Поиск источников информации по метаданным в предметном посреднике♣

© Малиновский И. А.

ИПИ РАН
to_ilia@mail.ru

Аннотация

В статье рассмотрены вопросы поиска источников информации в контексте инфраструктуры предметных посредников. В силу широкого распространения реестров метаданных источников информации, в качестве критерия поиска выбраны метаданные посредника и выражаемые с их помощью требования к информации посредника. Показана связь между схемами информации посредника и метаданными.

1 Введение

Использование предметных посредников для гетерогенных информационных источников предоставляет гибкий подход к интеграции информации. Архитектура посредника, рассмотренная в [5,6], определяет промежуточный слой, расположенный между информационными источниками (поставщиками информации) и потребителями информации. Посредники предоставляют унифицированный интерфейс для решения задач над множественными источниками информации, тем самым, освобождая пользователя от необходимости находить релевантные источники, задавать запросы к каждому из них по отдельности и вручную сопоставлять информацию из них.

Использование предметного посредника может осуществляться в разнообразных предметных областях: в науке, культуре, СМИ, электронной торговле, и т.д. Такие области могут получить ряд преимуществ от использования подхода предметных посредников, поддерживая интеграцию информации для конкретных классов задач предметной области. Применение инфраструктуры предметных посредников планируется, в частности, в проекте Российской Виртуальной Обсерватории (РВО) [3], в котором предметный посредник является ориентированным на описание предметной области для класса астрономических задач, решаемых над множественными, неоднородными источниками информации.

Для таких областей, согласно используемому подходу, модель предметной области посредника должна быть определена экспертами в этой области независимо от релевантных источников информации. Эта модель может включать спецификацию структур данных; используемых терминов (тезаурус); онтологических понятий; методов; процессов (потоков работ), и других характеристик предметной области. После того, как предметный посредник был специфицирован, поставщики информации могут предоставлять свою информацию для регистрации в предметном посреднике независимо друг от друга и в любое время. Пользователи могут ничего не знать о процессе регистрации и о тех источниках, которые были зарегистрированы. Пользователи должны знать только спецификацию предметного посредника, которая содержит определение понятий предметной области, структур данных, функций, процессов, согласованных внутри сообщества данной предметной области.

При регистрации формулируются спецификации источников (схемы, определения понятий, словари) в терминах метамодели предметного посредника и разрабатывают необходимые адаптеры. Во время этого процесса формируются выражения, которые определяют локальные схемы источников в канонической модели посредника, как взгляды над схемой посредника. Обеспечение регистрации источников независимыми поставщиками позволяет достигнуть свойства масштабируемости посредника.

К настоящему времени в мире накоплено большое количество источников информации, которые потенциально могут быть использованы в предметном посреднике. Например, только реестр метаданных астрономических источников информации NVO [3] содержит более 11000 записей. Очевидно, что регистрировать в посреднике необходимо только те источники, которые могут быть полезны в нем, а для осуществления этого необходима разработка методов и средств формулировки критериев релевантности источников информации, и осуществления по ним предварительного поиска. Общеизвестно, что технологии, решающие задачу публикации источников информации, такие, как,

