Грид-инфраструктура предметных посредников, движимых приложениями, для решения задач над множеством неоднородных распределенных информационных ресурсов

Д.О. Брюхов[I], А. Е. Вовченко[I], О.П. Желенкова[II], В.Н. Захаров[I], Л.А. Калиниченко[I], Д.О.Мартынов[I], Н.А. Скворцов[I], С.А. Ступников[I]

[I]Институт проблем информатики РАН
[II]Специальная астрофизическая обсерватория

Abstract

The article considers the middleware architecture of subject mediators in the hybrid grid-infrastructure for scientific problem solving over a set of heterogeneous distributed information resources (such as databases, services, ontologies) integrated by the mediators. The infrastructure is hybrid because it is constructed as a binding of the AstroGid Virtual Observatory system developed in the UK and of the middleware supporting subject mediators developed at the Institute of Informatics Problems of RAS. In the mediator middleware an approach driven by applications is used. The hybrid grid-infrastructure is planned to be used for solving of the Russian Virtual Observatory problems.

1 Введение

В различных областях науки наблюдается экспоненциальный рост объема получаемых экспериментальных (наблюдательных) данных. Сложность использования таких данных увеличивается еще и вследствие их естественной разнородности. Разнообразие (информационная несогласованность) получаемой информации вызывается не только большим числом организаций, производящих наблюдения, и их независимостью, но и разнообразием объектов наблюдения, непрерывным и быстрым совершенствованием техники наблюдений, вызывающим адекватные изменения структуры и содержания накапливаемой информации. Это приводит к необходимости использования неоднородной, распределенной информации, накопленной в течение значительного периода наблюдений технологически различными инструментами.

Увеличивающийся разрыв между исследователями и источниками данных и сервисов приводит к необходимости поиска новых путей создания информационных систем, в которых особое внимание было бы сосредоточено на специальных средствах организации решения задач над множеством распределенных информационных ресурсов (данных и программ), накапливаемых в разнообразных научных центрах. Разработан (разрабатывается) ряд инфраструктур, которые технически способствуют организации решения задач в такой среде. Среди них Веб-сервисы, Гриды данных, Семантический Веб, инфраструктуры онтологического моделирования, интеграции информационных ресурсов, интероперабельные инфраструктуры промежуточного слоя и др.

Настоящая статья в качестве применений ограничивается рассмотрением инфраструктур информационных систем для науки, которые в последнее время приобретают вид международных виртуальных обсерваторий (ВО). ВО, в свою очередь, могут быть отнесены к классу корпоративных информационных систем, поэтому результаты настоящей работы применимы и к последним. Анализ инфраструктуры Российской ВО (РВО) для астрономии дан в ее аванпроекте [1].

В Великобритании развивается система АстроГрид [2] как полная инфраструктура для создания систем решения научных задач в астрономических ВО. АстроГрид позволяет конструировать гибкие распределенные структуры ВО, компонентами которых могут быть разнообразные средства хранения данных и доступа к ним, средства размещения и временного хранения файлов в процессе решения задач группами пользователей, реестры метаданных, в которых регистрируются ресурсы ВО, средства программирования приложений и организации их интероперабельного исполнения, средства создания и исполнения потоков работ, которые позволяют решать сложные задачи в распределенной системе.

Остальная часть статьи структурирована следующим образом. В разд. 2 рассматривается методология интеграции неоднородных ресурсов в них. В разд. 3 рассматриваются особенности инфраструктуры системы АстроГрид. В разд. 4 приведено краткое описание объединенной

архитектуры АстроГрида и средств поддержки исполнительного слоя предметных посредников. Заключение статьи подводит итог обсуждению и намечает планы дальнейшего развития работы.

