Лекция 7

МЕТОДЫ ИНТЕГРАЦИИ ДАННЫХ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Что такое интеграция данных

Интеграция данных в информационных системах понимается как обеспечение единого унифицированного интерфейса для доступа к некоторой совокупности, вообще говоря, неоднородных независимых источников данных. Таким образом, для пользователя информационные ресурсы всей совокупности интегрируемых источников представляются как новый единый источник. Система, обеспечивающая пользователю такие возможности, называется *системой интеграции данных.*

Система интеграции данных освобождает пользователей от необходимости знания, данные из каких источников, кроме интегрированного, они используют, каковы свойства этих источников и как осуществить доступ к ним. Интегрируемыми источниками данных могут быть традиционные системы баз данных, поддерживающие различные модели данных (реляционные, объектные, объектно-реляционные, графовые и т.п.), разнообразные унаследованные системы, репозитории, веб-сайты, файлы структурированных данных. Обеспечение доступа к данным многих источников через единый интерфейс означает фактически, что речь идет о поддержке представления совокупности данных из множества независимых источников в терминах единой модели данных. Важно, наконец, заметить, что состав множества источников может быть наперед заданным или динамически пополняемым, источники данных могут обладать неизменным или обновляемым содержанием.

Краткая история

Разработка методов интеграции информационных ресурсов - одна из наиболее актуальных проблем в области информационных систем. Особенно большое внимание она стала привлекать в последние годы. Однако проблема интеграции данных отнюдь не является новой. Первые шаги в этой области относятся еще к середине 70-х гг., когда начались разработки распределенных систем баз данных и когда во многом благодаря отчету ANSI/X3/SPARC сформировались более четкие представления о многоуровневой архитектуре систем баз данных, о моделях данных как инструменте моделирования реальности и об отображении моделей данных.

Речь при этом шла, главным образом, о поддержке глобальной схемы для совокупности локальных баз данных, функционирующих в разных узлах сети под управлением СУБД, которые поддерживают одну и ту же или, в общем случае, разные модели данных. Позднее несколько более общая форма этой задачи была связана с созданием мультибаз и федеративных баз данных, хранилищ данных, различных репозиториев информационных ресурсов, а также веб-приложений. В последние годы в широко развернувшихся во многих странах разработках электронных библиотек (Digital Libraries) проблемы интеграции неоднородных данных стали играть ключевую роль, причем возникает также задача интеграции текстовых информационных ресурсов из различных независимых источников.

Многоаспектность проблемы

Проблема интеграции данных чрезвычайно многоаспектна и многообразна. Сложность и характер используемых методов ее решения существенным образом зависят от уровня интеграции, который необходимо обеспечить, свойств отдельных источников данных и всего множества источников в целом, требуемых способов интеграции.

Системы интеграции данных могут обеспечивать интеграцию данных на физическом, логическом и семантическом уровне. Интеграция данных на физическом уровне с теоретической точки зрения является наиболее простой задачей и сводится к конверсии данных из различных источников в требуемый единый формат их физического представления. В докладе обсуждаются, главным образом, два остальных случая. Интеграция данных на логическом уровне предусматривает возможность доступа к данным, содержащимся в различных источниках, в терминах единой

глобальной схемы, которая описывает их совместное представление с учетом структурных и, возможно, поведенческих (при использовании объектных моделей) свойств данных. Семантические свойства данных при этом не учитываются. Поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области обеспечивает интеграция данных на семантическом уровне.

Источники данных могут обладать различными свойствами, существенными для выбора методов интеграции данных — они могут поддерживать представление данных в терминах той или иной модели данных, могут быть статическими или динамическими и т.п. Множество источников интегрируемых данных может быть однородным или неоднородным относительно характеристик, соответствующих используемому уровню интеграции.

