010. – С. 83–95.
10. Metcalfe R.M., Boggs D.R. Ethernet: Distributed Packet Switching for Local Computer Networks. 1975. URL: http://ethernethistory.typepad.com/papers/EthernetPaper.pdf (дата обращения: 11.01.2011).
11. Адресуемая ячейка однородной структуры для решения дифференциальных уравнений в частных производных: пат. 2427033 Рос. Федерация. № 2010107933/08; заявл. 03.03.10; опубл. 20.08.11, Бюл. № 23. – 7 с.

Поступила 02.05.2012 г.

УДК 004.42

АРХИТЕКТУРА РАСПРЕДЕЛЕННОГО ВЫЧИСЛИТЕЛЬНОГО КОМПЛЕКСА ДЛЯ ДВУМЕРНОГО АНАЛИЗА ИЗОБРАЖЕНИЙ ДИСКОВ ДЕРЕВЬЕВ

И.А. Ботыгин, В.Н. Попов, В.А. Тартаковский*

Томский политехнический университет
*Институт мониторинга климатических и экологических систем СО РАН, г. Томск
E-mail: botygin@ad.cctpu.edu.ru

Разработана архитектура и алгоритмы реализации распределенного вычислительный комплекса обработки дендроэкологических данных. Представлена GPSS-модель комплекса для оценки эффективности его функционирования. Практическая реализация комплекса осуществлялась с использованием стека свободно распространяемых программных продуктов. Проиллюстрирована работа комплекса на таких задачах двумерного анализа изображений дисков деревьев, как вычисление азимута и среднеквадратичной ширины области максимального прироста, а также вычисление значений индексов прироста ширины годичных колец деревьев.

Ключевые слова:
Дендроэкология, дендрохронология, древесные спилы, годовые кольца деревьев, математическая модель, серверное приложение.

Key words:
Dendroecology, dendrochronology, tree stem disk, tree-rings, mathematical model, server application.

Программное обеспечение в области дендрохронологии

Проведенный обзор и сравнительный анализ существующих аппаратно-программных комплексов и систем для анализа и обработки данных в области дендрохронологии показал, что они являются ограниченными для использования и ориентированы на решение новых задач, связанных с неоднородностью окружающего пространства. Особенностью таких задач является большой объем данных, которые необходимо хранить, обрабатывать и сопоставлять между собой, пространственная распределенность мест сбора образцов.

В таких условиях оптимальным системным решением, обеспечивающим повышение вычислительной мощности, увеличение объема хранимых данных, является использование технологии распределенной их обработки. Эта технология подразумевает физическое распределение хранения и обработки данных в пространстве на нескольких вычислительных машинах, которые связаны между собой каналами передачи данных, координацию их вычислительных мощностей, использование стандартных протоколов и служб сетевого взаимодействия. В данное время существует множество инструментальных средств технологий распределенных вычислений, а также проектов, реализованных с их использованием, но для мониторинга климатических и экологических изменений на основе биоиндикации такие технологии не применялись. Таким образом, реализация технологии распределенных вычислений и разработка алгоритмического обеспечения для дендрохронологических исследований, связанных с двумерным анализом изображений спилов деревьев, даст возможность получать новые результаты и качественные оценки параметров окружающей среды.

В настоящей работе описывается реализация технологии распределенных вычислений и разработка алгоритмического обеспечения для дендрохронологических исследований. Безусловно, дендрохронологические исследования – это только часть мониторинга, моделирования и прогнозирования климатических и экосистемных изменений под воздействием природных и антропогенных факторов. Но особенность задач дендроэкологического анализа, заключающаяся в необходимости математической обработки очень большого объема данных (временные ряды наблюдений могут достигать сотен гигабайт), широком спектре решаемых задач, коллективной работе многих сотрудников на всех этапах дендроэкологических исследований, а также в необходимости хранения и систематизации больших объемов неоднородной структурированной информации (собственно хронологические ряды наблюдений, результаты обработки,

сопутствующие метеорологические, геологические, геофизические, аэрокосмические и т. п. ряды наблюдений), однозначно подразумевает в качестве одного из системных решений использование технологии распределенной обработки (*grid*-технологии), обеспечивающей динамическое изменение основных компонентов инфраструктуры системы обработки (от структур хранимых данных – до схем и алгоритмов решаемых задач) и повышение вычислительной мощности.

