ИНТЕГРАЦИЯ ДОКУМЕНТАЛЬНЫХ ИПС В ПОСРЕДНИКЕ НЕОДНОРОДНЫХ КОЛЛЕКЦИЙ ДАННЫХ ЭЛЕКТРОННЫХ БИБЛИОТЕК НА ПРИМЕРЕ ИПС ИРБИС

Леонтьев И.В., Калиниченко Л.А.,

Институт Проблем Информатики РАН

Максимов Н.В.

Российский Государственный Гуманитарный Университет

{ ileon@ipi.ac.ru, leonidk@synth.ipi.ac.ru }

Аннотация

К настоящему времени в мире накоплено большое количество разнородной информации, и актуальной стала проблема ее интеграции, что требует разработки новых методов, средств, применения передовых технологий. В проекте, в рамках которого проводится данная работа, интеграция коллекций данных осуществляется с помощью программного посредника, к которому каждая коллекция (Web-сервер, база данных, ИПС) подключается посредством соответствующего адаптера. На примере ИПС Ирбис в статье рассматриваются вопросы интеграции документальных ИПС, описана процедура регистрации новых коллекций, показан способ построения адаптеров, предназначенных для подключения коллекций библиографической информации к посреднику. Рассмотрены компонентная архитектура адаптера, его функции, особенности отображения системы типов и языка запросов посредника, основанного на модели Синтез, и ИПС Ирбис.

1. Введение

Одним из важных видов коллекций для создания электронных библиотек, наряду с различными базами данных и гипертекстовыми документами Web-серверов, являются базы данных в документальных ИПС. Настоящая работа выполняется при поддержке проектов РФФИ № 01-07-90084, № 00-07-90086. Задачей указанных проектов является исследование инфраструктур и создание прототипа среды, основными компонентами которой являются предметно-ориентированные посредники, обеспечивающие интегрированный доступ к неоднородным коллекциям данных [6]. Целью данной работы является рассмотрение вопросов интеграции документальных ИПС в посреднике на примере документальной ИПС Ирбис [10,15], предоставляющей доступ к коллекциям документов, накопленным в различных организациях Москвы (например, ИНИОН [16], ВНТИЦентр [2]). Коллекции различаются по тематике (экономика, право) и в общей сложности содержат более 6 млн. записей. Рассматриваются коллекции, в которых в качестве документов используются библиографические данные.

2. Посредник неоднородных коллекций данных

В основе создания посредника неоднородных коллекций лежит принцип канонической модели данных (Синтез [9]), в которую эквивалентным образом отображаются модели интегрируемых ресурсов и схемы их коллекций. В архитектуре посредника можно выделить два уровня представления данных – федеративный уровень, обеспечивающий интегрированное представление информации в определенной предметной области, и локальный уровень, соответствующий локальным подключаемым коллекциям.

Интеграция некоторой коллекции в посреднике предполагает выполнение следующих действий:

- получение локальной схемы коллекции в унифицированном представлении (каноническая модель);
- согласование контекстов коллекции и федеративной схемы, а также представление локального класса, соответствующего коллекции, в виде взгляда над классами федеративного уровня;
- подключение коллекции к посреднику посредством адаптера (или разработка адаптера, если его нет, для соответствующей ИПС).

Совокупность этих действий есть регистрация коллекции.

3. Федеративные и локальные схемы

Федеративный уровень содержит схему модели предметной области, в которой специализируется посредник. Пользователь задает запросы в терминах этой модели. В нашем примере мы ограничиваемся случаем, когда федеративный уровень содержит схему документа на языке Синтез. Схема документа на языке Синтез представляет собой спецификацию структуры документа ИПС в виде набора его атрибутов, типов, и поисковых возможностей для работы с документальными ИПС. Атрибутами типа документа являются поля документов, и для каждого поля задается его тип (текстовый, дата, число). Каждый тип атрибута характеризуется допустимыми операциями - в частности, текстовый тип предоставляет возможности контекстного поиска.