Что касается способов интеграции данных, то возможны два подхода — виртуальное или актуальное (материализованное) представление интегрированных данных. При первом подходе создается механизм доступа, который при обработке пользовательского запроса порождает данные в требуемом представлении непосредственно из источников данных. Полное материализованное представление интегрированных данных в терминах единого пользовательского интерфейса при этом не поддерживается. Виртуальный подход чаще всего применяется при использовании часто обновляемых источников данных. Напротив, при втором подходе на стадии интеграции формируется полное материализованное представление интегрированных данных, отчужденное от исходных источников и сосуществующее с ними. Именно это представление данных используется для обработки пользовательских запросов. Такой подход используется, в частности, в хранилищах данных.

Неоднородность источников данных

Неоднородность источников данных проявляется в системах интеграции данных в различных аспектах. При этом, естественно, идет речь о неоднородности характеристик источников, соответствующих используемому уровню интеграции данных.

Так, при интеграции на физическом уровне в источниках данных могут использоваться различные форматы файлов. На логическом уровне интеграции может иметь место неоднородность используемых моделей данных для различных источников или различаются схемы данных, хотя используется одна и та же модель данных. Одни источники могут быть веб-сайтами, а другие — объектными базами данных и т.д.

При интеграции на семантическом уровне различным источникам данных могут соответствовать различные онтологии. Например, возможен случай, когда каждый из источников представляет информационные ресурсы, моделирующие некоторый фрагмент предметной области, которому соответствует своя понятийная система, и эти фрагменты пересекаются.

Возникающие задачи

При создании системы интеграции возникает ряд задач, состав которых зависит от требований к ней и используемого подхода. К ним, в частности, относятся:

- Разработка архитектуры системы интеграции данных.
- Создание интегрирующей модели данных, являющейся основой единого пользовательского интерфейса в системе интеграции.
- Разработка методов отображения моделей данных и построение отображений в интегрирующую модель для конкретных моделей, поддерживаемых отдельными источниками данных.
- Интеграция метаданных, используемых в системе источников данных.
- Преодоление неоднородности источников данных.
- Разработка механизмов семантической интеграции источников данных.

Основной инструментарий

К числу основных средств, используемых для обеспечения интеграции информационных ресурсов, относятся конверторы данных, интегрирующие модели данных, механизмы отображения моделей данных, объектные адаптеры (Wrappers),

посредники (Mediators), онтологические спецификации, средства интеграции схем и интеграции онтологических спецификаций, а также архитектура, обеспечивающая взаимодействие средств, используемых в конкретной системе интеграции ресурсов.

Архитектура систем интеграции

В системах интеграции данных наибольшее распространение получила архитектура с посредником. На посредника возлагается задача поддержки единого пользовательского интерфейса на основе глобального представления данных, содержащихся в источниках, а также поддержку отображения между глобальным и локальным представлениями данных. Пользовательский запрос, сформулированный в терминах единого интерфейса, декомпозируется на множество подзапросов, адресованных к нужным локальным источникам данных. На основе результатов их обработки синтезируется полный ответ на запрос.

Используются две разновидности архитектуры с посредником - Global as View и Local as View. Первая из них (Global as View) предусматривает определение глобального представления интегрированных данных в терминах заданных представлений локальных источников. Такой подход более эффективен в случае, когда множество всех используемых источников предопределено. Если система интеграции предназначена для поддержки полного материализованного представления интегрируемых данных, процессы конверсии данных из источников в их единое глобальное представление осуществляются единовременно.

При использовании второй разновидности рассматриваемой архитектуры (Local as View) предполагается, что представление для каждого из локальных источников данных определяется в терминах заданного интегрирующего глобального представления. Хотя в этом случае усложняется отображение пользовательских запросов в среду локальных источников данных, такой подход допускает динамичность состава множества источников данных. Каждый новый источник может подключаться к системе как на стадии разработки, так и на стадии функционирования.

Интегрирующие модели данных

В качестве интегрирующих (называемых также *глобальными)* моделей данных для поддержки единого пользовательского интерфейса в системах интеграции чаще всего используются обычные широко используемые модели данных, например, реляционная или объектная. В связи с расширением разработок веб-приложений в качестве интегрирующей модели данных стала широко использоваться модель, основанная на стандартах XML.