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

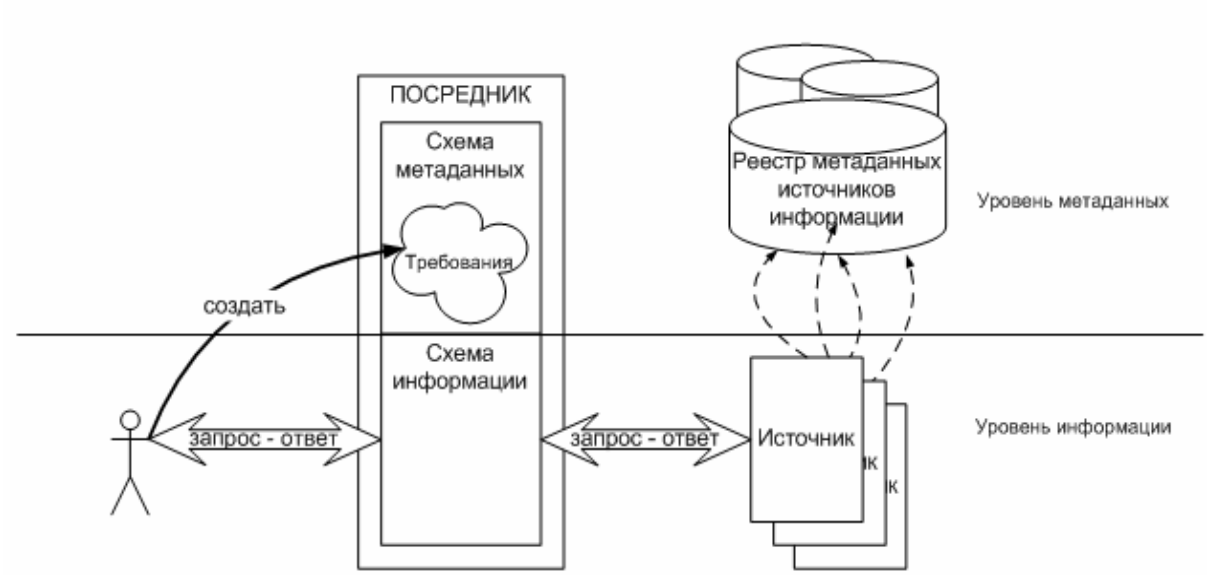

Рисунок 1. Общая схема инфраструктуры посредника

например, реестры UDDI [1515], или реестры, поддерживающие протокол обмена метаданными OAI-PMH [8], не обладают достаточными сервисами для поиска источников с нужными характеристиками. В ряде работ (например [14, 12, 10]) предлагаются методы, расширяющие возможности реестров использованием онтологий в спецификациях источников. В таких работах приводятся алгоритмы поиска в реестрах, расширенных соответствующим образом. Не известны работы, посвященные поиску источников информации в контексте задачи регистрации источников информации в посредниках. В настоящей работе рассматривается метод идентификации источников информации по метаданным для последующей их регистрации в посреднике. Работа ведется в рамках проекта создания предметных посредников [5]. Разработан прототип программного средства, реализующего поиск по метаданным предлагаемым способом релевантных посреднику источников, в соответствии с архитектурой посредника [6].

Оставшаяся часть статьи организована следующим образом: понятие метаданных конкретизируется во втором разделе. В третьем разделе обсуждается концепция модели требований к информации посредника. Раздел 4 посвящен детальному описанию механизма поиска по метаданным. В пятом разделе приводится краткий обзор родственных по теме работ.

2 Метаданные

Метаданные информационного источника описывают информацию в самом источнике. Метаданные определяют ортогональный основному уровню описания информации (который формируется такими понятиями, как классы, типы данных и др.) уровень описания свойств информации. Метаданные могут входить в стандарт Dublin Core, или его расширение, более специфическое для предметной области. Метаданные также могут описывать схему информационного источника, например реляционную, объектную или слабоструктурированную. Другими словами, схема данных источника или посредника является метаданными источника или посредника соответственно.

3 Концепция модели требований

Метаданные представляют собой первичные данные об источниках информации, доступные при помощи независимых от посредника реестров информационных источников. Поэтому естественным образом возникает необходимость в использовании метаданных источников для проверки релевантности информации в них задачам, которые решаются при помощи посредников.

Схема посредника определяет структуру информации и поведение, необходимые для решения задач посредника. Наряду с этим при определении посредника необходимо задание дополнительных требований (обычно нефункциональных) к необходимой информации. Такую схему, в отличие от схемы посредника, будем называть схемой метаданных информации посредника. Итак, посредник определяется своей *схемой информации и схемой метаданных*. Схемы информации и метаданных задаются в рамках одной и той же канонической информационной модели и одной и той же онтологии. Требования к необходимой информации задаются в виде предикатов над схемой метаданных посредника.

