ЕДИНЫЙ ИНТЕРФЕЙС ДОСТУПА К ГЕТЕРОГЕННЫМ БАЗАМ ДАННЫХ

Белодед Б. В., Глоба. Л. С., Терновой М. Ю., Штогрина Е. С.
Национальный технический университет Украины «Киевский политехнический институт», Институт телекоммуникационных систем
Украина, г. Киев, пер. Индустриальный, 2, 03056
тел.: +380-44-406-82-99, e-mail: ex3mst@gmail.com, lgloba@its.kpi.ua, maximter@mail.ru, L_Shtogrina@mail.ru

Аннотация — Предложен подход для интеграции реляционных баз данных на основе онтологии. Основополагающим моментом является маппинг реляционных схем в одну онтологию и последующая трансляция SPARQL-запросов в SQL-запросы. Результатом есть предоставление единого интерфейса доступа в виде SPARQL-точки доступа.

I. Введение

Развитие информационно-телекоммуникационных технологий влечет увеличение объемов информации, необходимой для работы корпоративных систем. Повысить эффективность работы корпоративной системы в целом можно за счет повышения эффективности работы каждого пользователя ее информационных ресурсов. Это возможно осуществить за счет увеличения скорости доступа к актуальной информации, необходимой конкретному пользователю в каждый момент времени.

Для доступа к данным существующие системы предоставляют различные средства формирования отчетов, а также графические средства построения запросов на получение данных. Чтобы воспользоваться этими средствами пользователь должен не только знать теоретические основы проектирования и использования баз данных (БД), но и детально разбираться в структурах хранения данных в своей корпоративной системе. И даже при наличии всех вышеперечисленных знаний и самостоятельном построении запросов пользователь будет тратить большую часть времени на процесс получения данных, а не на свои прямые обязанности, что значительно снизит эффективность его работы.

Вследствие этого возникает необходимость создания единого интерфейса доступа к данным для пользователей.

II. Основная часть

Как правило, информация хранится в гетерогенных источниках, в частности, в гетерогенных реляционных БД. Следовательно, требуется создание механизма обеспечивающего интеграцию нескольких гетерогенных реляционных БД на основе единой терминологии понятной пользователю. На сегодняшний день для описания и формализации предметных областей используются онтологии. В данном случае, онтология будет представлять собой иерархически структурированное множество терминов. Одним из ключевых моментов использования онтологии является концептуализация – первичная теоретическая форма, рассматриваемая независимо от словаря предметной области и конкретной ситуации. Основными компонентами онтологии являются классы (или понятия), отношения, функции, аксиомы, экземпляры [1].

Для описания онтологий используется формальный язык OWL (Ontology Web Language), принятый как стандарт консорциумом W3C (World Wide Web Consortium) [2]. Он в свою очередь фактически является надстройкой над RDF (Resource Description Framework) и RDFS (RDF Schema), и поддерживает эффективное представление онтологий в терминах классов и свойств, обеспечение простых логических проверок целостности онтологии и связывание онтологий друг с другом. Не вдаваясь в семантические нюансы OWL, можно сказать, что RDF является языком представления информации о ресурсах - метаданных. Под ресурсом понимается любая сущность – как информационная (файл), так и неинформационная (абстрактное понятие). RDF предоставляет универсальный способ разложения знаний на маленькие составные части, путем задания определенных правил касательно семантики этих элементов [3]. Идея состоит в том, чтобы одним способом можно было описать любой факт в структурированном виде, пригодном для автоматизированной обработки.

Базовой структурной единицей RDF является коллекция триплетов (троек), каждый из которых состоит из субъекта, предиката и объекта (S, P, O). Набор триплетов называется RDF-графом (рис. 1). В качестве вершин графа выступают субъекты и объекты, в качестве дуг — предикаты (или свойства). Направление дуги, соответствующей предикату в данном триплете, всегда выбирается так, чтобы дуга вела от субъекта к объекту.