При использовании в разных источниках данных неоднородных моделей данных часто для поддержки глобального представления данных создается специальная достаточно развитая интегрирующая модель данных. Экспериментальные разработки таких моделей начали проводиться еще с середины 70-х годов и ведутся до настоящего времени. Из ранних работ можно упомянуть. Из исследовательских моделей, созданных в последние годы, заслуживают внимания модели, обеспечивающие представление как структурированных, так и слабоструктурированных данных. Мощная в функциональном отношении модель данных воплощена в языке Синтез.

В разработках интегрирующих моделей данных используется также подход, основанный на интеграции моделей данных, поддерживаемых различными источниками. Такие интегрирующие модели обеспечивают одновременно и решение двойственной задачи — поддержку множества различных представлений одних и тех же данных. Известны относящиеся еще к началу 80-х годов проекты такого рода интеграции моделей. В качестве примера можно привести попытку интеграции в единой модели данных возможностей сетевой модели данных CODASYL и реляционной модели данных.

В 90-х годах появились разработки объектно-реляционной модели. В разработках флагманских SQL-серверов баз данных было реализовано объектное расширение языка SQL, «узаконенное» впоследствии в действующем стандарте языка —SQL:1999.

К этой же категории средств интеграции данных примыкает завершающаяся в настоящее время разработка расширения языка SQL - компонента новой версии стандарта языка SQL:200n, получившего название SQL/XML. Средства SQL/XML

обеспечивают возможности представления схем баз данных SQL и реляционных данных в форме XML-документов, а также реляционное представление информационных ресурсов XML в среде баз данных SQL.

Новая технологическая платформа Веб, основанная на стандартах XML, в последние годы привлекает внимание многих специалистов как эффективный инструмент интеграции информационных ресурсов во многих практически важных случаях. Большой интерес к среде XML связан не только с возможностями XML как языка описания данных, но и в значительной степени с возможностью использования его для транспорта сообщений в среде Веб.

Конструктивный интерес к средствам интеграции информационных ресурсов Веб и реляционных баз данных проявляют и разработчики новых информационных технологий для "Всемирной паутины". Разрабатываемый стандарт языка запросов XQuery платформы XML воплощает функциональность, свойственную интегрирующей модели данных. Базовая модель данных этого языка поддерживает иерархические и реляционные структуры данных и, таким образом, обеспечивает возможности для интеграции XML-данных и данных, содержащихся в реляционных базах данных. Она позволяет вместе с тем явным образом представлять огромные информационные ресурсы «скрытого» Веб — базы данных SQL, к которым в настоящее время обеспечивается доступ в среде Веб посредством интерфейса HTML-форм.

Механизмы отображения моделей данных

Неотъемлемым функциональным элементом архитектуры системы интеграции данных является механизм отображения моделей данных . Известен целый ряд работ, посвященных методам отображения моделей данных и построению отображения конкретных моделей. В некоторых системах, обеспечивающих интеграцию внешних источников данных в среду систем баз данных, используется понятие шлюза, представляющего собой по существу механизм отображения представления данных источника в среду системы базы данных. Стандартизация такого отображения для баз данных SQL обеспечивается спецификациями SQL/MED [31]. При интеграции данных в среде, основанной на платформе CORBA, используются объектные адаптеры (Wrappers), поддерживающие IDL-интерфейс к инкапсулированным информационным ресурсам и позволяющие тем самым
«объектизировать» необъектные ресурсы, например, унаследованные системы баз данных. Благодаря этому создается интегрированная интероперабельная объектная среда неоднородных информационных ресурсов.

Средства семантической интеграции данных

Наиболее распространенный подход к семантической интеграции данных основан на использовании семантических посредников (Mediators). Средствами посредников поддерживаются унифицированные метаописания интегрируемых источников данных. Как правило, семантические посредники разрабатываются для конкретной узкой предметной области. Механизмы посредников опираются на онтологические спецификации источников. Для посредника создается интегрированная онтология используемых источников. В таких системах необходима также интегрирующая модель данных с развитыми возможностями моделирования семантики данных.