Сравнительный обзор и анализ аппаратно-программных средств обработки и анализа годичных колец деревьев

Анализ научных исследований в области дендроэкологии, дендрохронологии и дендроклиматологии показывает, что в настоящее время для моделирования и анализа дендроэкологических данных используется достаточно широкий спектр аппаратно-программных средств. Использование программных средств зависит от задач исследования. Это может быть статистическая обработка первичных данных замеров характеристик прироста и получение надёжных обобщённых хронологий, сопоставление характеристик прироста с факторами внешней среды, моделирование процессов роста, изучение пространственного распределения характеристик прироста и визуализация результатов анализа. Для решения этих задач одни исследователи предпочитают использовать широко распространённые статистические пакеты (SAS, MatLab, SyStat, STATISTICA и др.), электронные таблицы (QUATTRO, LOTUS, Excel и др.), универсальные системы математической обработки результатов, предназначенные для численного и символьного решения математических задач различной сложности (MathCAD, Mathematica и др.) и интегрированные программные решения, такие как DPL, PRECON 5.17C, TREERING 3.0, LignoVision 1.32, TSAP-Win Professional 0.30, DendroClim 2002, WinDENDRO, DendroLab 470, PAST 32, OSM 3.10, MeasureJ2X, или решения на базе географических информационных систем (ESRI ARC/INFO, ArcView, Mapinfo, AutoCAD Map и др.) с форматами различных баз данных (Microsoft Access, Oracle, dBASE, FoxPro и др.) для связи обрабатываемых данных с географическим местоположением их сбора и отображением на электронных картах. В настоящее время широко распространено, в том числе и в России, оборудование и программное обеспечение для диагностики, контроля и исследования внутреннего состояния деревьев и древесины, разработанное компанией RINNTECH [1].

Рассмотренные программные комплексы и системы достаточно универсальны, ориентированы на широкий класс исследований и выполнение часто используемых операций обработки и анализа дендроэкологических данных (например, измерение ширины годичных колец и др.). Современные инструментальные средства автоматизируют многие этапы получения, сбора и обработки дендроэкологических данных, но и одновременно создают большие потоки информации, требующие не только обширных баз для их хранения, но и серьезных математических методов их обработки.

GPSS-модель для оценки эффективности функционирования распределенного вычислительного комплекса обработки дендроэкологических данных

Исследование задач обработки дендроэкологических данных и инструментально-программных средств их решения выявили трудоёмкость и сложность создания соответствующего программного обеспечения. Для успешного проектирования и реализации программного обеспечения обработки дендроэкологических данных должны быть построены его полные и непротиворечивые как функциональные, так и информационные модели, структура основных компонентов программного комплекса и алгоритмы их функционирования.

Для оценки эффективности функционирования распределенного вычислительного комплекса произвольной структуры была разработана его имитационная модель на языке GPSS World [2].

На рис. 1 отражена укрупненная модель распределенного вычислительного комплекса обработки дендроэкологических данных.

Основными объектами моделируемой системы являются: потоки входных заданий, вычислительные сегменты, вычислительные серверы (ВС) сегментов, коммуникационный сервер (КС), серверы баз данных (БД), менеджер БД. Вычисление моментов появления заданий (случайной величины, связанной с промежутком времени между появлениями двух соседних заданий) осуществляется по нормальному закону распределения. Интервалы обслуживания также являются случайной величиной.

Изменяемыми параметрами модели являются (всего 18): количество ВС, максимальное время обработки задания, количество типов заданий, средний интервал между заданиями, доля параллельных заданий, доля отклонённых заданий, среднее время обработки заданий КС, отклонение от среднего времени обработки заданий КС, вероятность сбоя ВС, вероятность восстановления ВС, среднее время восстановления ВС, отклонение от среднего времени восстановления ВС, количество КС, максимальное время обработки задания администратора, количество типов заданий, средний интервал между заданиями, среднее время восстановления ВС, отклонение от среднего времени восстановления ВС. Типовыми заданиями в настоящей модели являлись задачи вычисления азимута и среднеквадратичной ширины области максимального прироста годичных колец деревьев, вычисления значений индексов прироста ширины годичных колец деревьев и т. д.

