СЕМАНТИКА ЗАПРОСОВ В СИСТЕМАХ ДЕДУКТИВНЫХ БАЗ ДАННЫХ

Танянский С.С., Горпиненко Ю.С.
Харьковский национальный университет радиоэлектроники

Традиционный подход к реализации сценария: извлечения, преобразования и загрузки - extract, transform and load (ETL) требует от пользователя, во-первых, осуществления глубокой проработки задачи, связанной не только с описанием структуры и правил функционирования системы, но и с явным описанием процедур поиска решения. Во-вторых, эти процедуры, описанные на каком-либо формальном языке, необходимо "перевести" на входной язык системы управления базами данных (СУБД), что само по себе является трудоемкой задачей. Таким образом, применение традиционных языков программирования баз данных (например, языка SQL) для описания ETL процессов оказывается неэффективным.

Для обеспечения взаимной доступности данных из множества распределенных источников целесообразно применять распределенные архитектуры с промежуточным программным уровнем интеграции. Такой подход позволит достичь требуемого уровня гибкости, открытости и производительности распределенных информационных систем.

Системы управления базами данных, поддерживающие теоретико-доказательное описание базы данных (БД) и, в частности, обладающие способностью осуществлять логический вывод дополнительных фактов из множества явно заданных кортежей базовых отношений, путем применения определенных аксиом и правил вывода, называют дедуктивными базами данных (ДБД) [1].

Вывод в формальной логической системе является процедурой, которая из заданной группы выражений (программы DataLog) выводит семантически правильный результат. Эта процедура, представленная в определенной форме, и является правилом вывода [2]. Если выражения, образующие тело правила, является истинной, то гарантируется, что применение этого правила обеспечит получение истинного выражения в качестве заключения.

При обработке данных в дедуктивных системах, в частности выполнения запросов, вычисление множества ответов требует предварительного проведения логического вывода с учетом отображений источников и ограничений целостности виртуального хранилища. Такой вывод можно проводить в два этапа:

1. На первом этапе производится вычисление всех производных фактов без учета ограничений целостности виртуального хранилища. Результатом этапа будет множество извлеченных фактов EDB_{ret}, заданное над алфавитом виртуального хранилища $\mathbb{N}_{Var}$.
2. На втором этапе формируется множество $DDB_{ret} = \{EDB_{ret}, P\}$, составленное из множества извлеченных фактов EDB_{ret} и аксиом (ограничений целостности) виртуального хранилища. Для получения множества ответов относительно заданного набора данных, вычисляется запрос Q относительно DDB_{ret}. Полученный результат и будет представлять собой искомое множество данных (результат).

В общем, рассмотренный метод заключается в наполнении множества извлеченных фактов EDB_{ret}, путем применения правил отображения интегрированной системы данных к источникам.

Пусть задана система интеграции данных $\Psi = \{DDB, \bigcup_{i=1}^{m} DDB_{\Delta_i}, \mathcal{F}\}$, отображения $\mathcal{F}$ заданы в форме корректного комбинированного подхода или Global-Local-As-View (GLAV) отображения $C_\Delta \subseteq C$, где C_Δ, C - конъюнктивные правила вывода, в том числе с простыми ограничениями. Пусть $DBD_\Delta \overset{def}{=} \{\bigcup_{i=1}^{m} EDB_{\Delta_i}, \bigcup_{i=1}^{m} P_{\Delta_i}\}$ - объединение ДБД-источников и $\mathbb{C}$ - множество констант, над которым определены факты дедуктивной базы данных DBD_Δ. Множество извлеченных фактов EDB_{ret} определяется алгоритмически следующим образом:

пусть изначально множество EDB_{ret} пустое. Далее для каждого отображения $C_\Delta \subseteq C$ из множества $\mathcal{F}$ выполняются действия: из множества фактов S', являющихся логическим следствием из $C_\Delta(\ DDB_\Delta\ )$, добавляются в множество EDB_{ret} факты следующим образом: для каждой подцели конъюнктивного правила вывода C_Δ, образующих тело конъюнктивного запроса Q^{def}, задаваемой в форме предиката $p(t_1,t_2,\ldots, t_n)$, где p – предикатный символ, определенный в DDB_{Δ_i}, $t_1,t_2,\ldots, t_n$ – термы ($t_i \in \mathbb{N}_{Var} \cup \mathbb{C}$), добавляются в множество EDB_{ret} факты, при этом каждая свободная переменная запроса заменяется на соответствующую константу из вектора $\underline{t}$, а каждая экстенсиональная переменная - на сопоставленную ей новую константу из множества $\mathbb{C}^*$($\mathbb{C}^* \cap \mathbb{C} = \varnothing$), неиспользовавшуюся ранее.