Предполагается, что все источники изначально зарегистрированы в специальных *реестрах метаданных источников*, которые будут

использоваться посредником. В реестре источник представляется идентификатором, своими метаданными, семантика которых описывается определенной онтологией, и другой дополнительной информацией. Существование реестров не зависит от существования посредников. Общее представление инфраструктуры посредника дано на рисунке 1.

Для спецификации метаданных и требований к ним на этапе создания схемы посредника вводится тип (типы) метаданных M, инвариант которого, по определению, является тождественно истинным. Требования вводятся как подтипы M_i типа метаданных, при этом сущность требований – предикат требований – выражается инвариантом I_{Mi} данного подтипа. Отмечается, что в подтипе-требовании M_i экстенсионал и набор операций при наследовании не меняются. Предикатом требований может быть формула без кванторов, все переменные которой являются атрибутами метаданных.

Иллюстрация метода поиска будет приводиться на следующем примере. В качестве схемы данных посредника далее в тексте будем считать, что созданы некоторые классы cities, addresses, houses, persons, спецификации типов экземпляров которых для рассматриваемого примера не существенны.

Для иллюстрации будем использовать примеры на языке СИНТЕЗ [1]. Для простоты изложения пусть схема метаданных посредника состоит из одного типа метаданных посредника:

{M; in: type;
id: string;
ageOfInformation: integer;}

Также в схеме метаданных посредника пусть созданы следующие два требования:

{R1Mediator; in: type; supertype: M;
inv:{in: invariant; {ageOfInformation < 3}};},

{R2Mediator; in: type; supertype: M;
inv:{in: invariant; {ageOfInformation < 5}};}

Чтобы распространить требование на определенный класс посредника, необходим, во-первых, метакласс, инвариант собственного типа (т.е. тип метакласса как объекта) которого содержит данное требование и, во-вторых, зафиксировать, что данный класс посредника является экземпляром данного метакласса. Каждый класс посредника может быть экземпляром одного или нескольких метаклассов. Некоторые классы могут быть экземплярами одного и того же метакласса. По умолчанию, любой класс посредника является экземпляром метакласса, собственный тип которого равен M, что означает, что на данный класс посредника не распространяются никакие требования, т.к. инвариант типа M является тождественно истинным. Итак, пусть в посреднике созданы следующие метаклассы:

{meta1Mediator; in: metaclass;
class_section: R1Mediator;},

{meta2Mediator; in: metaclass;
class_section: R2Mediator;}

Пусть класс посредника cities является экземпляром метакласса meta1Mediator, класс посредника addresses экземпляром метакласса meta2Mediator, а класс посредника houses экземпляром обоих метаклассов.

4 Поиск источников информации

Перед поиском в реестре тех источников, которые являются релевантными установленным требованиям, реестр необходимо *зарегистрировать* в посреднике. На данном этапе реестр рассматривается как обыкновенный источник данных, и над ним выполняются все шаги регистрации источников информации (подробнее в работе [2]), за единственным исключением, регистрация реестра происходит в схеме метаданных посредника, тогда как регистрация источника происходит в схеме информации посредника.

Результатом регистрации является построение серии особых утверждений – взглядов (views) [4] – определяющих классы источника при помощи выражений над классами посредника. Такие выражения конструируются как конъюнктивные запросы над классами посредника, т.е. запросы с семантикой выбор-проекция-соединение (selection-projection-join). Взгляды позволяют *переписывать* [5, 6] запрос, данный в терминах спецификации посредника, в запрос в терминах спецификации источника, что делает возможным выполнение запроса в источнике. Корректность построенных взглядов доказывается формально с использованием понятия *уточнения* [5, 6].

Для рассматриваемого примера предположим, что метаданные некоторого источника S содержатся в реестре, а тип метаданных, используемый в реестре, совпадает с типом метаданных посредника (тип M). Пусть значения метаданных этого источника будут *<id = "example", age = 5>*.