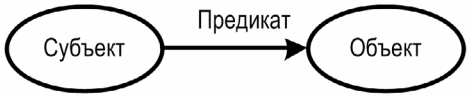

Рис. 1. RDF-триплет.
Fig. 1. RDF-triplet

В качестве языка запросов к онтологиям (RDF-хранилищам) консорциум W3C рекомендует использовать язык SPARQL. Необходимо заметить, что SPARQL-запросы синтаксически схожи с языком реляционных БД SQL. В свою очередь, реляционные БД являются наиболее часто используемыми хранилищами данных за счет наличия математической основы – реляционной алгебры и реляционного исчисления. Учитывая эти факторы, предлагается подход к интеграции гетерогенных реляционных БД на основе онтологии предметной области (рис. 2.).

Можно выделить следующие основные этапы работы информационной системы (рис. 3.):

1. Пользователь формирует запрос G в терминах предметной области.
2. Система взаимодействия P на основе запроса G формирует SPARQL запрос H к онтологии.
3. SPARQL запрос H на основе базы метаданных U преобразуется в SQL-запросы $L = \{l_k \mid k = \overline{1,n}\}$ к базам данных $БД = \{БД_k \mid k = \overline{1,n}\}$.

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine
 ISBN: 978-966-335-351-7. IEEE Catalog Number: CFP11788

4. Компоновка данных, полученных в результате выполнения запросов L для предоставления ответа $\widetilde{G}$ пользователю.

$\widetilde{G}$ - ответ, сформулированный в терминах понятных конечному пользователю, содержит данные из БД, поставленные в соответствие этим терминам.

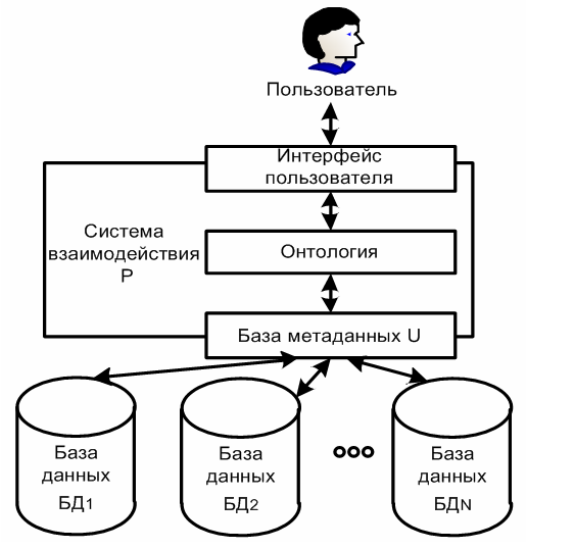

Рис. 2. Структурная схема интеграции баз данных
Fig. 2. The block diagram for the databases integration

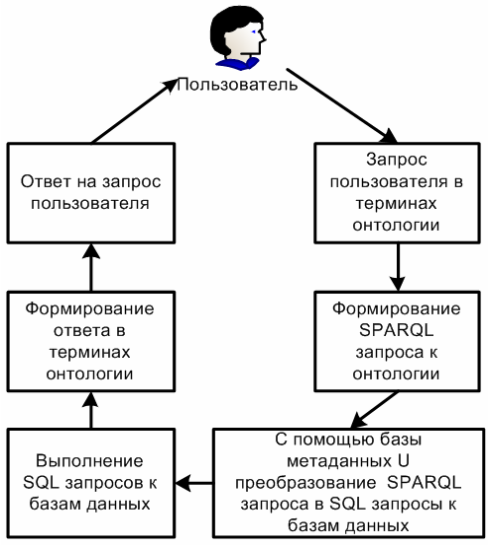

Рис. 3. Запрос на получение данных
Fig. 3. Data request (query)

Для реализации информационной системы предложенным способом необходимо решить следующие задачи:

Задача №1. Построение базы метаданных для интеграции БД и онтологии.

Задача №2. Преобразование запроса пользователя в SPARQL запрос к онтологии.

Задача №3. Преобразование SPARQL запроса в SQL запросы к БД.

Основной частью работы является интеграция реляционных БД с помощью онтологии, поэтому необходимо определить связующее звено между составными частями информационной системы – реляционными БД и онтологией, т.е. решить задачу №1.