В последние годы появился ряд публикаций, посвященных решению проблемы семантической интеграции данных из множества источников, в которых для представления глобальной схемы в системе интеграции данных предлагается использовать аппарат дескриптивных логик, воплощенный в языке описания онтологий OWL. Этой теме посвящено много работ, в которых онтология предметной области используется в качестве концептуальной схемы. Достоинство такого подхода заключается не только в том, что основой пользовательского интерфейса является при этом высокоуровневая семантическая модель данных, но и возможность рассуждений в терминах онтологии, служащейконцептуальной моделью.

Интеграция метаданных

Интеграция данных в информационной системе естественным образом

предполагает и интеграцию в той или иной форме метаданных, определяющих их источники**.**

Одной из традиционных задач интеграции метаданных в системах интеграции структурированных данных является задача интеграции схем. Трудности ее решения в конкретных ситуациях могут быть связаны с наличием конфликтов, например:

- Конфликтов неоднородности (используются различные модели данных дляразличных источников)
- Конфликтов именования (в различных схемах используется различная терминология, что приводит к омонимии и синонимии в именовании)
- Семантических конфликтов (выбраны различные уровни абстракции для моделирования подобных сущностей реального мира)
- Структурных конфликтов (одни и те же сущности представляются в разныхисточниках разными структурами данных).

Другая типичная довольно сложная задача - интеграция онтологических спецификаций информационных ресурсов.

Протоколы доступа как средство интеграции данных

В качестве инструмента интеграции данных могут выступать некоторые протоколы доступа к информационным ресурсам. В качестве примера можно назвать протокол доступа к распределенным ресурсам Z39.50. Он поддерживает единое иерархическое представление распределенных информационных ресурсов в среде архитектуры клиент-сервер и предоставляет пользователю единый интерфейс для доступа к ним. Разработанный первоначально для доступа к библиографическим данным, этот протокол впоследствии был адаптирован для доступа к интегрированным распределенным ресурсам иной природы, в частности, и к базам данных. В настоящее время Z39.50 обеспечивает доступ к базам данных в терминах языка SQL.

Интеграция текстовых ресурсов

Проблема интеграции коллекций текстовых информационных ресурсов сводится, главным образом, к интеграции метаданных их источников, каталогов, классификаторов, тезаурусов, онтологий и т.д. Как уже отмечалось, эта проблема приобрела особую актуальность в связи с разработками электронных библиотек. Интеграция здесь понимается как объединение коллекций текстовых документов из разных источников в рамках единого источника. Здесь наиболее интересны методы, предусматривающие материализованную интеграцию метаданных и виртуальную интеграцию собственно контента коллекций текстовых документов. Такой подход используется, например, в системе Соционет (http://socionet.ru).

Роль **стандартов** в системах **интеграции данных**

В системах интеграции данных широко используется ряд официальных международных стандартов, а также индустриальных стандартов де-факто. Среди них стандарты баз данных ISO/IEC SQL, ISO/IEC SQL/MED, стандарт объектных данных консорциума ODMG, стандарты CORBA и UML консорциума OMG, стандарты платформы XML консорциума W3C, стандарт Дублинского ядра консорциума OCLC и многие другие. Главное назначение стандартов в таких системах состоит в определении унифицированной модели данных (метаданных), являющейся основой единого интерфейса для доступа к интегрированным данным для приложений и/или конечных пользователей. Кроме того, некоторые стандарты позволяют «погрузить» интегрированные данные в некоторую полезную инфраструктуру и пользоваться для доступа к ним ее функциональностью. Такие возможности обеспечивают, например, стандарты XML и CORBA. В интеграции информационных ресурсов электронных библиотек в последние годы активно используются стандарты Инициативы открытых архивов (Open Archives Initiative, OAI; http://www.openarchives.org/).

Литература

1. Когаловский М.Р. Перспективные технологии информационных систем. — М.: ДМК Пресс. 288 с.