УДК 004.62:004.42

АЛГОРИТМ СОПОСТАВЛЕНИЯ СХЕМ ДАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ НЕФТЕГАЗОДОБЫВАЮЩЕГО ПРЕДПРИЯТИЯ

В.В. Вейбер, А.В. Кудинов, Н.Г. Марков

Томский политехнический университет
E-mail: webvad@tpu.ru

Предложен оригинальный алгоритм автоматического сопоставления XML схем – PSM, реализованный в рамках создания платформы интеграции производственных данный нефтегазодобывающей компании и предоставлены результаты экспериментов, подтверждающих его эффективность.

Ключевые слова:
XML схема, сопоставление схем данных, семантическое сопоставление.

Key words:
XML schema, schema matching, semantic matching.

Введение

При интеграции данных между несколькими независимыми информационными системами возникает необходимость однозначной идентификации экземпляров их сущностей. Само существование такой задачи обусловлено независимостью информационных систем – каждая система использует собственные правила именования сущностей и их атрибутов, в результате чего возникают рассогласования в наименованиях таблиц, полей и других объектов баз данных информационных систем, предназначенных для описания семантически эквивалентных объектов, процессов и явлений. При этом объектом анализа являются, чаще всего, не сами данные информационной системы, а их структурированное описание или *метаданные*. В свою очередь, под *схемой данных*, будем понимать любую структуру метаданных, которая описывает правила хранения данных и/или способ организации доступа к ним. Примером схемы данных является схема базы данных.

Сопоставление схем данных можно представить как ($S(s)$, $S(t)$, $\{Mi\}$), где $S(s)$ – исходная схема, $S(t)$ – целевая схема и $\{Mi\}$ – набор соответствий элементов схемы исходной по отношению к целевой схеме (рис. 1).

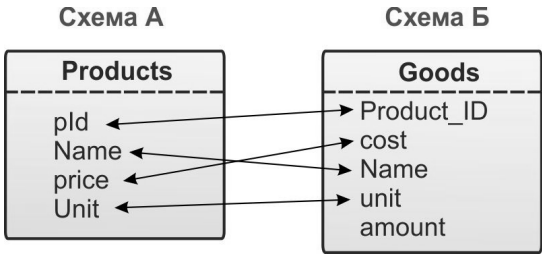

Рис. 1. *Пример схем данных и их сопоставления*

Сопоставление схем данных (*schema matching*) играет ключевую роль в решении различных задач: интеграции и трансформации данных, создании хранилищ данных, интеграции и сопоставлении данных на основе онтологий (в том числе при реализации концепции Semantic Web) и т. д. В процесе сопоставления участвует две разнородные схемы данных (возможно дополненные вспомогательной информацией), результатом сопоставления является набор соответствий (*mapping*), который отражает связь между элементами исходных схем [1, 2].

При решении задачи сопоставления схем необходимо учитывать ряд аспектов. *Во-первых*, задача сопоставления схем является задачей семантического анализа, при этом нужно учитывать смысл не только в именовании элементов схем, но в логике их построения, а также контексте конкретной задачи. *Во-вторых*, анализу требуется подвергнуть большое количество неупорядоченной или слабоструктурированной информации, например, документы описания схемы или образцы данных схемы.

Сопоставление схем вручную – длительный и дорогостоящий процесс, поэтому разработка алгоритма, автоматизирующего этот процесс, является актуальной задачей.

Цель данной работы: в рамках создания платформы интеграции производственных данных нефтегазодобывающей компании [3] разработать алгоритм сопоставления схем данных для обеспечения связывания логически идентичных экземпляров сущностей, поступающих из разных информационных систем, и минимизировать долю ручного труда в процессе этого связывания.

Существующие решения

В настоящее время предложено много подходов к решению задачи сопоставления схем данных [1, 2, 4–6]. Обычно эти подходы используют комбинации ранее созданных алгоритмов нечеткого сравнения строк (*matchers*).