В данной работе для формирования схемы документа федеративного уровня выбран набор атрибутов, включающий в себя некоторые атрибуты Dublin Core [5], MARC [11], BIB-1 [14] Z39.50, документов ИНИОН. Допускается пополнение набора атрибутов федеративного уровня.

Схема каждой коллекции (в частности, конкретная документальная ИПС) представляется в посреднике в виде описания на локальном уровне системы типов, связанных отношениями тип/подтип. Описания задаются на языке Синтез с помощью средств базы метаданных.

4. Особенности документальных ИПС

Главной особенностью документальных ИПС является наличие в языке запросов операторов для осуществления контекстного поиска (например, поиск слов внутри предложения, внутри всего текста, поиск слов на указанном расстоянии друг от друга)

Возможности контекстного поиска для поля задаются указанием его типа в базе метаданных как текстового. Текстовый тип федеративного уровня и текстовый тип локального уровня могут различаться количеством и видом допустимых операций.

В частности, текстовый тип федеративного уровня поддерживает следующие операции:

- совпадения строк,
- вхождения подстроки,
- следования слов без учета порядка,
- следования слов с учетом порядка,
- ограничения поиска по тексту некоторой областью (напр, предложением),
- использования в запросе операций с тезаурусом (расширение запроса синонимами, близкими словами, переводами на другой язык)

Текстовый тип локального уровня не обязательно может поддерживать полный набор указанных операций. В частности, ИПС “Ирбис” поддерживает следующие операции:

- вхождение подстроки,
- следование слов без учета порядка,
- следование слов с учетом порядка,
- ограничение контекстного поиска пределами предложения,
- поиск документов с общими значениями в текущем и заданном полях.

5. Согласование контекстов и выражение локальных типов как взгляда над федеративным уровнем.

Процесс регистрации осуществляется с помощью разработанных в рамках проекта инструментальных средств [1], и значительно облегчается за счет автоматического поиска возможных соответствующих компонент, полуавтоматического устранения структурных расхождений (конфликтов), наличия визуальных средств построения согласований типов. В основе процедуры регистрации, реализованной указанными средствами, используется согласование типов на основе теории уточнения [7].

При помощи указанного средства реализуются:

- онтологическая интеграция (установление взаимосвязей между понятиями, основанное на вербальных определениях этих понятий, заданных на естественном языке),
- разрешение структурных конфликтов сигнатур типов локального и федеративного уровня,
- проектирование реализации локального класса посредством композиции классов федеративного уровня.

6. Адаптеры для подключения коллекций

Для подключения документальной ИПС к посреднику нужно разработать соответствующий адаптер. Адаптер (wrapper) – это программный модуль, отвечающий за взаимодействие его клиента (программы-посредника) с соответствующим данному адаптеру информационным ресурсом. В качестве языка запросов адаптера используется подмножество языка SOQL. Язык запросов SOQL – это вариант ODMG OQL [13], включенный в состав языка Синтез. Адаптер получает и преобразует запросы, записанные на SOQL, в инструкции, понимаемые

информационным ресурсом. Затем результат (данные, полученные от информационного ресурса) “упаковываются” в виде объектов и возвращаются клиенту.

Адаптер выступает как преобразователь модели данных конкретной ИПС (“Ирбис”) в каноническую модель. Отображение исходной модели данных в целевую выражается в конечном счете в преобразовании схем коллекций в исходной модели данных в схемы в целевой модели данных, и в преобразовании операторов манипулирования данными целевой модели в последовательность операторов исходной модели данных. В рамках конкретной схемы требуется сохранение информации и операций локального уровня при таком преобразовании [8].

6.1. Интерфейс адаптера

Одним из основных требований при реализации адаптеров, помимо поддержки подмножества языка запросов SOQL, было обеспечение унифицированного интерфейса взаимодействия посредника с любым адаптером такого класса. Для адаптера разработан интерфейс, использующий объекты и сигнатуры методов, соответствующих подмножеству спецификации JDBC 1.0, а также разработаны средства для представления структур данных и их значений, не указанных в спецификации JDBC, но использующихся на интерфейсе посредника и адаптера (в частности, структуры данных set).