Приведенный метод легко расширить также для случая, когда $C_\Delta \in CD_S$, т. е., когда отображения могут содержать простые ограничения. Такие ограничения определяют дополнительную информацию, которую необходимо учитывать при логическом выводе для корректности и полноты множества продуцируемых фактов. Для учета ограничений, дополним определение множества извлеченных фактов EDB_{ret} для случая, когда Q_Γ содержит простые ограничения, добавив правило для каждого вектора констант $\underline{t} \in q_\Delta(DBD_\Delta)$ следующим образом: для каждой цели ограничения в определении конъюнктивного запроса с простыми ограничениями q_Γ^{def}, задаваемого в форме $op_1(u)$ или $(u\ op_2\ v)$, где $u \in \mathbb{N}_{Var}$, $v \in \mathbb{C}$, op_1 - унарный встроенный предикат, op_2 - бинарный встроенный предикат, добавим в EDB_{ret} аксиому в форме $op_1(u)$ или $(u\ op_2\ v)$ соответственно. При этом каждую свободную переменную запроса поменяем на константу из вектора $\underline{t}$, а каждую экстенсиальную переменную - на сопоставленную ей новую константу из множества $\mathbb{C}^*$.

Извлеченной ДБД для системы интеграции данных Ψ будем называть дедуктивную базу данных $DDB_{ret}(\Psi)$, описанную на языке DataLog и составленную из множества аксиом (ограничений целостности и правил вывода) виртуального хранилища P и множества извлеченных фактов $EDB_{ret}(\{\ DDB_{\Delta_i}\ \}_{i=1,\ldots m}, \mathcal{F})$, в обозначении

$$DDB_{ret}(\Psi) = \{EDB_{ret}(\bigcup_{i=1}^{m} DDB_{\Delta_i}\ , \mathcal{F}), P\}.$$

Пусть задана система интеграции данных Ψ, где отображения ДБД $\mathcal{F}$ заданы в форме корректных GLAV-отображений $C_\Delta \subseteq C$, где C - конъюнктивные правила вывода, в том числе с простыми ограничениями. В таком случае множество глобальных моделей системы интеграции данных Ψ совпадает с множеством моделей извлеченной ДБД $DDB_{ret}(\Psi)$, т. е. справедливо равенство $M(\Psi) = M(DBD_{ret}(\Psi))$.

Исходя из вышесказанного, также можно утверждать, что множество точных ответов на пользовательский запрос Q, относительно системы интеграции Ψ, совпадает с множеством ответов на этот запрос относительно извлеченной ДБД $DBD_{ret}(\Psi)$, т. е. выполняется равенство $Q(\Psi) = Q(DBD_{ret}(\Psi))$.

Таким образом, вычислив извлеченное множество фактов, можно не рассматривать источники данных и отображения. Для ответа на запрос достаточно учитывать только извлеченные факты и аксиомы глобальной ДБД. Следует отметить, что извлеченные факты могут быть не полны относительно аксиом, соответственно для вычисления ответа на запрос потребуется провести логический вывод на глобальном уровне. Более того, извлеченные факты могут противоречить аксиомам (извлеченная ДБД может быть противоречива), в таком случае система интеграции данных является противоречивой.

1. Дейт К. Введение в системы баз данных // К. Дейт / 6-е изд. Киев. М.: Диалектика, 1998. – 784 с.

2. Черри С. Логическое программирование и базы данных // С. Черри, Г. Готлоб, Л. Танка / М.: Мир. – 1992. – 352 с.