При инициализации поиска требования к метаданным посредника транслируются в запросы на языке, который поддерживается конкретным посредником, над классами схемы метаданных посредника. Используя результаты регистрации реестров в посреднике, этот запрос переписывается в терминах реальных реестров, а, следовательно, может быть выполнен. Для рассматриваемого примера будут созданы такие запросы:

req4cities(id):-
meta1Mediator(x) & ageOfInformation < 3;

req4addresses(id):- meta2Mediator(x) & ageOfInformation < 5;

req4houses(id):-

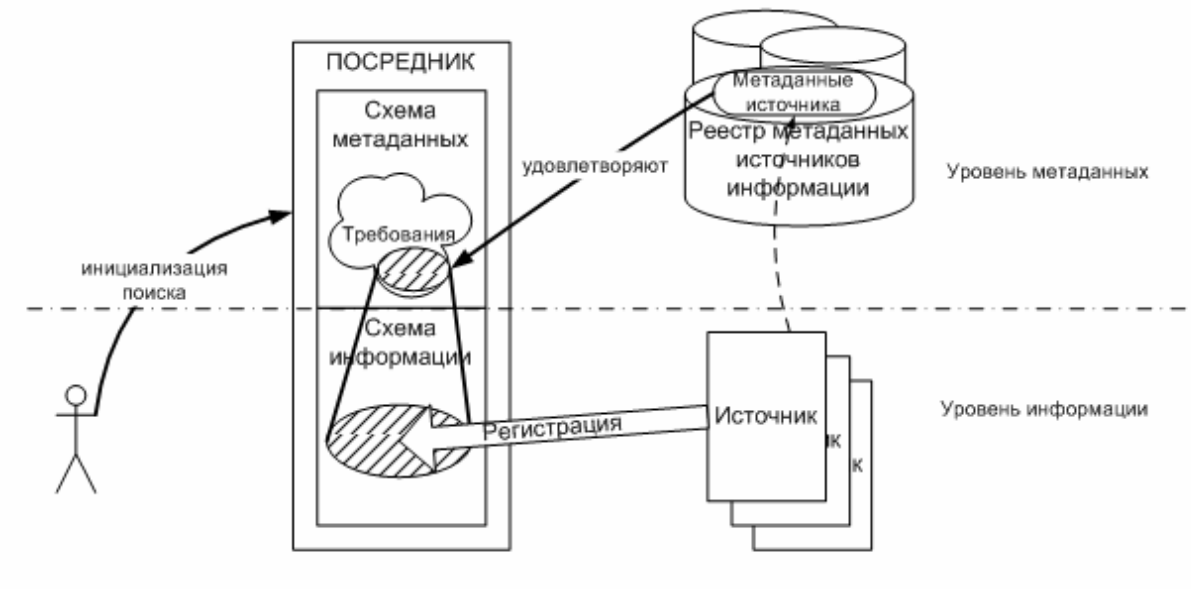

Рисунок 2. Интерпретация результатов поиска

meta1Mediator(x) & ageOfInformation < 3 & meta2Mediator(x) & ageOfInformation < 5;

Семантика запросов такова, чтобы в результате в req4cities были отобраны идентификаторы “id” тех источников, которые удовлетворяют требованиям, распространенным на класс посредника cities; в req4addresses находились id тех источников, которые удовлетворяют требованиям, которые распространены на класс посредника addresses; в req4houses находились id тех источников, которые удовлетворяют требованиям, которые распространены на класс посредника houses.

После выполнения запроса для каждого источника известно, требованиям каких классов посредника он удовлетворяет. Источники, которые оказались релевантными хотя бы одному требованию, допускаются к регистрации в посреднике. Во взглядах, которые будут строиться при регистрации источника, могут использоваться только те классы посредника, требованиям к которым он удовлетворяет. Другими словами, источник может регистрироваться только в той части схемы информации посредника, требованиям к которой удовлетворяют его метаданные. Данная ситуация схематично представлено на рисунке 2.