2 Методология интеграции неоднородных ресурсов в посредниках

Одним из широко распространенных взглядов на ВО является рассмотрение их как инфраструктур, предназначенных для интеграции данных и сервисов в различных исследовательских центрах, с целью предоставления таким образом всем ученым доступа к информации, необходимой для решения задач. Различаются два принципиально различных подхода к проблеме интегрированного представления описания предметной области задачи по отношению к множеству релевантных задаче информационных ресурсов: (1) двигаясь от ресурсов к задачам (схема посредника образуется как интегрированная схема множества ресурсов независимо от приложения) и (2) двигаясь от приложения к ресурсам (описание предметной области приложения образуется независимо от ресурсов (в терминах понятий, структур данных, функций, процессов), а затем релевантные приложению ресурсы отображаются в это описание). Первый подход, *движимый информационными ресурсами*, является не масштабируемым по отношению к числу ресурсов, не дает возможности достижения семантической интеграции ресурсов в контексте конкретного приложения, не ведет к доказательной идентификации релевантных приложению ресурсов, не способствует повышению стабильности спецификации посредника в процессе эволюции ресурсов, релевантных приложению. Эти недостатки являются характерными для подхода, при котором глобальная схема является взглядом (Global as View — GAV) [3, 4]. Схема GAV может служить в качестве базовой техники подхода, движимого информационными ресурсами.

Другой подход (*движимый приложениями*) предполагает создание предметного посредника, который поддерживает взаимодействие между приложением и ресурсом на основе определения прикладной области (определения посредника). Второй подход имеет очевидные преимущества по отношению к подходу, движимому информационными ресурсами. Процесс регистрации неоднородных информационных ресурсов в предметном посреднике в подходе, движимом приложениями, основан на технике GLAV[5] , комбинирующей два подхода: LAV (Local as View) — когда локальная схема ресурса является взглядом над схемой посредника — и GAV. Согласно LAV [15], схемы регистрируемых ресурсов рассматриваются как материализованные взгляды над виртуальными классами посредника. В этом случае GAV взгляды служат для разрешения различных конфликтов между спецификациями ресурсов и посредника и обеспечивают правила трансформации результатов запроса в формате ресурса в представление в посреднике. Подобная техника регистрации обеспечивает стабильность спецификации приложения ВО при изменении конкретных информационных ресурсов и их фактического присутствия (удаление ресурса, добавление новых ресурсов и пр.), а также масштабируемость посредников по отношению к числу ресурсов, регистрируемых в них. В версии GLAV голова определения правила взгляда LAV может содержать произвольный запрос над схемой ресурса аналогично GAV.

Настоящая статья основана, главным образом, на подходе, движимом приложениями.

3 Особенности инфраструктуры системы АстроГрид

Цель системы АстроГрид — поддержка инфраструктуры для решения научных задач в астрономических ВО. Для решения научных задач АстроГрид предоставляет средства доступа к астрономическим каталогам, цифровым обзорам и архивам изображений, а также к реестрам метаданных, в которых регистрируются ресурсы ВО. Система обеспечивает размещение файлов и хранение промежуточных результатов в процессе решения задач группами пользователей, средства доступа к реестрам метаданных, в которых регистрируются ресурсы ВО, средства создания и исполнения потоков работ, средства взаимодействия между различными внешними приложениями в распределенной среде ВО. Далее рассматриваются основные компоненты АстроГрида.

Registry (реестр) представляет собой коллекцию метаданных—XML-документов, описывающих ресурсы, которые могут использоваться при решении задач с помощью ВО. Registry реализован на основе стандарта OAI PMH, специализированного IVOA (Альянс Международной Виртуальной Обсерватории [6]) для нужд ВО.

Community обеспечивает регистрацию и персональную аутентификацию пользователей.

MySpace представляет собой виртуальное хранилище данных, к которым могут иметь доступ все сервисы системы АстроГрид.

Common Execution Architecture (CEA — Общая исполнительная архитектура) определяет способ оформления приложения в виде сервиса АстроГрида.

DataSet Access (DSA) реализует подключение базы данных к системе АстроГрид.

