ОНТОЛОГИИ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ. ЛИНГВИСТИЧЕСКИЙ АСПЕКТ[1]

A RELATIONAL DATABASE ONTOLOGY. THE LINGUISTIC ASPECT

Биряльцев Е.В. *(ltbd@ksu.ru),* ***Гусенков А.М.*** *(ltbd@ksu.ru)*
Казанский государственный университет

Рассматривается задача представления структуры реляционных баз данных в формализме онтологий для решения задачи поиска реляционных баз данных по запросам, задаваемым пользователями. Предлагается базовая онтология реляционных баз данных, включающая концепты, отношения и функции интерпретации. Показывается, что для выполнения запросов в реальных базах данных необходимо расширение онтологии лексико-семантическими отношениями между определениями столбцов таблиц баз данных. Рассматриваются виды лексико-семантических отношений, существующие в реальных базах данных.

Значительная часть научной, коммерческой, культурной и иной информации в настоящее время представлена в табличной форме и, при хранении в электронном виде, находится в реляционных базах данных. Как для общегражданских, так и для корпоративных целей достаточно актуально иметь возможность производить в них произвольный поиск по запросам, задаваемым непосредственно пользователями.

Структура хранимой в реляционных базах данных информации существенно отличается от сплошного текста. Если для поиска в сплошных текстах в простейшем случае достаточно линейного просмотра текста в поисках ключевых слов, то организовать поиск в реляционной базе данных подобным образом невозможно. Доступ к реляционным базам предполагает знание пользователем структуры базы данных и производится либо через заранее предопределенные формы (QBE), либо посредством языка запросов SQL. Необходимо генерировать синтаксически правильные и семантически осмысленные запросы (как правило, на соответствующем базе данных диалекте SQL) и анализировать текст ответа. В настоящее время подобная генерация осуществляется программистом, что определяется следующими причинами:

- Разбиение данных по таблицам может быть осуществлено очень большим количеством способов, конкретный вариант разбиения конечному пользователю неизвестен, его изучение и составление соответствующего структуре запроса требует значительного времени и квалификации, которые могут у конечного пользователя отсутствовать.
- Названия таблиц, столбцов и словарных элементов могут отличаться от употребляемых пользователем, даже в пределах одного языка. Перебор пользователем всех возможных формулировок запроса практически неосуществим.

Вместе с тем, пользователь, являющийся специалистом в предметной области, в состоянии сформулировать семантически правильный запрос с точностью до наименований элементов требуемой информации и их распределения по объектам базы данных. Такие запросы могут быть сгенерированы также программными средствами в рамках систем электронного бизнеса.

Таким образом, возникает задача разработки методов автоматической трансформации семантически правильного запроса, составленного с терминах предметной области в синтаксически и терминологически корректный запрос к конкретной базе данных, предположительно содержащей требуемые сведения, его выполнения и обратной трансформации запроса в термины, понятные пользователю или системе, выдавшей запрос.

Проблема состоит из 2-х подзадач:

1. Согласования терминологических базисов исходного запроса и терминов, в которых описана структура реляционной базы данных.

2. Генерация исполняемого запроса к базе.

Для решения обеих подзадач необходимо иметь универсальный механизм хранения структуры реляционной базы данных и терминов, в которых она описана, и универсальных механизмов преобразования пользовательского запроса в исполняемые SQL-предложения.

Для описания структуры реляционных баз данных существует достаточное количество формализмов – DDL SQL, ER-диаграммы, UML и другие диаграммные техники. Их мощности вполне достаточно для решения

[1] Работа выполнена при поддержке гранта РФФИ 06-07-89219-а.

