УДК 004.75

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ С ЦЕЛЬЮ ИНТЕГРАЦИИ ДАННЫХ В РАМКАХ АВТОМАТИЗИРОВАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ ВУЗОВ

Бубарева О.А., Попов Ф.А., Ануфриева Н.Ю.

Бийский технологический институт (филиал) ГОУ ВПО «Алтайский государственный технический университет им. И.И. Ползунова», Бийск, e-mail: angel@bti.secna.ru

Рассмотрена проблема интеграции информации в процессе разработки и сопровождения интегрированной автоматизированной информационной системы (ИАИС) вуза, предложен подход к ее эффективному решению на основе использования семантических моделей данных, построенных с использованием онтологий. Отличительной особенностью подхода является расширение семантической модели за счет включения в нее сведений о таких предметных областях деятельности вуза, как управление качеством образования, ИТ-область, управление процессами. Использование данного подхода при построении ИАИС позволило объединить данные из разных источников с сохранением качества интегрированной информации, в т.ч. ее целостности и достоверности.

Ключевые слова: интеграция, онтологический подход, автоматизированная информационная система, качество данных

USE OF ONTOLOGIES TO INTEGRATE DATA WITHIN AN AUTOMATED INFORMATION SYSTEM UNIVERSITIES

Bubareva O.A., Popov F.A., Anufrieva N.Y.

Biysk institute of technology (branch) of the state educational institution of the higher occupational education «the Altay state technical university of I.I. Polzunova», Biisk, e-mail: angel@bti.secna.ru

The problem of information integration in the process of developing and maintaining integrated automated information system (IAIS), the University, proposed an approach for its effective solution based on the use of semantic data models built using ontologies. A distinctive feature of the approach is to extendthe semantic model by the inclusion of information on these subjectareas the University, as quality management education, IT area, the management processes. Using this approach in constructing the IAIS allowed to combine data from different sources while maintaining the quality of integrated information, including its integrity and authenticity.

Keywords: Integration, the ontological approach, an automated information system, quality of data

Проблема интеграции баз данных (БД), унификации доступа к централизованным данным с целью формирования агрегированной информации для последующего анализа особенно актуальна в настоящее время для вузов в связи с разработками в них и особенностями построения интегрированных автоматизированных информационных систем (ИАИС), обеспечивающих автоматизацию и информатизацию всех видов вузовской деятельности [3].

Данная проблема кратко описана в работе [4]. Она особенно остро стоит на этапе сопровождения ИАИС, когда необходимо консолидировать сведения из нескольких, независимо разработанных ИС, уже существующих или вновь разрабатываемых. Обусловлено это тем обстоятельством, что разработчики ИАИС по мере наращивания их функциональности вынуждены постоянно заниматься корректировкой программ и моделей данных, что значительно повышает сложность процессов разработки, сопровождения и эксплуатации систем такого рода.

При этом средства интеграции должны обеспечивать не только унифицированный интерфейс к унаследованным и новым информационным системам, входящим в состав ИАИС, но и создание информационной инфраструктуры для доступа к корпоративным ресурсам и системам, опирающейся на единые принципы взаимодействия и управления доступом к данным.

Источники данных в ИАИС можно подразделить на транзакционные (операционные) базы данных, хранилища и витрины данных. Операционные источники данных, пополняемые через системы сбора и обработки информации, как правило, не согласованы друг с другом, поэтому для использования в рамках одной ИАИС требуется их консолидация, т.е. объединение. Хранилища и витрины данных (Data Warehouse и Data marts) есть те источники, куда поступает консолидированная информация и из которых извлекаются сведения, необходимые для делового анализа и принятия управленческих решений.

В архитектуре современной ИАИС можно выделить следующие уровни:

– сбор данных и их первичная обработка;

– извлечение данных из операционных источников, их преобразование, консолидация и загрузка в хранилища и витрины, выполняемые с помощью ETL-инструментов (ETL-Extraction, Transformation, Loading), а

также технологий управления содержанием корпорации – ECM (ECM – enterprise content management). Большинство решений ECM направлено на консолидацию и управление неструктурированными данными, такими, как документы, отчеты и web-страницы;

– размещение данных в предметно-ориентированных, интегрированных, некорректируемых, зависимых от времени хранилищах;

– представление данных в витринах, предназначенных для проведения целевого делового анализа;

– собственно анализ данных, реализуемый с помощью так называемых BI-инструментов (Business Intelligence Tools);

– Web-портал, обеспечивающий, с учетом ограничений доступа, пользователей как внутри вуза, так и в любой точке мира необходимой для анализа информацией.

Упомянутая выше консолидация данных является одним из способов их интеграции, понимаемой как обеспечение единого унифицированного интерфейса для доступа к некоторой совокупности неоднородных независимых источников данных [2].