Функциями адаптера являются:

- получение запроса от посредника в терминах локального уровня,
- преобразование запроса в соответствии с представлением коллекции на локальном уровне базы метаданных в запрос, поддерживаемый документальной ИПС,
- отправка запроса к коллекции,
- получение результата и его преобразование в представление, установленное интерфейсом адаптера.

6.2. Преобразование типов данных посредника и типов данных коллекции

Типы данных, которые предоставляет пользователю посредник, могут отличаться от типов данных, хранящихся в каждой из коллекций. Посредник использует объектную модель предметной области, в которой документы из ИПС представлены с помощью системы типов, атрибутов и методов. Документальная ИПС позволяет осуществлять поиск документов с использованием контекстного поиска, что отражено в системе типов и методах для работы с типом. Поля документа соответствуют атрибутам в спецификации типа документа, и также имеют свой собственный тип, например, базовый (строковый, целочисленный, дата), множественный (набор строк) или структура.

Пример. Пусть запрос к одной из коллекций под управлением документальной ИПС Ирбис возвращает HTML-документ вида:

Автор(ы): Артемьева И. Л.; Гаврилова Т. Л.; Клещев А. С.;
Коллективн.автор: ВИНИТИ

Переводн.заглавие: Система логических соотношений с атомарными объектами
Дата публикации: 0.0.96
П/серия: Информационные процессы и системы
Ключевые слова: экспертные системы; математические модели; системы логических соотношений; expert systems; mathematical models; logical correlation systems;

В данном примере поля документа “Автор(ы)” и “Ключевые слова” являются множественными, с фиксированным разделителем “;” между отдельными значениями. Указание разделителя для каждого из множественных полей при подключении новой коллекции позволяет адаптеру выявлять значения, формировать из них объект типа Set и передавать посреднику. Доступ к значениям типа Set осуществляется с помощью методов поэлементного извлечения, а также поиска и сравнения подмножеств. Для преобразования типов в адаптере используется трансформационная таблица, допускающая и поэлементное преобразование структурных типов.

Поле “Дата публикации” отображается в значение типа Date в формате, например, DD-MM-YY, также заданном при настройке адаптера. При необходимости в адаптере могут быть реализованы вспомогательные функции преобразований формата и/или значений.

Поля документа, допускающие контекстный поиск, отображаются в специальный текстовый тип с соответствующими операциями, например, CONTAINS (вхождение подстроки в текст), NEAR (близость слов), WITHIN (ограничение поиска некоторой областью). В частности, ИПС Ирбис поддерживает два различных оператора, учитывающих близость слов – оператор CTX, для которого порядок вхождения слов не важен, и NEAR, для которого порядок следования поисковых слов в тексте важен. Операторы CTX[n] и NEAR[n] отображаются соответственно в NEAR(n, FALSE) и NEAR(n, TRUE) текстового типа в посреднике.

Имена полей документа в ответе ИПС Ирбис и атрибутов в схеме документа посредника могут также различаться. Для их согласования в трансформационной таблице адаптера указаны имена, по которым будут выявляться поля в ответе (HTML документе) ИПС Ирбис, и имена атрибутов в возвращаемом посреднику объекте.

Таким образом, адаптер сформирует Java-объект, соответствующий спецификации типа документа из конкретной коллекции ИПС Ирбис, и поместит в него данные. В данном случае атрибуты возвращаемого посреднику объекта имеют типы:

Authors : set{String};
Col_author : String;
Title : Text;
Pubdate : Date;
Series : String;
Keywords : set{String};

6.3. Преобразование адаптером запросов посредника в запросы ИПС Ирбис

Адаптер использует в качестве языка запросов подмножество языка SOQL. Посредник формирует запрос в виде SELECT…FROM…WHERE, используя в условии FROM имя класса-коллекции, и запрашивая требуемые поля (через атрибуты) и операции над полями (через вызовы методов) документа.