Исходными данными для сопоставления схем могут являться:

- названия элементов схемы;
- данные (значения);
- общие шаблоны или фразы в данных;
- ограничивающие условия схемы;
- структурная информация;
- специфика предметной области – синонимы, схожие названия и т. п.

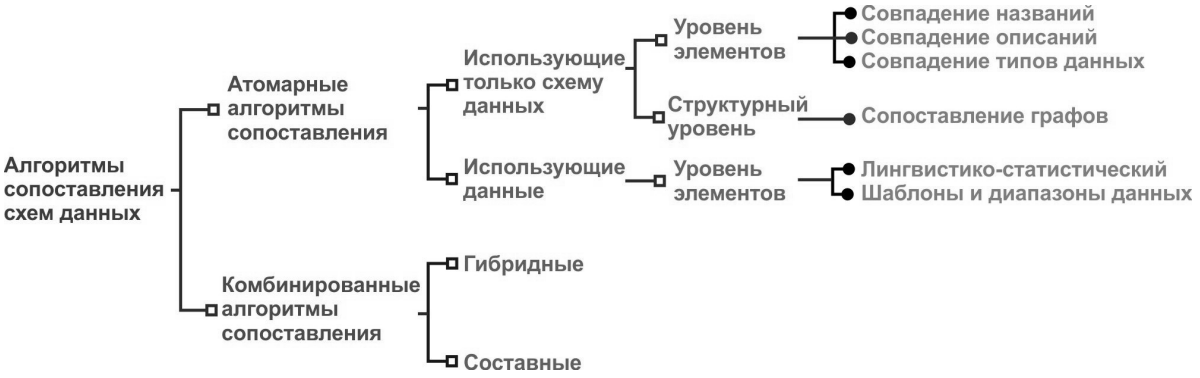

Рис. 2. *Классификация алгоритмов, используемых для сопоставления схем данных*

Классификация, составленная по результатам анализа алгоритмов сопоставления схем данных [1, 2, 4–6], представлена на рис. 2.

Чаще всего самый весомый вклад в результаты сопоставления вносит синтаксический подход (совпадение названий). Методы синтаксического сравнения строк определяют сходство между элементами, основываясь только на эквивалентности названий. Наиболее простой метод для определения равенства имен – точное соответствие строк. Другой, более универсальный вариант, это использование методов нечеткого сравнения строк. Алгоритмы, на основе таких методов были разработаны и применены в других, смежных областях, таких как полнотекстовый поиск, автоматическое исправление опечаток и ошибок и т. д. Наиболее распространенными являются алгоритмы синтаксического нечеткого сравнения строк, такие как алгоритм редактирования расстояния (Левенштейна), алгоритм *N*-грамм и алгоритм SoundEx. Так алгоритм Левенштейна вычисляет минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую [7]. Алгоритм *N*-грамм сравнивает строки в соответствии с их множеством *N*-грамм, т. е. последовательностей символов *N*, например, 2-грамм, 3-грамм [7]. Алгоритм SoundEx, напротив, вычисляет *фонетические* сходства между строками и может быть полезен при сравнении слов, которые написаны по-разному, но имеют одинаковое произношение и значение [7].

Помимо синтаксической информации, содержащейся в схемах данных, важную роль играет семантическая информация: формат и типы данных, допустимые значения, ссылочные ограничения и т. д. Семантическое соответствие обусловлено также терминологическими отношениями (синонимы, гиперонимы и гипонимы). Например, слова «автомобиль» и «машина» являются синонимами и, следовательно, соответствуют друг другу. Это требует использования дополнительных источников, таких как словари, тезаурусы, онтологии.

Очевидно, что эффективным будет алгоритм сопоставления схем данных, комбинирующий сразу несколько подходов [2]. Такие алгоритмы уже разработаны, например, в рамках проекта COMA+ [2], но детали их реализации закрыты, описание большинства методов отсутствует, что делает невозможной их модификацию для учета специфики конкретных прикладных задач. Поэтому задача разработки эффективного алгоритма сопоставления схем данных является актуальной.