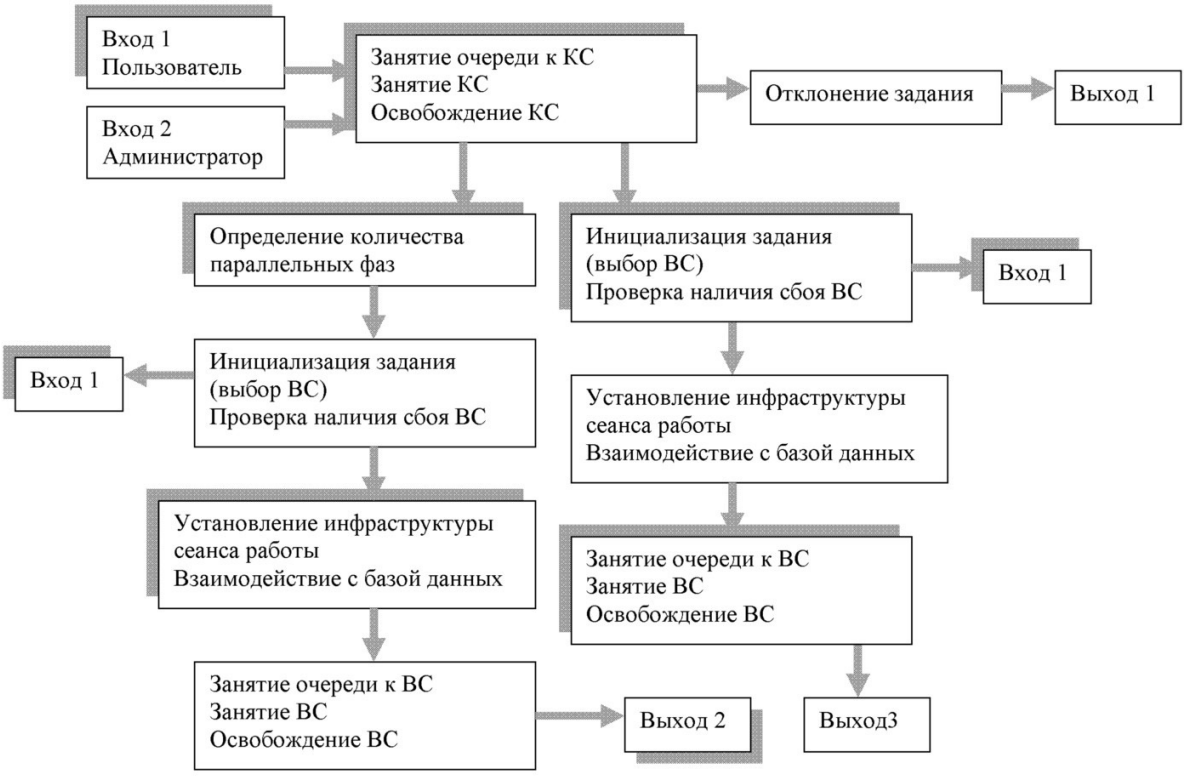


Рис. 1. *GPSS-модель для оценки эффективности функционирования распределенного вычислительного комплекса обработки дендроэкологических данных*

В работе проведено исследование зависимостей таких критериев как (всего 9): количество заданий общее, отклонённых, без параллелелизма, с параллелизмом, аварийных ситуаций с восстановлением, без восстановления, с восстановлением (для заданий с параллелизмом), без восстановления (для заданий с параллелизмом), заданий в очереди, от следующих параметров (всего 5): количество ВС, максимальное время обработки задания, количество типов заданий, интервал между заданиями, количество КС (рис. 2).

По оси Y приведены полученные значения исследуемых критериев – абсолютные численные значения количества, а по оси X – абсолютные численные значения соответствующих изменяемых параметров.

Во всех экспериментах моделирование осуществлялось в течении 24 ч (модельное время).