Постановка задачи №1

Дано:

1. $БД = \{БД_k \mid k = \overline{1,n}\}$ - множество баз данных, где $БД_k$ – реляционная база данных;
2. $O = \{X, R\}$ - онтология предметной области, где $X = \{X_i \mid i = \overline{1,m}\}$ – множество понятий, обозначающих объекты, процессы или явления предметной области; $R = \{R_j \mid j = \overline{1,s}\}$ - множество отношений между понятиями предметной области (предикаты);

Найти:

U - база метаданных, которая содержит данные для интеграции БД и онтологии.

Решением для этой задачи является построенная база метаданных, в которой будет описана связь онтологии и реляционной БД. Процесс создания такой базы метаданных заключается в маппинге реляционных схем БД в одну онтологию. Реляционным сущностям БД, к которым необходимо предоставить доступ пользователю, ставятся в соответствие объекты онтологии.

В простейшем случае, любая реляционная схема может быть представлена в виде RDF-схемы путем преобразования всех первичных и внешних ключей в IRI (Internationalized Resource Identifier), назначение IRI-предиката каждому атрибуту, и rdf:type предиката для каждой строки, привязав ее к IRI RDF-класса, который соответствует таблице. Таким образом, каждое значение атрибута таблицы (одна запись), который не является частью первичного или внешнего ключа, является объектом триплета. IRI первичного ключа является субъектом. А сам атрибут (столбец) – предикатом RDF-триплета (рис. 4.).

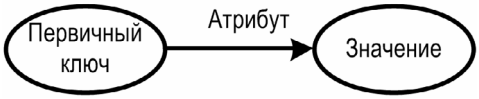

Рис. 4. Маппинг триплета
Fig. 4. Triple mapping

Результатом такого преобразование всех таблиц БД будет RDF-граф. Каждая отдельная реляционная схема будет иметь свой граф. Для создания обобщенной онтологии эти графы необходимо соединить и представить в виде одного графа (не обязательно связного). Если между реляционными объектами разных БД прослеживается логическая связь, то в терминах RDF необходимо произвести выделение предиката для определения связи между субъектом и объектом (рис. 5).

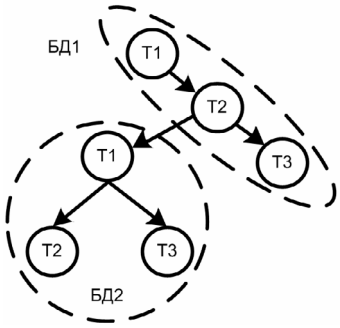

Рис. 5. Составление объединенного графа.
Fig. 5. Graph composing

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine

Сам по себе, маппинг реляционных БД в одну онтологию не является однозначным, то есть, он различен в зависимости от набора таблиц, к которым необходимо предоставить доступ, и от выработанной концепции по виденью всей системы. Один и тот же набор таблиц можно представить в виде разных наборов объектных сущностей в онтологии. На этом этапе разработчик имеет возможность гибко настраивать представление путем объединения схожих объектов в одно целое, и выделении обнаруженных связей между сущностями.

Поскольку конечной целью является предоставление пользователю интерфейса, позволяющего оперировать терминами предметной области при составлении запроса на получение данных, возникает задача описания интерфейса «пользователь-онтология». Сама по себе, эта задача включает разработку интуитивно понятного пользовательского интерфейса, который не будет требовать от пользователя ни знаний по информационным технологиям, ни составления словаря, в котором объектам онтологии будут поставлены в соответствие термины предметной области. Вид пользовательского интерфейса зависит от предметной области и требует индивидуального подхода в каждом конкретном случае. Информационная система должна преобразовывать запрос составленный пользователем, с помощью такого интерфейса, в запрос к онтологии, т.е. должна решать указанную ранее задачу №2 «Преобразование запроса пользователя в SPARQL запрос к онтологии».

Постановка задачи №2:

Дано:

1. Запрос пользователя G в терминах предметной области на получение данных;
2. Словарь предметной области, описывающий элементы онтологии;
3. $O = \{X, R\}$ - онтология предметной области

Найти:

SPARQL-запрос на получение данных.

При использовании интерфейса взаимодействия, на основе запроса, поставленного пользователем, формируется SPARQL-запрос в соответсвии с требуемой функциональностью.