При консолидации важным является вопрос выявления в источниках данных изменений с момента последней передачи их в хранилище, что обусловлено необходимостью поддерживать данные в актуальном состоянии.

Другой способ интеграции – федерализация данных (интеграция в реальном времени), заключается в их извлечении из источников на основании внешних требований, в процессе которого над ними и осуществляются все необходимые преобразования. Если приложение генерирует запрос к информации, то процессор федерализации данных извлекает ее из соответствующих источников, интегрирует таким образом, чтобы она соответствовала единому представлению и требованиям запроса, и отправляет результаты приложению, от которого пришел запрос. Интеграция корпоративной информации (EII – Enterprise information integration) – пример технологии, поддерживающей федеративный подход к интеграции данных. Один из ключевых элементов такой технологии – метаданные, используемые процессором федерализации данных при извлечении их из источников.

Говоря другими словами, при использовании этого метода образуется единое виртуальное информационное пространство, данные в котором могут содержаться в различных источниках, однако информация об их расположении недоступна запрашивающей стороне. Федерализация поддерживает данные в актуальном состоянии и дает возможность ИС оставаться независимыми, но характеризуется трудностями при извлечении и согласовании больших массивов данных, а также сложностью сопровождения ИАИС при изменении одной из входящих в нее систем.

Недостатками метода можно считать семантическую неоднородность при интеграции данных из разных источников, вследствие чего нарушается их качество (полнота, точность).

Третий способ интеграции – распространение данных заключается в переносе информации из одного места в другое при наступлении определенных событий. Обмен данными при этом ведется оперативно, а передаются они синхронно или асинхронно. Большинство технологий синхронного распространения данных поддерживает двусторонний обмен ими между первичными и конечными источниками. Примерами технологий, поддерживающих распространение данных, являются интеграция корпоративных приложений (EAI – Enterprise application integration) и тиражирование корпоративных данных (EDR – Enterprise data replication).

Часто при интеграции информации используется так называемый гибридный подход, суть которого рассмотрена ниже на примере интеграции данных о клиентах (CDI – customer data integration), в качестве которых могут рассматриваться студенты вуза. В этом случае могут быть использованы одновременно как консолидация, так и федерализация данных. Общие данные о студентах (имя, адрес и т.д.) могут быть консолидированы в одном складе, а данные, которые относятся к определенному приложению (например, сведения об успеваемости), могут быть федерализированы. Такой гибридный подход может быть расширен также за счет распространения данных.

Интегрируемыми источниками данных могут быть традиционные системы БД, поддерживающие различные модели данных, репозитории, веб-сайты, файлы структурированных данных и др. При этом возникают проблемы, обусловленные необходимостью хранить, анализировать, обобщать, осуществлять поиск и представлять пользователю на основе удобных пользовательских интерфейсов, с учетом уровня его подготовки, в наглядном виде мультимедийные, неполные и неточные данные, обеспечивать эффективные способы обнаружения моделей существующих наборов данных (Data Mining).

В целом необходимо отметить: обеспечение доступа к данным многих независи-

мых источников через единый интерфейс означает, что речь идет об их единой модели данных.

Разрешается проблема интеграции данных в рамках ИАИС путем последовательного решения следующих задач:

• Разработка принципов построения системы интеграции данных с учетом их семантики;

• Создание глобальной модели данных, являющейся основой единого пользовательского интерфейса в ИАИС;

• Разработка методов отображения локальных моделей данных, поддерживаемых отдельными источниками, в глобальную модель;

• Интеграция понятий, используемых в системах источников данных.

При этом могут иметь место конфликты, обусловленные использованием различных моделей данных для различных источников, в частности: использование различных терминов для обозначения одних и тех же понятий; различного рода семантические конфликты; одни и те же сущности реального мира представляются в разных источниках разными структурами данных.

Все известные методы интеграции данных можно разделить на группы, образующие в совокупности шестиуровневую структуру: интеграция вручную (Manual Integration), общий интерфейс пользователя (Common User Interface), интеграция средствами приложений (Integration by Applications), интеграция средствами программного обеспечения промежуточного слоя (Integration by Middleware), унифицированный доступ к данным (Uniform Data Access), общие системы хранения (Common Data Storage).

В качестве примера общих систем хранения можно назвать операционные БД, создаваемые для хранения «атомарных» данных. Системы распределенных БД (федеративные системы), построенные с использованием аппарата метаданных – пример унифицированного доступа к данным. Программное обеспечение промежуточного слоя может включать компоненты для перемещения данных, для оценки их качества и предварительного анализа, а также для работы с метаданными. Пример общего пользовательского интерфейса – различного рода Интернет-порталы. И, наконец, интеграция вручную осуществляется непосредственно пользователем тогда, когда использование автоматизированных методов невозможно или они не дают нужного эффекта.