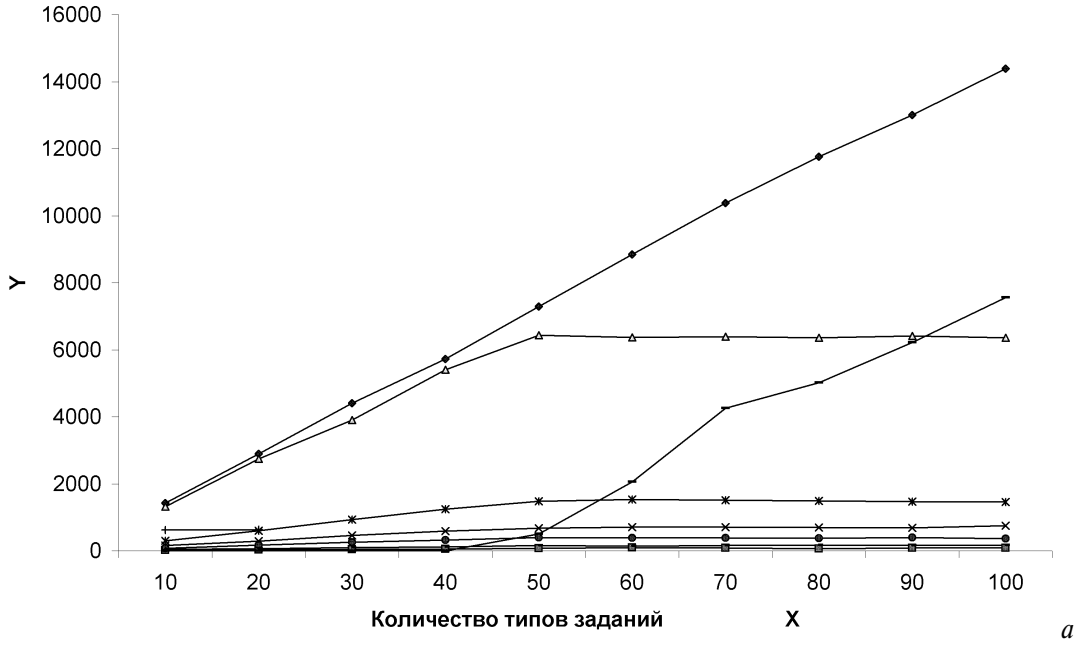
При выбранных в модели параметрах быстродействия ВС: их количество (свыше 100), максимального времени обработки задания (до 100 единиц), количество КС (до 10) практически не влияет на исследуемые критерии. Количество типов выполняемых заданий значительно влияет на число выполненных заданий, а также на очередь к КС. Изменение интервала поступления заданий значительно влияет на исследуемые критерии, особенно на начальных этапах его увеличения.

Распределенный вычислительный комплекс обработки дендроэкологических данных

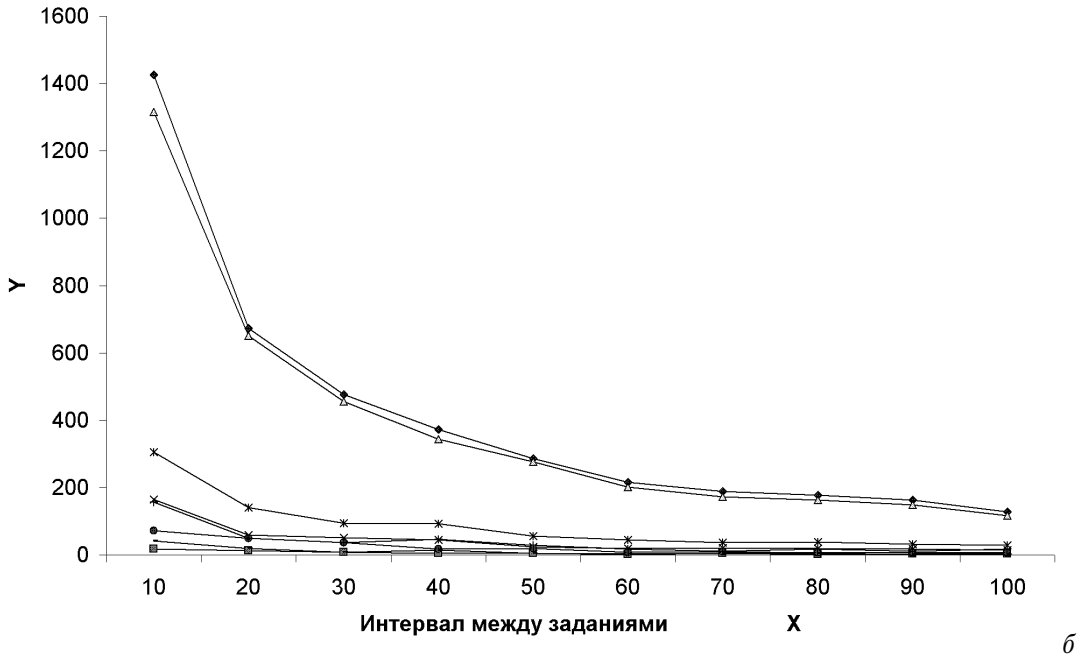
На основе результатов моделирования был спроектирован распределенный вычислительный комплекс обработки дендроэкологических данных (рис. 3). Архитектура комплекса базируется на компонентах, используемых при построении современных инструментальных средств распределенных вычислений (*grid*-систем). Но внутренняя структура комплекса, содержание функциональных компонентов (*middleware*) и протоколы их взаимодействия являются оригинальными. Архитектура разработанного комплекса виртуализирует три основных технических ресурса, из которых строится высокопроизводительный центр обработки данных (вычислительные системы, системы хранения данных и глобальные коммуникации), а затем собирает их в единый виртуальный компьютер, чтобы предоставлять его ресурсы в виде сервисов пользователям центра.