задач проектирования баз данных и поддержки доступа к базам данных, рассматриваемых изолированно, каждой в своей собственной системе атрибутов. При решении задач извлечения информации из реляционных баз данных сторонним пользователем или разработчиком, возникают проблемы [1], связанные с особенностями наименований артефактов баз данных, в первую очередь таблиц и столбцов. Эти особенности требуют рассмотрения наименований артефактов не просто как атомарных имен, а как самостоятельных объектов обладающих, возможно, собственной структурой и имеющих между собой семантически нагруженные связи. Такие связи невыразимы в традиционных формализмах представления структуры реляционных баз данных, таким образом, мы не можем не только решить поставленную задачу, но и сформулировать ее. Подходящим формализмом для представления подобной сложноструктурированной информации являются онтологии.

Рассмотрим подзадачу генерации запроса к базе данных на основе представления ее структуры в виде онтологии. Известны подходы к описанию структуры конкретных реляционных баз данных с помощью онтологий [2] для конкретных предметных областей.

Применительно к решаемой задаче можно предложить следующее универсальное представление структуры реляционных БД в формализме онтологий. Основой являются универсальные (не зависящие от конкретной базы данных) концепты ТАБЛИЦА, СТОЛБЕЦ, КЛЮЧ, ДОМЕН, соответствующие основным объектам баз данных, и универсальные отношения между ними:

ТАБЛИЦА *содержит* СТОЛБЕЦ
ТАБЛИЦА *имеет первичный* КЛЮЧ
ТАБЛИЦА *имеет внешний* КЛЮЧ
КЛЮЧ *содержит* СТОЛБЕЦ
СТОЛБЕЦ *имеет тип* ДОМЕН

Объекты (таблицы, столбцы, ключи и домены) конкретной база данных во втором случае представляется как экземпляры универсальных концептов соответствующего типа. Любая информация в реляционной базе данных имеет вид набора таблиц, возможно связанных между собой общими ключами.

Задачу интеграции можно определить следующим образом – задано отношение из атрибутов базы данных, отличное от существующих. Требуется определить, можно ли, используя операции объединения по ключу и проекции, получить требуемое отношение из существующих. Задача сводится к известному классу задач о блуждании по ориентированному графу, где вершинами являются экземпляры концепта ТАБЛИЦА, а дугами – наличия общего ключа ориентированных посредством отношения “*имеет первичный*“ и “*имеет вторичный*“. Подходы к решению данного класса задач хорошо известны [3] и представляю лишь вычислительную сложность при большой размерности задачи.

Вернемся к задаче согласование терминологического базиса. Дальнейшее рассмотрение будем вести на примере реальных баз данных из области нефтяной промышленности, исследованных в рамках работ по гранту РФФИ 06-07-89219-а. Типичное описание структуры таблиц рассматриваемых баз данных приведено в Таблице 1. Очевидно, что идентификаторы столбцов реальных баз данных неинформативны для задачи идентификации их семантики. Более информативными являются определения столбцов. В определениях обычно используются лексика литературного языка и профессиональные термины из предметной области, к которой относятся базы данных.

Идентификатор столбца	Описание столбца
NC	Номер скважины
GOD	Код пласта
MES	Год
PL	Месяц
SPEX	Способ эксплуатации
DN	Добыча воды
DW	Добыча нефти
DG	Добыча газа
KDEX	Часов работы
PLB	Плотность попутно добытой воды

Таблица 1. Типичное описание таблицы в реляционной базе данных

Вместе с тем, анализ показывает, что употребление специалистами этих терминов при описании структуры баз данных и формулировании запросов к ней далеко не однозначно. Рассмотрим эту неоднозначность на при-

мере (Таблица 2) сопоставления описания структуры таблицы базы данных и построенной на ее основе отчета с заголовками столбцов.

Анализ таблицы 2 демонстрирует многочисленные лексико-семантические отношения между определениями столбцов. Необходимо отметить, что отношение синонимии, наиболее простое для анализа и представления, встречается достаточно редко. В приведенном примере синонимами, в профессиональном контексте, является пара терминов «Способ эксплуатации» и «Метод разработки».