Например, запрос, полученный адаптером от посредника, в виде:

SELECT doc.title, doc.authors, doc.pubdate
FROM Informatics doc
WHERE title.CONTAINS(NEAR(artificial, intellect,5,FALSE)) OR
keywords.CONTAINS(expert systems)

будет преобразован к запросу на языке ИПС Ирбис:

(TI: (ARTIFICIAL ctx[5] INTELLECT) или (KW: EXPERT SYSTEMS))

При этом, с помощью предварительно настроенных в адаптере таблиц соответствия, по имени класса в условии FROM определяется Интернет-адрес сервера, имя коллекции на этом сервере и другая вспомогательная информация, требуемая для установления соединения, например, необходимость авторизации пользователя. Затем запрос на языке ИПС Ирбис отправляется по HTTP-протоколу на требуемый сервер.

6.4. Архитектура адаптера ИПС Ирбис

Архитектура адаптера разработана таким образом, чтобы обеспечить настраиваимость адаптера для работы с другими ресурсами при минимальной его модификации. В архитектуре адаптера выделены следующие компоненты:

1) Интерфейсный модуль взаимодействия адаптера с посредником
Функциями данного модуля являются:

- Реализация в адаптере объектов и методов JDBC интерфейса,
- Получение запроса на SOQL, проведение лексического и синтаксического разбора и построение дерева запроса,
- Передача запроса трансформационному модулю адаптера,
- Получение от трансформационного модуля адаптера данных и передача их в виде Java объектов, соответствующих типам Синтеза, в посредник.

2) Трансформационный модуль адаптера
Функциями трансформационного модуля адаптера являются:

- разбор дерева запроса SOQL и определение сервера, базы данных, и интерфейсного объекта IrbisServerConnector по имени класса в условии FROM,
- отображение имен классов, типов и атрибутов, указанных в SOQL запросе, в имена, используемые в коллекции,
- преобразование типов данных Синтеза в типы данных коллекции (и обратно),
- формирование запроса в синтаксисе ИПС Ирбис по запросу на SOQL,

- передача запроса интерфейсному объекту IrbisServerConnector и получение порции документов из коллекции в виде объекта DocStore.

3) Интерфейсный модуль взаимодействия с серверами ИПС Ирбис

Функцией данного модуля является обеспечение взаимодействия адаптера с серверами ИПС Ирбис. Каждому серверу соответствует некоторый объект класса IrbisServerConnector. Интерфейсный модуль с Web-сервером ИПС осуществляет формирование CGI-запроса к требуемому серверу ИПС (в частности, ИПС Ирбис), его кодирование посредством URLEncoder и отправку POST-запроса.

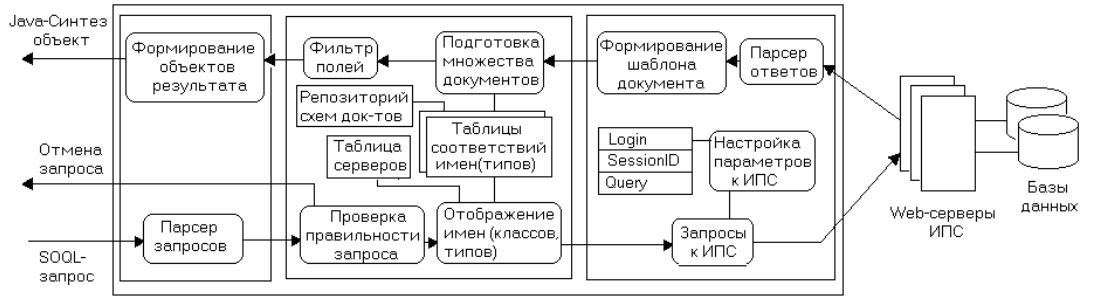

рис.1 Схема адаптера

Адаптер получает запрос на SOQL, преобразует запрос в форму для передачи через CGI интерфейс серверу Ирбис, получает документы в форме HTML, осуществляет разбор их структуры и формирует объекты в соответствии с JDBC –интерфейсом.