В структуре распределенного комплекса обработки дендроэкологических данных выделены вычислительные серверы, коммуникационный сервер и сервер базы данных.

Планирование и диспетчирование процессами обработки дендроэкологических данных возложено на специальный коммуникационный сервер. Взаимодействие всех пользователей с комплексом



а



б

Количество заданий общее
Количество отклоненных заданий
Количество заданий без параллелелизма
Количество заданий с параллелизмом
Количество аварийных ситуаций с восстановлением
Количество аварийных ситуаций без восстановления
Количество аварийных ситуаций с восстановлением (для заданий с параллелизмом)
Количество аварийных ситуаций без восстановления (для заданий с параллелизмом)
Количество заданий в очереди

Рис. 2. *Результаты моделирования распределенного вычислительного комплекса обработки дендроэкологических данных зависимости критериев от: а) количества типов заданий; б) интервала между заданиями*

обработки осуществляется только через коммуникационный сервер. Основной задачей коммуникационного сервера является обеспечение оптимальной загрузки имеющихся в его распоряжении вычислительных серверов и обеспечение режима работы в реальном времени (*online*) пользователей. В такой ситуации пользователю уже не важно, на каком конкретном узле сети исполняется его задача; он просто потребляет определенное количество виртуальной процессорной мощности, имеющейся в сети.

На вычислительных серверах комплекса производится математическая обработка данных.

Для полноценного функционирования комплекса анализа дендроэкологических данных организован автоматизированный сбор, систематизация и хранение научной информации в области дендроэкологического мониторинга, а также формирование и ведение базы дендроэкологических данных. Выполнение этих задач возложено на менеджер данных, управляющий работой распределённых серверов баз данных.

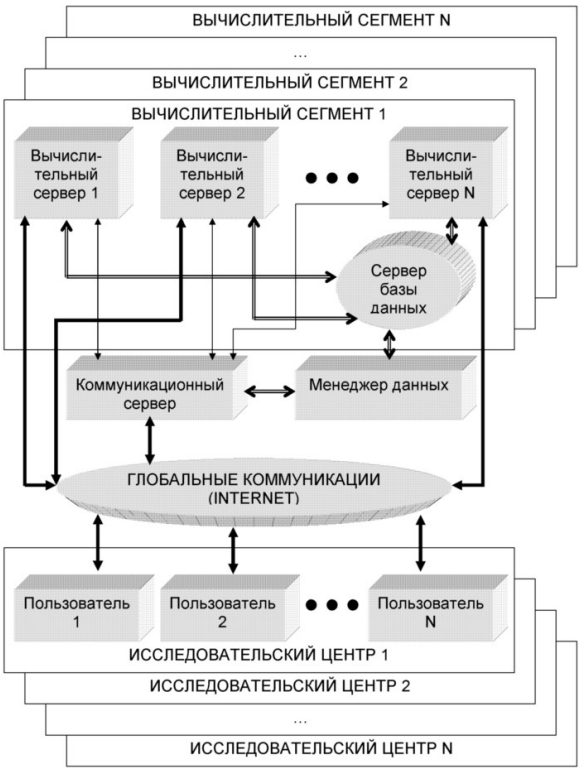


Рис. 3. *Распределенный вычислительный комплекс обработки дендроэкологических данных*

На рис. 4 представлена схема разработанного алгоритма выбора вычислительного сервера коммуникационным сервером.