Наличие такого интерфейса не исключает возможности использования сторонних SPARQL-приложений, которые будут взаимодействовать с информационной системой через SPARQL-точку доступа. Это позволяет предоставить возможность поиска данных в реляционных БД согласно концепции Semantic Web. Необходимо заметить, что зачастую такие системы оперируют распространенными онтологиями, такими как FOAF (Friend Of A Friend) [4], SIOC (Semantically-Interlinked Online Communities) [5], и т.д.. Это означает, что в процессе маппинга для общепринятых понятий предметной области необходимо использовать именно эти онтологии.

После того как построен запрос к онтологии необходимо, с помощью базы метаданных для интеграции БД и онтологии, преобразовать этот запрос в запросы к БД. то есть решить задачу № 3 «Преобразование SPARQL запроса в SQL запросы к БД».

Постановка задачи №3:

Дано:

1. Запрос пользователя в терминах онтологии на получение данных;
2. $БД = \left\{ БД_k \mid k = \overline{1,n} \right\}$ - множество баз данных;
3. $O = \{X, R\}$ - онтология предметной области
4. U - база метаданных;

Найти:

SQL-запрос на получение данных.

Как указано ранее, SPARQL-запросы синтаксически схожи с SQL-запросами, однако принцип описания предикатов у них отличается. Само по себе, выполнение SPARQL-запроса на выборку (SELECT) состоит в поиске в графе совпадений по указанному условию, то есть поиск триплетов по предоставленной «маске». Если совпадение найдено, то выводится соответствующий результат.

Данная задача решается путем использования математического аппарата и подходов, предложенных в [6-8]. Основным моментом есть то, что сам процес трансляции является функцией от произведенного ранее маппинга.

III. Выводы

В работе предложен подход к созданию единого интерфейса доступа к гетерогенным базам данных, основанный на онтологии предметной области. Данный подход предоставляет единый унифицированный интерфейс для получения данных. При этом конечному пользователю предоставляется возможность формулировать свой запрос в терминах предметной области.

IV. Литература

[1] *Добров Б. В.* Онтологии и тезаурусы: модели, инструменты, приложения / Б. В. Добров, В. В. Иванов, Н. В. Лукашевич, В. Д. Соловьев. [Электронный ресурс]. – Электрон. текстовые данные. – Режим доступа: http://www.intuit.ru/department/expert/ontoth/1/

[2] *OWL* Web Ontology Language Overview / Deborah L. McGuinness (ed.), Frank van Harmelen (ed.) // W3C Recommendation 10 February 2004. [Электронный ресурс]. – Электрон. текстовые данные. – Режим доступа: http://www.w3.org/TR/owl-features/

[3] *Resource* Description Framework (RDF): Concepts and Abstract Syntax / Graham Klyne (ed), Jeremy J. Carroll (ed) // W3C Recommendation 10 February 2004. [Электронный ресурс]. – Электрон. текстовые данные. – Режим доступа: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[4] *The Friend* of a Friend (FOAF) project. [Электронный ресурс]. – Электрон. текстовые данные. – Режим доступа: http://www.foaf-project.org/

[5] *The SIOC* (Semantically-Interlinked Online Communities). [Электронный ресурс]. – Электрон. текстовые данные. – Режим доступа: http://sioc-project.org/

[6] *Chebotko A.* Semantics Preserving SPARQL-to-SQL Translations / A. Chebotko, S. Lu, F. Fotouhi // Technical Report, TR-DB-112007-CLF, 2007.

[7] *Cyganiak R.* A relational algebra for SPARQL / R. Cyganiak // Technical Report, HP Laboratories Bristol, 2005

[8] *Elliott B.* A complete translation from SPARQL into efficient SQL. / B. Elliott, E. Cheng, C. Thomas-Ogbuji, Z. M. Ozsoyoglu // Proceedings of the 2009 International Database Engineering Applications Symposium on IDEAS 09, 31. ACM Press.

UNIFIED ACCESS INTERFACE TO HETEROGENEOUS DATABASES

Bilodid B. V., Globa L. S.,
Ternovoy M. Y., Shtogrina O. S.
National Technical University of Ukraine
"Kyiv Polytechnic University",
2, Industrialniy Lane, Kyiv, 03056, Ukraine
Ph.: 380-44-2417699, e-mail: ex3mst@gmail.com,
lgloba@its.kpi.ua, maximter@mail.ru,
L_Shtogrina@mail.ru

Abstract —.The ontology-based approach for relational databases integration is proposed. Mapping of the relational schemas into one ontology and further translation of SPARQL-queries into SQL-queries are fundamental point of approach. The result provides a unified access interface in the form of SPARQL-end point.