Алгоритм PSM

Предлагаемый алгоритм разработан в рамках создания платформы интеграции производственных данных нефтегазодобывающего предприятия [3] с целью автоматического сопоставления схем данных информационных систем, используемых в нефтегазовой отрасли. Платформа предназначена для интеграции данных на основе общей метамодели, базирующейся на отраслевом стандарте PRODML [8]. Это обуславливает специфику алгоритма PSM (PRODML Schema Matcher) – одна из схем данных заранее известна – это схема PRODML.

Алгоритм разработан для поиска соответствий между двумя схемами данных, представленных в формате XSD (XML Schema Definition). По классификации, приведенной на рис. 2, он относится к алгоритмам, использующим только схему данных, т. е. имена элементов, типы данных и структурные свойства элементов схем. Логическая схема работы алгоритма представлена на рис. 3. Основные шаги алгоритма PSM реализуют нижеперечисленные функции.

- Разбиение каждой строки названия элементов на подстроки по набору правил. Необходимость разбиения названий элементов схем обусловлена тем, что они могут быть составными (Camel-стиль, Pascal-стиль), т. е. не иметь пробельных разделителей между словами. Например, fileName можно разбить на подстроки file и Name.
- Расчет соответствия между списками подстрок названий элементов с использованием синтаксических алгоритмов (Левенштейна, *N*-грамм и комбинированного).
- Расчет семантического соответствия между списками названий элементов схем с использованием словаря. Предметная область применения алгоритма PSM и целевая схема данных известны заранее, поэтому создание словаря си-

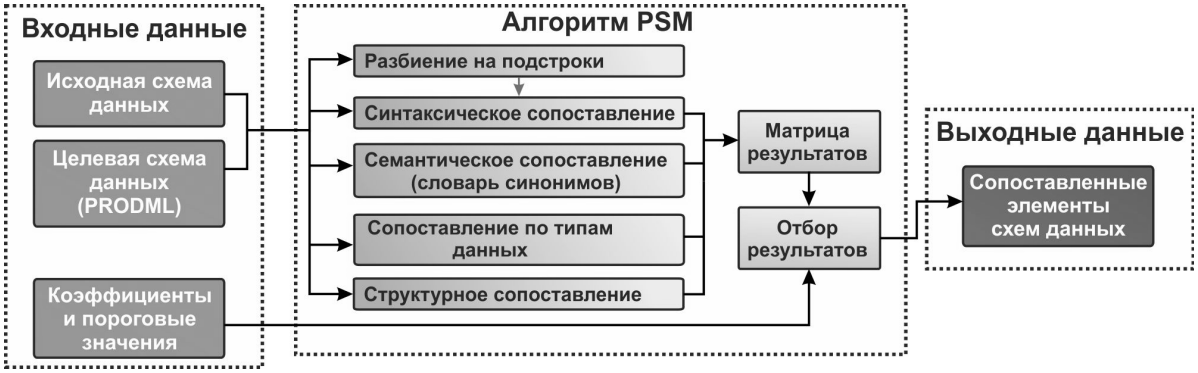

Рис. 3. *Логическая схема работы алгоритма сопоставления схем данных PSM*

нонимов элементов схемы PRODML позволяет существенно повысить качество получаемых результатов. Кроме того, словарь позволяет сделать алгоритм применимым для сопоставления схем данных на любом языке.

- Вычисление соответствия между элементами схем по типам данных. Для сопоставления типов данных использована таблица коэффициентов совместимости типов данных [7].
- Вычисление соответствия между элементами схем на основе структурных зависимостей.
- Выбор соответствий между элементами исходных схем данных на основе матрицы результатов предыдущих блоков. Выходные данные блоков алгоритма формируют матрицу результатов, на основе которой формируется конечный результат сопоставления элементов исходных схем данных.