Workbench—это клиент системы АстроГрид, который позволяет работать с астрономическими ресурсами.

4 Гибридная архитектура АстроГрида и исполнительных средств поддержки слоя предметных посредников

В результате анализа архитектур средств поддержки слоя посредников и АстроГрида выработаны следующие основные решения гибридной архитектуры:

– посредник реализуется как CEA-приложение и регистрируется в реестре АстроГрида при помощи сервера-адаптера приложений. Метаданные, которыми характеризуется регистрируемый посредник, расширяются схемой посредника;

– на интерфейсе посредника имеются методы задания запросов на Syfs – подмножестве языка формул в языке СИНТЕЗ [7] и на ADQL[8];

– результаты запросов к посреднику представляются в формате VOClass, для которого формат VOTable [6] является подмножеством, и сохраняются в MySpace;

– для подключения к посредникам DSA ресурсов, уже зарегистрированных в системе АстроГрид, используется адаптер к DSA ресурсам;

– для осуществления поиска по метаданным ресурсов или приложений в реестрах системы АстроГрид используется адаптер к реестрам АстроГрида, реестр регистрируется в посреднике как обычный ресурс;

– в качестве клиентского интерфейса и приложения используется Workbench;

– для поддержки астрономических онтологий, принятых в системе АстроГрид, разрабатываются средства онтологической интеграции на основе UCD (Unified Content Description) — унифицированных дескрипторов контента, широко используемых в астрономии. В данном разделе рассматриваются, в основном, те компоненты гибридной архитектуры, которыми дополняется система АстроГрид, а также исполнительные средства слоя посредников.

Посредник может регистрироваться в реестрах АстроГрида в двух вариантах: как CEA-приложение или как DSA-компонент. Стоит отметить, что один и тот же посредник может быть зарегистрирован одновременно как CEA и как DSA.

Посредник, зарегистрированный как CEA-приложение, выполняет запрос на языке Syfs над объектной схемой посредника и возвращает результат в формате VOClass. Для реализации этой возможности разработан компонент *M-CEA*, который инкапсулирует конкретный посредник и может быть зарегистрирован в реестре АстроГрида как CEA приложение.

Посредник, зарегистрированный как DSA, выполняет запрос на языке ADQL над уплощенной схемой посредника и возвращает результат в формате VOClass. Поскольку схема посредника объектная, а ADQL—язык запросов к реляционной схеме, необходим специальный компонент посредника — *преобразователь схем*, который по объектной схеме посредника строит ее уплощение (реляционное представление). Запрос на ADQL формулируется в терминах плоской схемы, а затем компонентом *ADQL2Syfs* транслируется в запрос на языке Syfs к исходной объектной схеме посредника. *Преобразователь схем* и *ADQL2Syfs* используются разработанным компонентом *M-DSA*, который инкапсулирует конкретный посредник и может быть зарегистрирован в реестре АстроГрида как DSA-приложение. Оформление посредника как DSA — важная особенность, так как предоставляет пользователям АстроГрида возможность видеть посредник в привычном виде как DSA-источник.

В любом случае посредник оформляется в качестве CEA-приложения, которое может быть использовано как шаг потока работ или как самостоятельное приложение. Таким образом, обеспечивается возможность использования посредника в системе АстроГрид, а также возможность удобного просмотра схемы посредника прямо из клиента Workbench системы АстроГрид.

Адаптеры обеспечивают преобразование запросов на языке планов посредника в запросы на языке ресурса, получение результата запроса от другого ресурса, а также преобразование объектов результирующего множества в объекты схемы посредника.

В действующем прототипе разработаны адаптер реляционных баз данных (*RelationalWrapper*), адаптер к астрономическому каталогу данных SDSS (*SDSS Wrapper*) с поддержкой возможности

выполнения XMatch на сервере SDSS, адаптер DSA ресурсов (*DSA Wrapper*) и адаптер реестров АстроГрида (*RegistryWrapper*).