Из сказанного видно, что интеграция данных возможна на синтаксическом и семантическом уровнях. При этом возможности интеграции на синтаксическом уровне позволяют лишь интерпретировать множества разнообразных данных как данные из одного источника. Очевидно, что в случае гетерогенных данных, к тому же имеющих разную структуру или вообще неструктурированных, зачастую описывающих одну и ту же проблемную область с использованием различных терминов и понятий, для успешного решения задачи интеграции их семантика должна быть явным образом выражена и сохранена вместе с этими данными.

В связи с этим актуальным при интеграции данных является использование семантически ориентированных технологий, таких как онтологии и дескрипционная логика [1,5]. При этом использование онтологий не только позволяет создавать модели данных, адекватные реальному миру, но и соответствует общему направлению работ в области стандартизации World Wide Web в рамках проекта семантического Web, что позволяет рассматривать проблемы интеграции данных и интеллектуализации WWW с единой точки зрения. Разработанные для этих целей унифицированная модель данных RDF (Resource Decription Framework) и язык веб-онтологий OWL (Web Ontology Language [6]) предоставляют богатые возможности семантического описания распределенных в Интернет сведений.

В семантической модели данных, построенной с использованием онтологий, базовыми структурными элементами являются *глоссарий* и *таксономии, определяющие в совокупности* понятия и классы понятий предметных областей деятельности вуза (образовательная, научно-исследовательская, планово-финансовая и административно-хозяйственная деятельности, управление персоналом, административное управление, управление качеством образования, ИТ-область, управление процессами), отношения между этими понятиями, а также иерархию их классов. Каждый класс при этом характеризуется определенными свойствами, описывающими различные его характеристики с учетом ограничений на свойства. Механизм изменения онтологических описаний реализуется с помощью инструментов создания понятий и отношений между ними, а также редактирования и удаления экземпляров понятий.

Онтологии ИТ-области и области управления процессами при этом занимают особое место во множестве онтологий предметных областей деятельности вуза: первые предназначены для реализации идеи отображения предметных областей на область информационных технологий, вторые обусловлены тем, что все виды де-

ятельности вуза могут быть представлены в виде множества взаимосвязанных бизнес-процессов. В качестве примеров понятий области ИТ можно отметить следующие: База данных; Система управления базами данных; Информационная система; Портал; Операционная система; Вычислительная сеть; Пользователь; Проект; др. Понятия области управления процессами: Бизнес-процесс; Сценарий; Событие; Маршрут; Условие; др.

Описания всех понятий, физических и логических связей между источниками данных и потребляющими эти данные подсистемами ИАИС хранятся в репозитории метаданных. Понятия идентифицируются их именами; в рамках ИАИС имя понятия должно быть уникальным. В репозитории метаданных хранятся также ссылки на все процедуры и сервисы, обеспечивающие поддержание качественной информации.

В целом рассмотренный подход к интеграции данных с использованием онтологий в рамках ИАИС вузов обеспечивает не только консолидацию распределенной информации с сохранением ее качества, но и позволяет автоматизировать изменения в семантике понятий и в источниках данных. Данный подход был успешно применен при построении ИАИС Бийского технологического института, что позволило объединить данные из разных источников с сохранением качества интегрированной информации, в т.ч. ее целостности и достоверности.

Список литературы

1. Бездушный А.А. Математическая модель интеграции данных на основе дескриптивной логики: автореф. дис. ... канд. физ.-мат. наук. – М., 2008. – 21 с.

2. Бубарева О.А., Использование интеграции информации для анализа несопоставимых источников данных в информационно-управляющих системах / О.А. Бубарева, Ф.А. Попов // Единая образовательная информационная среда: проблемы и пути развития: труды VIII Международной научно-практической конференции-выставки. – Томск: ТГУ, 2009. – С. 136–137.

3. Бубарева, О.А. К вопросу проектирования автоматизированной системы управления учебным процессом вуза // Телематика'2010: телекоммуникации, веб-технологии, суперкомпьютинг: сборник статей участников Всероссийского конкурса научных работ студентов и аспирантов. – СПб: СПбГУ ИТМО, 2010. – С. 72–76.

4. Бубарева О.А. Решение проблемы интеграции данных при построении интегрированной автоматизированной информационной системы вуза / О.А. Бубарева, Ф.А. Попов, Н.Ю. Ануфриева // Международный журнал экспериментального образования. – 2011. – №5. – С. 90–92

5. The Description Logic Handbook: Theory, Implementation and Applications / F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P.F. Patel-Schneider // Cambridge University Press. – 2003.

6. OWL Web Ontology Language [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/owl-features/ (дата обращения: 17.03.11).

Рецензенты:

Цхай А.А., д.т.н., профессор, зав. кафедрой математики и прикладной информатики Алтайской академии экономики и права, г. Барнаул;

Лебедев А.С., д.т.н., доцент, начальник лаборатории ОАО «ФНПЦ «Алтай», г. Бийск.

Работа поступила в редакцию 07.07.2011.