7. Заключение

В данной работе предложен подход для интеграции документальных информационно-поисковых систем в посреднике неоднородных коллекций данных, описана процедура регистрации новых коллекций. На примере адаптера для ИПС Ирбис показан способ построения адаптеров, предназначенных для подключения различных документальных коллекций библиографической информации к посреднику цифровой библиотеки. Рассмотрены компонентная архитектура адаптера, его функции, особенности преобразования системы типов и языка запросов посредника и ИПС Ирбис. Реализация адаптера выполнена на языке Java.

8. Литература

[1] Briukhov D.O., Kalinichenko L.A., Skvortsov N.A., “Information Sources registration at a subject mediator as a compositional development”, Preliminary version of the paper to be published in Proceedings of the Fifth East European Symposium on Advances in Databases and Information Systems (ADBIS'01)

[2] Всероссийский Научно-технический информационный центр (ВНТИЦ), http://www.vntic.org.ru

[3] Chang C.-C.K., Garcia-Molina H., Paepcke A., “Predicate Rewriting for Translating Boolean Queries in a Heterogeneous Information System”, ACM Transactions on Information Systems, Vol.17, No.1, January 1999

[4] Davis J., Lagoze C., “Dienst: A Protocol for a Distributed Digital Document Library”, Cornell University Computer Science Department, Technical Report 94-1418, June 1994

[5] Dublin Core Standard, http://purl.oclc.org/metadata/dublin_core/

[6] Kalinichenko L., Briukhov D., Kravchenko D., Zakharov V., “Infrastructure of the subject mediating level aiming at semantic interoperability of heterogeneous digital library collections”, Proceedings of the 2nd Russian Conference on Digital Libraries, 2000

[7] Kalinichenko L.A. “Compositional Specification Calculus for Information Systems Development”, Proceedings of the East-West Symposium on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia, September 1999, Springer Verlag, LNCS

[8] Kalinichenko L.A. “Method for data models integration in the common paradigm”, Proceedings of the First East-European Symposium ADBIS'97, St.Petersburg, September 1997

[9] Калиниченко Л.А. “СИНТЕЗ: Язык определения, проектирования и программирования интероперабельных сред неоднородных ресурсов”, ИПИ РАН, Москва, 1993

[10] Максимов Н., Васина Е., Голицина О., Храмцов П. и др., “Информационные ресурсы Интернет. ИПС IRBIS”, РГГУ, 1999

[11] MARC Standard for bibliographic data, http://lcweb.loc.gov/marc/

[12] Paepcke A., Cousins S., Garcia-Molina H., Winograd T. et al., “Towards Interoperability in Digital Libraries”, IEEE Computer Magazine, May, 1996

[13] The Object Data Standard ODMG 3.0, Morgan Kaufmann Publishers, 2000

[14] Z39.50 BIB-1Attribute Set Specification, http://lcweb.loc.gov/z3950/agency

[15] Эпштейн М.Я., “Использование современных информационно-поисковых систем”, “Информационные ресурсы России”, #1 (38), 1998

[16] Электронная библиотека ИНИОН (http://www.inion.ru)

Integration of Information Retrieval Systems in the DL Mediator of Heterogeneous Data Collections using IRS IRBIS as an example

Leontiev I.V., Kalinichenko L.A.,

Institute for Problems of Informatics of the Russian Academy of Sciences, Moscow

Maximov N.V.,

Russian State University for the Humanities, Moscow

A growth of internet technologies and a broad range of information sources, used in organizations, libraries and scientific centers, lead IT researchers to develop new models, methods and tools for data integration. The necessity in heterogeneous information sources integration, such as relational and object databases, bibliographic collections, Web servers and information retrieval (IR) systems, becomes clear. The approach used in our project is based on the mediator architecture, to which heterogeneous data collections are attached via wrappers. In the article, we discuss the aspects of information retrieval systems integration. The procedure of new collection's registration, and an approach for construction of bibliographic IRS wrappers are also shown using the example of IR system IRBIS. A component-based architecture of the wrapper, wrapper's functions, and mapping of the mediator's query language and type system, based on the SYNTHESIS data model, into IRBIS are described.