Другие соответствия демонстрируют более сложные отношения.. Чтобы установить соответствие столбцов «Год» и «Месяц» в таблице и столбца «Период эксплуатации» в отчете необходимо учитывать, что «Год» и «Месяц» являются гипонимами лексемы «Период», а также, что в описании столбцов «Год» и «Месяц» подразумевается отнесение их к периоду добычи нефти, а не к другому событию или интервалу, например к вводу скважины в эксплуатацию.

Столбцы таблицы	**Лексико-семантическое отношение**	**Столбцы отчета**
Номер скважины	Меронимия	Объект разработки
Код пласта		
Год	Гипонимия	Период эксплуатации
Месяц		
Способ эксплуатации	Синонимия	Метод разработки
Добыча воды	Конверсив, Гипонимия	Тип флюида
Добыча нефти		Отдача флюида
Добыча газа		
Часов работы	Антонимия	Процент простоя скважины

Таблица 2. Пример сопоставления таблицы базы данных и отчета на ее основе

Для соотнесения столбцов «Добыча воды», «Добыча нефти», «Добыча газа» таблицы и столбцов «Тип флюида» и «Отдача флюида» отчета необходимо учитывать, что «вода», «нефть» и «газ» являются гипонимами термина «Флюид» в данном контексте, а «добыча» и «отдача» являются, также в профессиональном контексте, конверсивами, выражающими точку зрения на процесс: «Залежь отдает нефть скважине» «Скважина добывает нефть из залежи».

Соответствие столбцов «Часов работы» и «Процент простоя скважины» основано на антонимии терминов «Работа» и «Простой». Зная время простоя скважины и месяц, к которому он относится, можно установить время работы. Дополнительную сложность придает тот факт, что количественное выражение работы и простоя скважины выражаются различными единицами – в абсолютном исчислении (часы) и в относительном (проценты).

Соответствие между парой столбцов «Номер скважины» и «Код пласта» в таблице и столбцом «Объект разработки» отчета отражает тот факт, что объектом разработки, в профессиональном контексте, - это добыча флюида из конкретного пласта на конкретной скважине. Объект разработки, таким образом, включает, как составные части, пласт и скважину. В этом случае между терминами присутствует отношение меронимии.

Достаточно часто, в исследуемых базах данных, между определениями сопоставляемых столбцов наблюдаются отношения метонимии типа «Объект»-«Роль», «Процесс» -«Результат» и некоторые другие.

Таким образом, для решения поставленной задачи, онтологию, описывающую структуру реляционной базы данных, необходимо дополнить лексической онтологией, описывающей термины предметной области и лексико-семантические отношения между ними. Для реального применения лексической онтологии при поиске сопоставимых столбцов в интегрируемых базах данных необходимо ее заполнение для конкретных предметных областей. Насыщенные профессионализмами тексты определений столбцов ставят под сомнение применимость общеупотребительных словарных источников. Достаточно полным источником для ряда отношений (гипонимии, меронимии, ролевой метонимии), может являться общепризнанная логическая модель предметной области. В большинстве предметных областей такая модель отсутствует, однако в области нефтяной и газовой промышленности существует и развивается модель POSC Epicenter [4], содержащая более тысячи сущностей и несколько тысяч атрибутов. Альтернативным источником являются сами исследуемые базы данных. В приведенном примере лексемы «вода», «нефть» и «газ», с одной стороны, являются элементами наименований столбцов таблицы, с другой стороны, являются значениями столбца «Тип флюида». Сопоставление подобных фактов позволяет построить гипонимические отношения исходя из анализа самих интегрируемых баз данных.

Лексико-семантическая онтология дополняется правилами интерпретации, определяющими возможность и порядок применения подстановок для семантически эквивалентной трансформации наименований столбцов.

Рассмотрим наиболее часто встречающиеся лексико-семантические отношения и связанные с ними правила подстановки.

