ПРИВЕДЕНИЕ ДАННЫХ. ОНТОЛОГИЧЕСКИЙ ПОДХОД

О.М. Атаева, А.Н. Бездушный

Введение

С каждым годом количество накопленной информации под управлением различных систем, в различных форматах увеличивается. Все эти системы и форматы сильно различаются между собой и можно сказать, что накопленные знания сильно различаются между собой как по физической, так и по логической структуре. С развитием технологий возникает необходимость в разных областях научных знаний объединения информации из этих источников в хранилища данных. Одним из наиболее важных процессов в технологии хранилищ данных является ETL – процесс, который извлекает информацию из исходного источника данных, информация в котором может быть представлена в различных моделях данных, приводит ее к каноническому виду, задаваемому источником назначения, а затем загружает в нее преобразованную информацию.

Существует большое разнообразие подходов к решению задач возникающих на многоуровневом этапе процесса приведения данных к каноническому виду, которые эффективно используются в различных приложениях, но недостатком всех этих подходов является то, что с увеличением источников данных и потоков данных, поступающих из них, эффективность этих решений падает. Особенно остро стоит эта проблема для источников научной информации, из-за сложности используемых понятий и отношений между ними. Также, они подвержены более частому изменению структур данных, что неизбежно приводит к необходимости внесения существенных доработок в уже имеющиеся решения. Учитывая эти особенности источников научных данных, разрабатываемые решения привидения, должны легко адаптироваться и для таких изменений. Процесс приведения данных является также важным этапом в анализе работы существующих систем при анализе доступности и полезности информации, хранимой в системе.

«Качество данных» основные направления приведения данных

Все работы по приведению данных проводятся в тесной связи с попыткой дать ответ на вопрос «Что такое качество данных?». Определение «качества данных», которое используется нами, охватывает основные характеристики качества на уровне данных, такие как: полнота, точность, уникальность, каждая из которых, в свою очередь, является многомерной величиной и включает детализирующие характеристики. Рассматриваемое нами определение не является жестким, так как дальнейшее развитие работ может выявить необходимость для его расширения или уточнения.

Основные направления приведения данных

Выделяют несколько основных направлений подзадач возникающих в процессе приведения данных:

- *профайлинг данных* ориентирован на анализ отдельных атрибутов объектов,
- *нормализация/стандартизация значений* в соответствие задаваемой канонической схемой,
- *выявление недостающих значений*, такими значения могут быть:
  - *пропущенные*, которые соответствуют *обязательным* атрибутам
  - *отсутствующие,* которые соответствуют *дополнительным* атрибутам
- *вычисление недостающих объектов,* на которые ссылаются другие объекты в потоке,
- *выявление недостоверных объектов* с восстанавливаемыми недостоверными значениями *обязательных* атрибутов, объекты их содержащие признаются как недостоверные,
- *выявление недостоверных значений*, ориентировано на выявление восстанавливаемых недостоверных значений *дополнительных* атрибутов,
- *выявление фиктивных значений* относится к выявлению невосстанавливаемых значений *дополнительных* атрибутов,
- *выявление фиктивных объектов* относится к выявлению невосстанавливаемых значений *обязательных* атрибутов, объекты их содержащие признаются как фиктивные,
- *выявление дублирующих объектов* во множестве объектов потока данных.

Показано что каждое направление привидения соответствует определенной характеристике «качества данных».

Постановка задачи

Область привидения данных плохо формализована. Некоторые вопросы, хорошо исследованы, но моделей, описывающих все компоненты построения процесса приведения в целом, не существует. Это осложняет построение систем для управления процессами приведения при большом количестве потоков данных из разнородных источников. В работе приводится описание формальной модели, на основе которой создана система управления процессом привидения данных.

Для формального определения общей задачи приведения данных вводится понятие базиса процесса приведения W как тройки $W=(\Gamma,T,U)$, где Γ поток данных, T множество действий, U множество условий.

Γ состоит из объектов вида $A(a_1,...,a_n)$, типы значений атрибутов a_i могут быть как элементарные базовые типы, так и другие объекты из Γ. В соответствии с типом значений выделяются следующие виды атрибутов: атрибуты с одним значением элементарного или объектного типа данных, многозначные компоненты, имеющие множество однотипных значений, которые могут составлять - мультимножество, множество, список, массив.

По своему использованию, назначению, по отношению к тем или иным функциям атрибуты подразделяются на следующие пересекающиеся подмножества: обязательные, классифицирующие, озаглавливающие, идентифицирующие, поисковые и описательные атрибуты. Каждое направление приведения данных использует определенные подмножества в процессе своей работы

Множество $T=T_v\cup T_\varepsilon$, где $T_v=\{t_1,...,t_m\}$ состоит из действий t_i, которые могут менять структуру или значения объектов $A\in\Gamma$, а множество $T_\varepsilon=\{\varepsilon\}$ состоит из одного действия ε, которое никак не влияет на объекты из Γ..