Для обеспечения возможности использования DSA ресурсов, зарегистрированных в реестрах АстроГрида, был разработан адаптер к DSA ресурсам. Этот адаптер играет одну из ключевых ролей по реализации гибридной архитектуры, обеспечивая возможность конструирования посредников над ресурсами системы АстроГрид.

Реестр регистрируется в посреднике как обычный ресурс, к которому могут быть сформулированы запросы, что позволит идентифицировать ресурсы, уже зарегистрированные в системе АстроГрид, и использовать их для конструирования новых посредников.

Подробно гибридная архитектура посредников и АстроГрида рассмотрена в [9].

5 Заключение

В настоящей статье рассматриваются первые результаты создания промежуточного слоя предметных посредников в гибридной грид-инфраструктуре виртуальной обсерватории для решения научных задач над множеством интегрируемых посредниками неоднородных распределенных информационных ресурсов. Введение такого промежуточного слоя призвано решить ряд семантических проблем взаимодействия ученого, решающего задачу в некоторой предметной области, и разнообразных релевантных задаче результатов наблюдений и средств их обработки. Гибридная архитектура ВО реализована как объединение инфраструктуры системы поддержки ВО АстроГрид, разработанной в Великобритании, и средств поддержки исполнительного слоя предметных посредников, созданных в ИПИ РАН. В исследованной архитектуре предметных посредников реализован подход, движимый приложениями, при котором для класса приложений формируется спецификация предметной области независимо от существующих информационных ресурсов. Далее происходит идентификация ресурсов, релевантных задаче, и их регистрация в посреднике на основе техники GLAV.

Полученные результаты свидетельствуют о перспективности исследованного подхода, существенное развитие которого планируется в ряде направлений. Планируется использование гибридной грид-инфраструктуры при решении разнообразных задач РВО. Возможности разработанной инфраструктуры были продемонстрированы при решении задачи поиска далеких галактик [10,11].

Литература

1. Briukhov D. O., et al. Information infrastructure of the Russian Virtual Observatory (RVO). 2nd ed. — IPI RAN, 2005. 173 p.
2. AstroGrid. http://www.astrogrid.org.
3. Ullman J.D. Information integration using logical views // 6th International Conference on Database Theory (ICDT'97) Proceedings, 1997.
4. Alon Y. Halevy. Answering queries using views: A survey // VLDB J. 2001. Vol. 10. No. 4.
5. Friedman M., Levy A., Millstein T. Navigational plans for data integration // National Conference on Artificial Intelligence (AAAI) Proceedings, 1999.
6. Virtual observatory architecture overview. Version 1.0, IVOA Note 15 June 2004.
7. Kalinichenko L. A., Stupnikov S. A., Martynov D. O. SYNTHESIS: A language for canonical information modeling and mediator definition for problem solving in heterogeneous information resource environments. — M.: IPI RAS, 2007. 171 p.
8. IVOA Astronomical Data Query Language Version 1.01.
http://www.ivoa.net/Documents/WD/ADQL/ADQL-20050624.pdf.
9. Д.О. Брюхов, А. Е. Вовченко, О.П. Желенкова, В.Н. Захаров, Л.А. Калиниченко, Д.О.Мартынов, Н.А. Скворцов, С.А. Ступников. Архитектура Промежуточного слоя Предметных Посредников для Решения Задач над Множеством Интегрируемых Неоднородных Распределенных Информационных Ресурсов в Гибридной Грид-Инфраструктуре Виртуальных Обсерваторий. Информатика и ее Применения, 2008. Т. 2. Вып. 1. С. 2-34
10. Гибридная архитектура АстроГрида и Посредников.
http://synthesis.ipi.ac.ru/synthesis/projects/Z_Mediator/index2_html
11. Задача поиска далеких галактик в системе АстроГрид.
http://synthesis.ipi.ac.ru/synthesis/projects/Z_DGS/index2_html