

УДК681.3.07

Математические основы задачи поиска отображений между элементами схем данных

*Белошицкий Д.А., аспирант
кафедра «Компьютерные системы и сети»,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана*

*Научный руководитель: Брешинов А.В., д.т.н., профессор
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
v.suzev@bmstu.ru*

Введение

В общем случае задача интеграции данных сводится к нахождению отображений семантически эквивалентных объектов схем-источников. Поиску и формированию отображений предшествует этап проведения преобразований схем данных. Совокупность подобных преобразований в большей степени определена и описана эмпирически, тогда как для построения полноценной теории необходимо строгое формальное описание.

Для построения формальных преобразований используется аппарат математической логики и аппарат дескрипционных логик. В работе рассмотрены некоторые аспекты применения аппарата математических логик – модели *LAV* и *GAV*. Также приведена классификация подходов к интеграции данных и определена общая архитектура системы интеграции.

Классификация подходов к интеграции данных

На схеме ниже показана классификация подходов для интеграции данных.



Рис. 1. Классификация подходов к интеграции данных

Подходы, основанные на схемах, используют метаинформацию о структурах данных, а методы, основанные на содержимом – анализируют содержимое структур данных. На уровне элементов анализируются атомарные элементы схем, а на уровне структур – их комбинации. Лингвистический подход подразумевает анализ названий атрибутов, их описаний, частоты встречаемости слов, поиск ключевых терминов. При анализе ограничений целостности анализируются ключевые атрибуты, типы данных.

Важным при этом является накопление знаний о предыдущих попытках интеграции и их повторное использование.

Общая архитектура системы интеграции данных изображена на рисунке ниже.



Рис. 2. Архитектура системы интеграции

Средство импорта/экспорта схем данных позволяет извлекать описания структур таблиц баз данных-источников. Средство преобразования во внутренний формат

позволяет преобразовывать метаописания схем в формат данных системы. операция отображения схем данных реализует основную задачу системы – поиск конфликтов в схемах данных и их разрешение. В результате получается консолидированная схема данных. База знаний используется для повышения качества интеграции с каждой новой попыткой.

Теоретические основы интеграции данных

Формализация системы интеграции данных.

Из приведенного выше следует, что основными элементами системы интеграции являются следующие: глобальная схема данных, источники и отображением. Поэтому формализованное определение системы интеграции следующее:

$$I = \langle G, S, M \rangle;$$

I – система интеграции;

G – глобальная схема, описываемая языком L_G над алфавитом A_G ;

S – множество схем источников, описываемых на языках L_{S_i} над алфавитами A_{S_i} ;

M – отображение между G и S , которое представлено набором предикатов:

$$\begin{aligned} q_S &\rightarrow q_G, \\ q_G &\rightarrow q_S \end{aligned}$$

где q_S и q_G – это два запроса над схемами S и G соответственно, описываемых языками $L_{M,S}$ и $L_{M,G}$ над алфавитами A_S и A_G соответственно.

Далее назовем D базу данных-источник, которая соответствует одной из схем s_D множества S и удовлетворяет всем ограничениям этой схемы. В этом случае глобальная база данных B для I соответствует локальной базе данных-источнику D , если выполняются два условия:

1. B удовлетворяет всем ограничениям G ;
2. B находится с D в отношении M .

Для данной базы данных-источника D для I , ответ $q^{I,D}$ на запрос q в I представляет собой набор кортежей t , таких, что $t \in q^B$ для каждой глобальной базы данных B для I соответствующей D . В этом случае $q^{I,D}$ называют *множеством четких ответов* на запрос q для I в базе данных D . Существует также задача нахождения множества возможных ответов на запрос q^B для базы данных B .

Существуют два базовых подхода для отображения схем данных: *local-as-view (LAV)* и *global-as-view (GAV)*. Обе этих модели основаны на логике предикатов первого порядка.

Local as view (LAV)

В системе интеграции данных $I = \langle G, S, M \rangle$, основанной на подходе LAV, отображение M ставит в соответствие каждому элементу s из S запрос q_G над множеством G . Таким образом, LAV-отображение – есть набор предикатов первого порядка для каждого элемента s из $S : s \rightarrow q_G$.