Источник S, который рассматривается в качестве примера, будет находиться в результатах только запроса req4addresses (т.к. значение age источника равно 4) и, следовательно, при его регистрации может быть использованы только следующие классы посредника: addresses и persons (т.к. на класс persons не распространялось никаких требований).

Интересным побочным результатом регистрации метаданных, в качестве которых может выступать схемы данных, является получение трансформационной схемы [9], которая в дальнейшем может стать основой для построения адаптера соответствующего источника.

5 Анализ существующих работ

В ряде статей [14, 12, 10], которые посвящены расширению возможностей поиска источников информации, а именно, достижению семантической интероперабельности описаний ресурсов, задача решается использованием некоторой общей онтологии. Перед разработчиками этих методов стоят специфические требования, которые заключаются в необходимости обеспечить полную автоматизацию процесса поиска. Эти требования диктуются предметной областью – поиск и композиция веб-сервисов в инфраструктуре Semantic Web. Во многих статьях предлагаются алгоритмы поиска соответствий в XML документах и приводятся их оценки [13]. Идея объектного трейдинга в [11] развивается в сторону введения некоторых специальных нечетких мер близости, определенных на IDL описаниях, дополняющих основной поиск по метаданным.

В то же время, известно лишь немного статей, обсуждающих задачу поиска источников информации в целом. В работе [7] формализуется задача поиска веб-сервисов, вводятся формальные понятия цели поиска сервиса. Разрабатываемая в этой работе модель является достаточно общей и может служить каркасом для многих других частных методов поиска.

Предлагаемая в настоящей работе модель и метод поиска рассматривает задачу поиска источников информации шире, чем поиск веб-сервисов. На модель не накладываются ограничения, связанные с необходимостью осуществить полностью автоматизированный поиск, поэтому стадия регистрации реестра является

автоматизированной. Примером рассматриваемой области применения предлагаемого подхода является e-science (в частности, проект Российской Виртуальной Обсерватории), где обычно полная автоматизация поиска невозможна из-за более высоких требований к логическому языку спецификаций, чем в Семантическом Вебе. При необходимости предлагаемая модель может быть легко ограничена, например, использованием дескриптивных логик вместо логики первого порядка или некоторыми вариантами нечеткого соответствия типов вместо операции уточнения.

Рассмотренные выше методы поиска не могут быть использованы напрямую для поиска в инфраструктуре предметного посредника, т.к. в последнем случае необходимо прозрачно отслеживать взаимосвязь между критериями поиска, заданными в терминах метаданных и результатом поиска, который, в конечном счете, является соответствием между источником и некоторым подмножеством (возможно, пустым) схемы посредника. Описанные выше методы ограничиваются поиском на определенном множестве критериев (метаданные), в терминах которых задается понятие релевантности источника целям поиска. В инфраструктуре предметного посредника понятие релевантности определено как возможность источника предоставить информацию, необходимую посреднику, поэтому понятие релевантности и критерии предварительного поиска по требованиям к информации, выраженным в метаданных, находятся на ортогональных уровнях измерения информации.

6 Заключение

В статье представлен метод поиска источников информации в инфраструктуре предметных посредников. Решаемая задача потребовала расширения структуры посредника за счет добавления уровня метаданных, ортогонального основному уровню спецификации информации. Требования к необходимой в посреднике информации задаются в терминах метаданных. Результат поиска, в конечном счете, является соответствием между источником и некоторым подмножеством схемы посредника. На основе данной модели поиска было разработано программное средство поиска источников информации, удовлетворяющих заданным при помощи метаданных требований. Это средство рассматривается как часть средств регистрации источников информации в посреднике.