I. Introduction

Development of information and telecommunication technologies leads to the increasing of necessary information amount in the corporate systems. The efficiency increasing is possible by speeding-up the access to the relevant information which user needs at the moment.

For data access a user must know not only the theoretical foundations of design and use of databases (DB), but also must understand in detail the data storage structures in the corporate system. So, it is necessary to create a unified data access interface for the end users.

II. Main Part

Typically, the information stored in heterogeneous sources and particularly in heterogeneous relational databases. Hence, there is a need to design a mechanism for integration multiple heterogeneous DB. Nowadays ontologies are used for the description and formalization of the subject areas. In this case, ontology is a hierarchically structured set of terms. The main components of ontology are classes (or concepts), relations, functions, axioms, instances [1].

OWL (Ontology Web Language) is used for an ontology description/ OWL is based on RDF (Resource Description Framework) and RDFS (RDF Schema) [2], and provides an effective representation of ontologies in terms of classes and properties, provides the simple logic verifications of ontology. RDF provides the unified way for knowledge decomposition into the small components by the specific rules setting about the semantics of these components [3]. The idea is to provide one way for any fact description in a structured form which is suitable for automated processing.

Basic building block of RDF is a collection of triplets, each of which consists of subject, predicate and object (S, P, O). Set of triplets is called RDF-graph (Fig. 1). The graph nodes are the subjects and the objects. The arcs are predicates (or properties). The direction of the arc, which corresponds to predicate in the triplet, is always chosen so that the arc points from subject to object.

The language SPARQL is used as a query language for ontologies (RDF-repositories). The SPARQL-query is syntactically similar to the relational databases query language - SQL.

Taking into account these factors, the ontology-based approach for the heterogeneous relational databases integration is proposed (Fig. 2.).

There are following basic stages of information system operation (Fig. 3.):

1. A user creates a query G in terms of the subject area.
2. Interaction system P forms SPARQL query H to the ontology.
3. Using metadata base U interaction system P transforms SPARQL query H into the SQL queries $L = \{l_k \mid k = \overline{1,n}\}$ to the databases $DB = \{DB_k \mid k = \overline{1,n}\}$
4. The obtained data are aggregated and are given to the $\widetilde{G}$ to the user.

$\widetilde{G}$ - answer, formulated in terms understandable to the end user.

To implement an information system in this way, we need to solve the following problems:

Problem #1. To design the metadata base for the integration of databases and ontologies.

Problem #2. To convert a user query into SPARQL query to the ontology.

Problem #3. To convert SPARQL query into SQL query to database.

Problem definition #1

Given:

1. $DB = \{DB_k \mid k = \overline{1,n}\}$ - set of databases, where DB_k – relational database;
2. $O = \{X, R\}$ - the ontology of the knowledge domain, where $X = \{X_i \mid i = \overline{1,m}\}$ – the set of concepts that denote objects, processes or phenomena of the knowledge domain; $R = \{R_j \mid j = \overline{1,s}\}$ - the set of relations between domain concepts (predicates);

Find:

U - meta-database which contains data for the integration of databases and ontology.

Problem definition #2

Given:

1. User query G in terms of knowledge domain for data acquisition;
2. Knowledge domain vocabulary, which consists of the description of the ontology elements;
3. $O = \{X, R\}$ - knowledge domain ontology

Find:

SPARQL-query H for data acquisition.

Problem definition #3

Given:

1. User SPARQL-query H in terms of ontology for data acquisition;
2. $DB = \{DB_k \mid k = \overline{1,n}\}$ - set of the databases;
3. $O = \{X, R\}$ - the ontology of the knowledge domain
4. U - meta-database;

Find:

SQL-query for data acquisition.

III. Conclusion

In this paper, the unified access interface to heterogeneous databases, which based on the using of knowledge domain ontology, is proposed. This approach provides a unified interface for data acquisition. The end user can formulate a request in terms of the knowledge domain.

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine
 ISBN: 978-966-335-351-7. IEEE Catalog Number: CFP11788