Смысл LAV заключается в том, что для каждого источника s данных формируется представление в терминах глобальной схемы G . Для того, чтобы более точно характеризовать каждый источник данных, LAV-отображение расширяется за счет того, что к каждому источнику добавляется спецификация, называемая $as(s)$. Она показывает, насколько точным является отображение $s \rightarrow q_G$. Далее возможны три варианта значений $as(s)$:

Частичный ($as(s)=sound$):

Источник s – частичный, если $s^D \subseteq q_G^D$. Таким образом, если кортеж данных находится в s^D , то есть локальной схеме, то он однозначно находится и в q_G^D , то есть в глобальной схеме. на языке предикатов первого порядка данное утверждение имеет вид:
 $\forall x \ s(x) \rightarrow q_G(x)$

Полный ($as(s)=complete$): $\forall x \ q_G(x) \rightarrow s(x)$

Точный ($as(s)=exact$): $\forall x \ s(x) \leftrightarrow q_G(x)$

Global as view (GAV)

В системе интеграции данных $I = \langle G, S, M \rangle$, основанной на подходе LAV, отображение M ставит в соответствие каждому элементу g из G запрос q_s над множеством S . Таким образом, GAV-отображение – есть набор предикатов первого порядка для каждого элемента g из $G : g \rightarrow q_s$.

В этом случае значения $as(s)$ определены следующим образом:

Частичный ($as(s)=sound$): $\forall x \ q_s(x) \rightarrow g(x)$

Полный ($as(s)=complete$): $\forall x \ g(x) \rightarrow q_s(x)$

Смысл GAV заключается в том, что каждый элемент глобальной схемы находит свое отражение в соответствующем представлении над источником.

Обработка запросов в GAV.

Рассмотрим пример обработки запросов в GAV. Пусть задана система интеграции данных $I = \langle G, S, M \rangle$, в которой G представлено следующим набором отношений:

employee(Ecode, Ename, Ecity)

company(Ccode, Cname)

employed(Ecode, Ccode)

и ограничений целостности:

key(employee) = {Ecode}

key(company) = {Ccode}

employed[Ecode] $\subseteq$ employee[Ecode]

employed[Ccode] $\subseteq$ company[Ccode],

отображение M определено следующим образом:

employee $\rightarrow \{x, y, z \mid s_1(x, y, z, w)\}$

company $\rightarrow \{x, y \mid s_2(x, y)\}$

employed $\rightarrow \{x, w \mid s_3(x, w)\}$,

а S состоит из следующих элементов:

$s_1^D :$	12	calvin	rome	21
	15	alice	hong kong	24
$s_2^D :$	AF	hotdog corp.		
	BN	banana ltd.		
$s_3^D :$	12	AF		
	16	BN		

Теперь предположим, что пользователь создал запрос q, желая узнать коды сотрудников (атрибут {Ecode}) : $\{x \mid \text{employee}(x, y, z)\}$. Результатом данного запроса по идее должно являться объединение $s_1^D \cap s_2^D \cap s_3^D$ в проекции по x, т.е {12}, но, принимая во внимание ограничение employed[Ecode] $\subseteq$ employee[Ecode], результатом запроса является множество {12, 16}, хотя информации о сотруднике с кодом 16 нет: ограничение employed[Ecode] $\subseteq$ employee[Ecode] допускает такую ситуацию.

Нерешенные задачи

1. Математическая формализация операций этапов сравнения, согласования и реструктуризации схем данных на базе аппарата дескрипционных логик и логик предикатов первого порядка.

2. Доказательство непротиворечивости и функциональной полноты полученных преобразований.

3. Построение работающих моделей LAV и GAV и преобразования $I_{LAV} \leftrightarrow I_{GAV}$

4. Решение задачи построения глобальной схемы над базами данных из разных областей за счет логического программирования, то есть автоматического доказательства теорем.

5. Решение задачи доказательства полноты запроса: необходимо определить, возвращает ли запрос в терминах глобальной схемы 100% информации из источников, или же лишь подмножество ответов.

6. Разработка и обоснование архитектуры системы интеграции.

Список литературы

1. Дейт К. Дж. Введение в системы баз данных: Пер. с англ. - М.: Наука, 1980.-464 с.
2. Брешенков А.В. Базы данных. Проектирование баз данных на основе информации табличного вида. LAP LAMBERT Academic Publishing GmbH & Co. KG Dudweiler, rbr, 66123 Saarbrücken, Germany, 2011, 394 с.
3. A Survey of Approaches to Automatic Schema Matching, Erhard Rahm, Philip A. Bernstein, The International Journal on Very Large Data Bases, 10(4):334-350
4. Data Integration: A Theoretical Perspectives, Maurizio Lenzerini Dipartimento di Informatica e Sistemistica Universita di Roma "La Sapienza" , Via Salaria 113, I-00198 Roma, Italy