Множество $U=U_{Lex}\cup U_{Syn}\cup U_{Sem}$ состоит из непересекающихся подмножеств лексических U_{Lex}, синтаксических U_{Syn} и семантических U_{Sem} условий, состоит из условий накладываемых на Γ согласно схеме приемника S_x и дополнительных условий определяемых экспертом.

Для каждого условия $u\in U$ определена пара $(\varepsilon,t)\in T_\varepsilon\times T_v$, где ε определяет выполнение условия u для объектов $A\in\Gamma$, а действия $t\in T_v$ приводит те объекты из Γ, для которых условие u не выполняется, в соответствие с условием u.

Для построения концептуальной модели задачи привидения данных вводятся дополнительные сущности для оценки эффективности приведения и ее оптимизации, определяются связи между введенными сущностями, вводятся ограничения на допустимые сочетания компонентов модели и некоторые их свойства. Определяются условия для выделения допустимого порядка применения действий для множества рассматриваемых условий, при котором достигается лучший результат.

Реализация

Представленная формальная модель описывает в комплексе все необходимые компоненты приведения данных и позволяет описать задачи приведения данных в виде потоков работ, так как определяет ключевые сущности необходимые для описания потока работ, на основе онтологии BWW. Модель делает возможным создание гибких настраиваемых систем приведения для различных источников и потоков данных. На основе данной модели была разработана первая версия системы приведения данных на базе ИС «Научный Институт РАН».

Цель процесса приведения данных достигается посредством использования множества независимых действий, которые могут быть использованы в различной последовательности в зависимости от определенных в задаче условий и от «качества данных», которое должно быть достигнуто. Порядок и контекст, в котором используются эти действия, определяют их семантику и влияние на «качество».

Типы событий, которые могут инициировать и синхронизировать действия в процессе привидения данных, могут быть временные, прерывающие процесс или другие события инициированные пользователем. Действия потока работ при привидении данных не выполняются сразу же после их инициирования, а могут зависеть также от текущего состояния потока данных. Таким образом, состояние входящих данных так же является необходимым условием для координации действия.

В потоке работ процесса привидения данных могут использоваться различные действия по своей семантике, но стратегия процесса разрабатывается независимо от того, что именно выполняют действия, некоторые процессы могут объединяться, некоторые разбиваться на более мелкие. Поток работ приведения данных обладает достаточной гибкостью для динамической настройки и определения координирующих их действий и событий.

В качестве используемых ресурсов выступают как эксперты, которые определяют требования ограничения и стратегии процесса приведения, данных также компьютерные системы, которые участвуют в процессе управления данными.

При автоматизации задачи приведения данных в рамках проекта ИС «НИ РАН» создана настраиваемая система приведения данных не зависящая от типов обрабатываемых данных и их структуры. На данный момент в реализации поддерживаются направления профайлинг данных, нормализация/стандартизация, выявление дубликатов. Система работает в режиме наведения порядка в хранилище. Описание сценария работ по привидению данных записывается в определенном XML синтаксисе и подается пользователем через веб –

форму. Задание последовательности работ в виде правил освобождает пользователя от знания деталей реализации алгоритмов приведения. Система настраивается под любые типы ресурсов и позволяет конструировать процесс приведения данных, комбинируя при необходимости различные методы и определяя правила их использования. Для каждого процесса приведения определяются его количественные характеристики, которые можно предварительно рассчитать на соответствующих тестовых множествах. Дальнейшая работа предполагает развитие системы в рамках расширения поддерживаемых направлений процесса приведения, расширение библиотек используемых методов и т.д.

Одной из ключевых идей является проведение четкой грани между логическими требованиями к трансформации данных и их реализацией, и концентрация усилий разработчика на описании потока работ приведения данных. В построении системы используется опыт построения более ранних систем приведения данных, решения для которых специализированы и набор применяемых правил и операций ограничен рамками предметной области. В настоящее время приходится переосмысливать эти подходы в применении к онтологической модели данных и качественно иному подходу к построению информационных систем.

ЛИТЕРАТУРА:

1. Andreas Maier, J. Aguado, A. Bernaras, I. Laresgoiti, C. Pedinaci, N. Pena, T. Smithers. Integration with Ontologies. Conference Paper WM2003, April 2003, Luzern
2. Heiko Müller, Johann-Christoph Freytag. Problems, Methods, and Challenges in Comprehensive Data Cleansing. 2003
3. Erhard Rahm, Hong Hai Do. Data Cleaning: Problems and Current Approaches. 2000
4. Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data Quality Assessment. 2004
5. David A. Garvin. Completing on the Eigth Dimensions of Quality. 1987
6. H. Galhardas, D. Florescu, D. Shasha, E. Simon, C. Saita: Declarative Data Cleaning: Language, Models and Algorithms. 2001
7. Wand Y., Weber R. An ontological model of an information system // IEEE Transactions on Software Engineering Journal. – 1990. – 16(11). – P. 1281-1291.
8. Атаева О.М., Шиолашвили Л.Н. Методы очистки интегрируемых данных// Современные проблемы фундаментальных и прикладных наук: Труды XLIX научной конференции. /Моск. физ.-тех. ин-т. – М., 2006.