Литература:

1. Калиниченко Л. А. “СИНТЕЗ язык определения, проектирования и программирования интероперабельных сред неоднородных информационных ресурсов” (вторая редакция), РАН, ИПИРАН, Москва 1993.
2. Briukhov D. O., Kalinichenko L. A., Skvortsov N. A. “Information Sources Registration at a Subject Mediator as Compositional Development” Proceedings of the Conference on Advances in Databases and Information Systems (ADBIS), Vilnius, September 2001.
3. Briukhov D.O., Kalinichenko L.A., Zakharov V.N., Panchuk V.E., Vitkovsky V.V., Zhelenkova O.P., Dluzhnevskaya O.B., Malkov O.Yu., Kovaleva D.A. Information Infrastructure of the Russian Virtual Observatory (RVO). M.: IPI RAN, 2005.
4. Halevy Y. “Answering Queries Using Views: A Survey”. The VLDB Journal, vol. 10(4), p 270-294, 2001.
5. Kalinichenko L.A. Mediation Infrastructure and Digital Libraries. Proceedings of the International Conference on Digital Libraries. New Delhi, February, 2004
6. Kalinichenko L.A., Briukhov D.O., Skvortsov N.A., Zakharov V.N., “Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections”, Institute of Problems of Informatics RAS, Second All-Russian Conference Digital Libraries 2000
7. Keller U., Rub´en Lara, Holger Lausen, Axel Polleres, and Dieter Fensel. “Automatic Location of Services”. Proceedings of the 2nd European Semantic Web Conference, Heraklion, Greece, May 2005.
8. The Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
9. Papakonstantinou Y., Ashish Gupta, Hector Garcia-Molina, Jeffrey Ullman “A Query Translation Scheme for Rapid Implementation of Wrappers”, Deductive and Object-Oriented Databases (DOOD) 95
10. Rajasekaran P., John Miller, Kunal Verma, Amit Sheth. “Enhancing Web Services Description and Discovery to Facilitate Composition” Proceedings of SWSWPC (In conjunction with ICWS'2004).
11. Reichl P., Dirk Thißen, Claudia Linnhoff-Popien. “How to enhance service selection in distributed systems”. Proceedings of the Intl. Conf. Dist. Computer Communication Networks---Theory and Applications, pages 114-123, Tel-Aviv, November 1996
12. ShaikhAli A., Omer F. Rana, Rashid Al-Ali, and David W. Walker. “UDDIe: An Extended Registry for Web Services”. Proceedings of the Workshop on Service Oriented Computing: Models, Architectures and Applications at SAINT Conference. IEEE Computer Society Press, 2003.

13. Smiljanic M., Maurice van Keulen, and Willem Jonker. “Using Element Clustering to Increase the Efficiency of XML Schema Matching”. Proceedings of the 2nd International Workshop on Challenges in Web Information Retrieval and Integration (WIRI 2006) In conjunction with The 22nd International Conference on Data Engineering (ICDE 2006), April 3rd, 2006, Atlanta, Georgia, USA
14. Srinivasan N., Massimo Paolucci, Katia Sycara. “Adding OWL-S to UDDI, implementation and throughput”. Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004) 6-9, 2004, San Diego, California, USA.
15. Universal Description, Discovery and Integration http://www.uddi.org/specification.html

Information Sources Search using Metadata at Subject Mediator

Malinovsky Ilia

The paper presents an approach for heterogeneous information sources identification at subject mediators. Subject mediators provide uniform interface for multiple heterogeneous information sources. It is known that growth of amount of information sources looks like an explosion so necessity for source search methods becomes obvious.

To achieve this intention the traditional schema of mediator information is extended with a metadata schema. It is composed of metadata types and metaclasses. Requirements for mediator information are introduced as invariants of metadata types. A metaclass serves as a relation between a requirement and classes of the mediator schema, the requirement apply to. A result of search is a correspondence between information source and some subset (possibly empty) of mediator’s schema.

Well-known search methods use some set of criteria and these criteria determine the term of the relevance of information source to the goal of search. Methods like those do not handle the needs for search at subject mediators. At subject mediators, the term of the relevance is determined as a possibility of information sources to provide information that is necessary for mediators. In other words, at the subject mediators, the term of the relevance and criteria of search constitutes an orthogonal level of information representation.

♣ работа выполняется при поддержке грантов РФФИ: 06-07-89188-а, 05-07-90413-в, 06-07-08072-офи.