Конверсивы, как правило, образуют взаимосвязанные пары, что позволяет рассматривать их как наиболее простое, в рассматриваемом аспекте, лексико-семантическое отношение. Правило для конверсивов сводится к простой замене лексем, образующих конверсивную пару.

При подстановке антонимов необходимо учитывать, образуют антонимы взаимоисключающую пару или содержат спектр значений, отражающий интенсивность проявления некоторого фактора. Для первого случая мы можем делать подстановку аналогично конверсивам, второй случай несколько сложнее и зависит от контекста. Вопрос антонимии здесь фактически сводится к таксономии по одной или нескольким шкалам и адекватности использования наименования одного из таксонов вместо другого.

Подстановка синонимов также зачастую осложняется контекстными зависимостями. В одном профессиональном контексте две лексические единицы могут быть синонимами, в другом – нет.

При использовании отношений гипонимии и меронимии для лексических подстановок необходимо учитывать, что перемещение вверх по иерархии гипонимов и меронимов не имеет смысла, так при неограниченном повышении уровня абстракции или размера объекта сопоставятся любые лексические единицы. Возможность использования лексической единицы А вместо лексической единицы В определяется входимостью А в иерархию гипонимов/меронимов В либо наоборот (является ли А частью или частным случаем В или наоборот). Если оба вхождения отсутствуют, замена принципиально невозможна, если одно из вхождений присутствует, адекватность замены также определяется контекстом.

Метонимия, во всех своих многообразных проявлениях, является наиболее сложным для интерпретации случаем. Если синонимы, антонимы, конверсивы, гипонимы и меронимы принадлежат, как правило, к одной тематической группе, то такие распространенные метонимические отношения как объект-роль, процесс-результат, цель-средство и т.д. могут вывести за пределы тематической группы, а при неоднократном применении покрыть весь словарь. Вместе с тем, анализ реальных баз данных показывает, что метонимия, особенно в виде отношения объект-роль, является распространенным явлением в наименованиях артефактов баз данных.

Рассмотренные аспекты использования лексико-семантических отношений не позволяют предложить полностью автоматическую процедуру сопоставления наименований артефактов из различных баз данных, вместе с тем они дают мощный интерактивный инструмент для аналитика, занимающегося разработкой интеграционных процедур. При интерактивном сопоставлении наименований артефактов аналитик, в случае отсутствия полного аналога наименования включает поиск по лексико-семантическим отношениям, автоматически выполняемым системой. Аналитик, исходя из контекста, выбирает наиболее адекватный вариант соответствия. При работе с большими базами данных, содержащих тысячи артефактов, полностью устраняется рутинная компонента и достигается многократное повышение производительности труда.

Использование инструментария, разработанного на основе излагаемых принципов, применялось для разработки таблиц соответствия словарных элементов при интеграции производственной и финансовой подсистем ОАО «Татнефть». Общий объем словарных элементов составил около 50 тысяч значений. Таблицы соответствия были построены одним аналитиком за неделю работы, что дало подтвержденный экономический эффект от базового варианта (ручное сопоставление) около 270 тысяч рублей.

Список литературы

1. Биряльцев Е.В., Гусенков А.М., Галимов М.Р. Особенности лексико-семантической структуры наименований артефактов реляционных баз данных. // Труды казанской школы по компьютерной и когнитивной лингвистике TEL-2005, Казань: Изд-во Казанского государственного университета, 2006, Вып.9. С.4-12,
2. Жучков А.В. и др. Новые технологии для понятийных сетей, создаваемых в рамках МНТП «Вакцины нового поколения и диагностические системы будущего». // Электронные библиотеки. 2003 т. 6, вып. 6. < http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part6/ZATGS>
3. Джеймс.А Андерсен.. Дискретная математика и комбинаторика: Пер. с англ. // М.: Издательский дом «Вильямс», 2003
4. General E&P Standards. < http://www.energistics.org/posc/General_Stds.asp?SnID=381164644>