Алгоритм использует информацию, которая хранится в профилях (метаописаниях) вычислительных серверов (таблица процессоров, таблица соответствия задач и таблица задач).

Основные шаги алгоритма:

Шаг 1. Формирование списка вычислительных серверов для решения задачи.

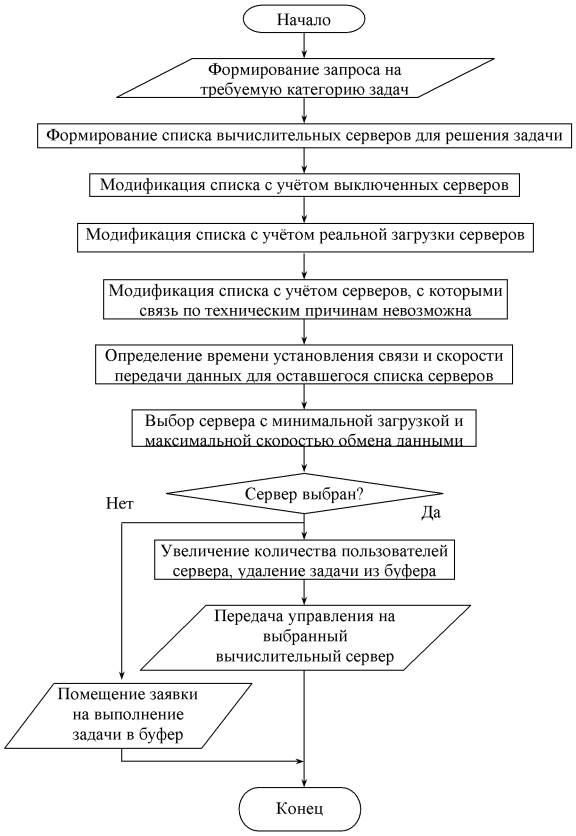


Рис. 4. *Алгоритм функционирования коммуникационного сервера*

Шаг 2. Исключение из списка выключенных серверов.

Шаг 3. Исключение из списка занятых серверов.

Шаг 4. Исключение из списка серверов, с которыми связь по техническим причинам невозможна.

Шаг 5. Определение времени установления связи и скорость передачи данных с серверами.

Шаг 6. Определение сервера с максимально свободной вычислительной мощностью.

Шаг 7. При наличии нескольких серверов с одинаковой вычислительной мощностью выбирается тот сервер, с которым скорость обмена данными выше.

Шаг 8. Увеличение на единицу значения поля Количество пользователей.

Шаг 9. При отсутствии серверов со свободными вычислительными мощностями заявка на выполнение задачи помещается в буфер коммуникационного сервера.

Шаг 10. Передача управления на выбранный вычислительный сервер.

Критерий выбора заключается в поиске такого сервера, нагрузка на который минимальна, а скорость обмена данными – максимальна.

На рис. 5 представлена схема разработанного алгоритма менеджера данных (СУБД – система управления базой данных).

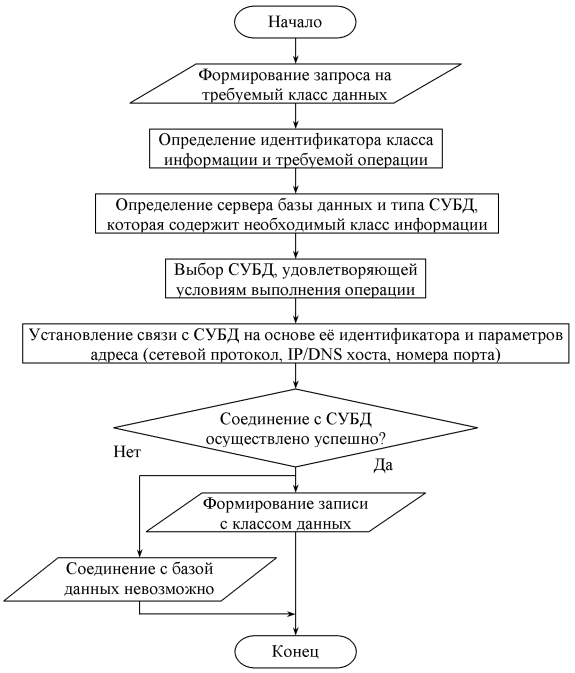


Рис. 5. *Алгоритм функционирования менеджера данных*

На менеджер данных (рис. 6) возложена работа по взаимодействию с серверами баз данных. Именно менеджер данных координирует распределение и использование информации, находящейся в локальных базах данных и тем самым виртуализирует накопители данных, объединяя их в единый логический информационный ресурс. Функционирование менеджера данных основывается на информации, формируемой в системной базе данных в таблицах серверов баз данных, таблицах соответствия классов информации и описания классов информации и СУБД.

Для практической реализации комплекса в настоящей работе использовался стек программных продуктов LAMP – набор свободно распространяемого инструментария: HTTP сервер Apache 1.3.14, SQL СУБД MySQL 3.22.21, язык сценариев PHP 4.2, а также сервер баз данных Oracle 8i для постоянного хранения информации. Отметим, что *grid*-системой, реализованной с использованием стека LAMP, является, например, компонент GridPrems французского *grid*-проекта Grid'5000.

Таким образом, с использованием разработанного алгоритма двумерного анализа изображений дисков деревьев созданный распределенный вычислительный комплекс обработки дендроэкологических данных способен решать такие задачи, как вычисление азимута и среднеквадратичной ширины области максимального прироста, а также вычисление значений индексов прироста ширины годичных колец деревьев. Также, с помощью разработанных программных средств математического анализа годичных колец деревьев появилась возможность выявления изменений параметров окружающей среды, отраженных в приросте дерева, и решения задач дендроэкологической диагностики с использованием дополнительной картографической и метеорологической информации. Более подробная информация о решенных с помощью комплекса задачах двумерного анализа изображений дисков деревьев представлена в работах [3, 4].

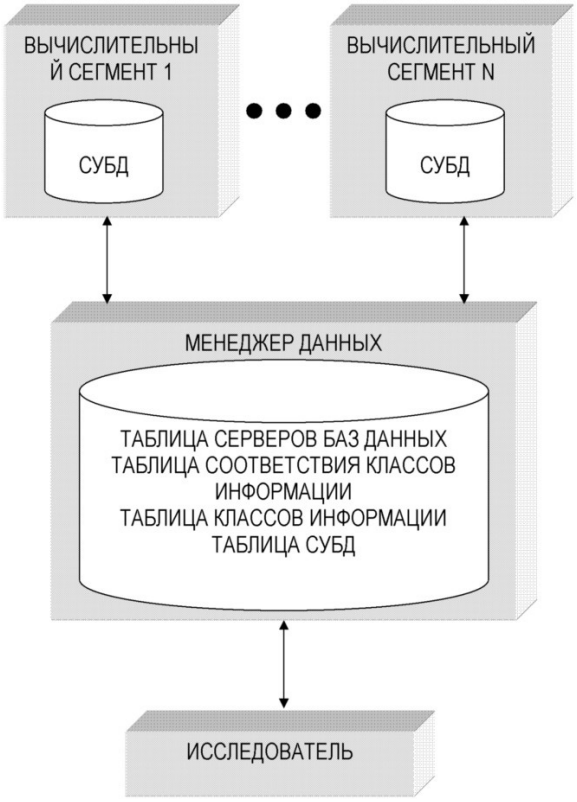


Рис. 6. *Менеджер данных*

Выводы

1. Разработаны алгоритмы и программные средства обработки дендроэкологических данных, которые могут быть применены в мониторинге экологического состояния окружающей среды.
2. Разработанная и постоянно пополняемая база данных изображений годичных колец деревьев обеспечивает использование полученной невозобновимой информации в будущих работах на основе новых методов и разработок.
3. Предложенная функциональная структура распределенного вычислительного комплекса может быть использована для решения широкого круга задач, требующих значительных вычислительных мощностей.
4. Использование разработанного распределенного вычислительного комплекса для решения задач дендроэкологического анализа позволяет упростить их технологические аспекты решения и значительно увеличить мощность вычислительного комплекса.
5. Созданные программные средства могут быть использованы в качестве лабораторной базы

в различных учебных дисциплинах, связанных с охраной окружающей среды и экологией человека и рекомендованы к внедрению в лесохозяйственных организациях и при подготовке специалистов в области лесоведения в университетах России.

СПИСОК ЛИТЕРАТУРЫ

1. RINNTECH // Products and services for tree and wood analysis. 2010. URL: http://www.rinntech.de/ (дата обращения: 27.08.2010).
2. GPSS // Имитационное моделирование систем. 2010. URL: http://www.gpss.ru/ (дата обращения: 27.08.2010).
3. Ботыгин И.А., Попов В.Н., Тартаковский В.А. Математические модели в задачах обработки дендроэкологических данных. Ч. I // Известия Томского политехнического университета. – 2011. – Т. 319. – № 5. – С. 118–122.
4. Ботыгин И.А., Попов В.Н., Тартаковский В.А. Математические модели в задачах обработки дендроэкологических данных. Ч. II // Известия Томского политехнического университета. – 2011. – Т. 319. – № 5. – С. 123–125.

Поступила 29.06.2011 г.

УДК 004.4:004.89

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИИ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

Ле Хоай, А.Ф. Тузовский

Томский политехнический университет
E-mail: lehotomsk@yahoo.com

Рассматривается использование онтологий в семантических электронных библиотеках, дается их определение и назначение. Анализируются виды онтологий таких библиотек, в том числе системы организации знаний и структурная таксономия. Обосновывается вариант набора онтологий для разработки семантических электронных библиотек.

Ключевые слова:
Онтология, контрольный словарь электронных ресурсов, семантические технологии, электронная библиотека, семантическая электронная библиотека.

Key words*:*
Ontology, glossary of electronic resources, semantic technology, electronic library, semantic digital library.

Под электронными библиотеками (ЭБ) понимаются информационные системы, позволяющие автоматизировать работу пользователей с электронными ресурсами (ЭР), такими, как документы, изображения, аудио-, и видеофайлы и т. д. С появлением семантических технологий (СТ), предоставляющих средства работы с семантикой документов, возникла возможность разработки подходов к автоматизации работы с этими ресурсами на новом уровне. Разработка семантических электронных библиотек (СЭБ) представляет собой решение комплекса задач, целью которых являются повышение возможностей взаимодействия с пользователями и расширение функциональности ЭБ, особенно в поиске данных. Многие ЭР содержат метаданные, в том числе документы включают данные об авторе, издании, дате создания и т. д. Такие метаданные часто хранятся в виде XML файлов, позволяющих выполнять универсальное описание ЭР. Язык XML позволяет описывать только структуру объектов, а не их семантику. В свою очередь, СТ представляют возможность описывать семантику ЭР (аннотировать их) и выполнять программную обработку таких метаданных.

Для описания семантики метаданных необходимо использовать онтологические модели (онтологии), определяющие наборы понятий предметной области и их взаимосвязи. Для работы с онтологиями в СТ имеются специальные языки для описания семантики (OWL, RDFS, RDF) [1–3] и запросов к семантическим данным (SPARQL) [4], а также набор инструментов редактирования и работы с семантическими данными (Protégé, Sesame, Jena и т. д.) [5–7].

При разработке СЭБ с использованием СТ одной из наиболее важных и сложных задач является разработка онтологий, описывающих области знаний, с которыми связано функционирование ЭБ и содержание имеющихся в них ресурсов. В данной статье рассматриваются особенности и проблемы, связанные с задачей построения онтологий для СЭБ, и обосновывается базовый набор таких онтологий.

1. Использование онтологии в СЭБ

В области искусственного интеллекта под онтологией понимается специальная система понятий и взаимосвязей между ними, описывающая определенную предметную область. Содержание понятий определяется с помощью концептов. Формально в онтологии концепт отождествляется с объектом (классом), имеющим связи с другими класса-