Алгоритм PSM формирует окончательный результат (набор соответствий элементов схем) на основе результатов соответствий полученных 4-мя методами (*matcher*). Для расчета комбинированного значения соответствий нескольких методов могут использоваться различные стратегии агрегации:

- максимальное значение среди всех методов используется как результирующее;
- минимальное значение – как результирующее;
- среднее значение всех методов как результирующее;
- весовые коэффициенты – каждый метод наделяется своим весом.

Стратегия использования весовых коэффициентов является наиболее гибкой, экспериментально с использованием этой стратегии были получены высокие результаты.

На псевдокоде работу алгоритма можно представить в следующем виде:

1 for($s \in N_S$)
2 for($t \in N_T$){
3 sSim(s,t)=SemanticSimilatary(s,t)
4 dictSim(s,t)=DictionarySemanticSimilatary(s,t)
5 dtSim(s,t)=DataTypeSimilatary(s,t)
6 stSim(s,t)=StructureSimilatary(s,t)
7 }
8 for($s \in N_S$)
9 for($t \in N_T$){
10 combSimilitaryMatrix(s,t)=α_1*sSim(s,t)+
+α_2*dictSim(s,t)+α_3*dtSim(s,t)+α_4*stSim(s,t)
11 if (similitaryMatrix (s, t)>threshold)
12 Result=Result+(s, t, similitaryMatrix (s, t))
13 }

Здесь N_S и N_T – исходная и целевая схемы данных; s и t – элементы исходной и целевой схем данных; *threshold* – пороговое значение; α_1, α_2, α_3, α_4 – весовые коэффициенты, определяющие вклад каждого из методов в общий результат.

Оценка качества автоматического сопоставления схем данных

Для оценки качества автоматического сопоставления схем данных сопоставление должно быть предварительно произведено вручную и использовано в качестве эталона. Множество полученных соответствий состоит из множества B – истинных положительных (правильных соответствий), множества C – ложных положительных соответствий, множества A – ложных отрицательных и множества D – истинных отрицательных.

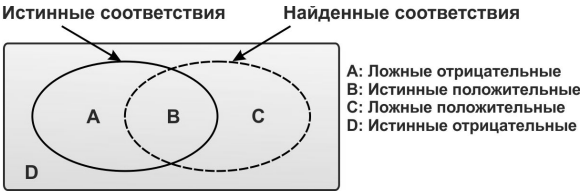

Рис. 4. *Множества соответствий схем данных*

Точность отражает долю правильных соответствий среди всех найденных (рис. 4):

Точность=|B|/(|B|+|C|).

Полнота показывает долю правильных найденных соответствий среди всех правильных (рис. 4):

Полнота=|B|/(|A|+|B|).

Иногда бывает полезно объединить точность и полноту в одной усреднённой величине. Для этой цели среднее арифметическое не подходит, т. к., например, алгоритму сопоставления схем достаточно вернуть вообще все возможные результаты, чтобы обеспечить равную единице полноту при близкой к нулю точности, и среднее арифметическое точности и полноты будет не меньше 1/2. Среднее гармоническое не обладает этим недо-

статком, поскольку при большом отличии усредняемых значений приближается к минимальному из них. Поэтому хорошей мерой для совместной оценки точности и полноты является F-мера (*F-measure*), которая определяется как взвешенное гармоническое среднее точности и полноты [7]:

F-мера (α)=|B|/((1−α)*|A|+|B|+α*|C|)

F-мера (α)=Точность*Полнота/((1−α)*Точность+α*Полнота)

Коэффициент α позволяет задавать относительную значимость точности и полноты, при равной значимости α=0,5.

Исследование эффективности алгоритма PSM

Разработанный алгоритм сопоставления схем данных PSM был опробован на ряде тестовых, специально сгенерированных схем, а также на наборе реальных схем данных для хранения информации по гидродинамическим исследованиям скважин, используемых популярными в нефтегазовой отрасли программными продуктами, такими как ПК «БАСПРО» [9], MES «Магистраль-Восток» [10] и продукты фирмы Schlumberger. Эффективность алгоритма PSM оценивалась только с точки зрения качества (F-меры) в сравнении с качеством работы комбинации аналогичных по принципу работы ал-

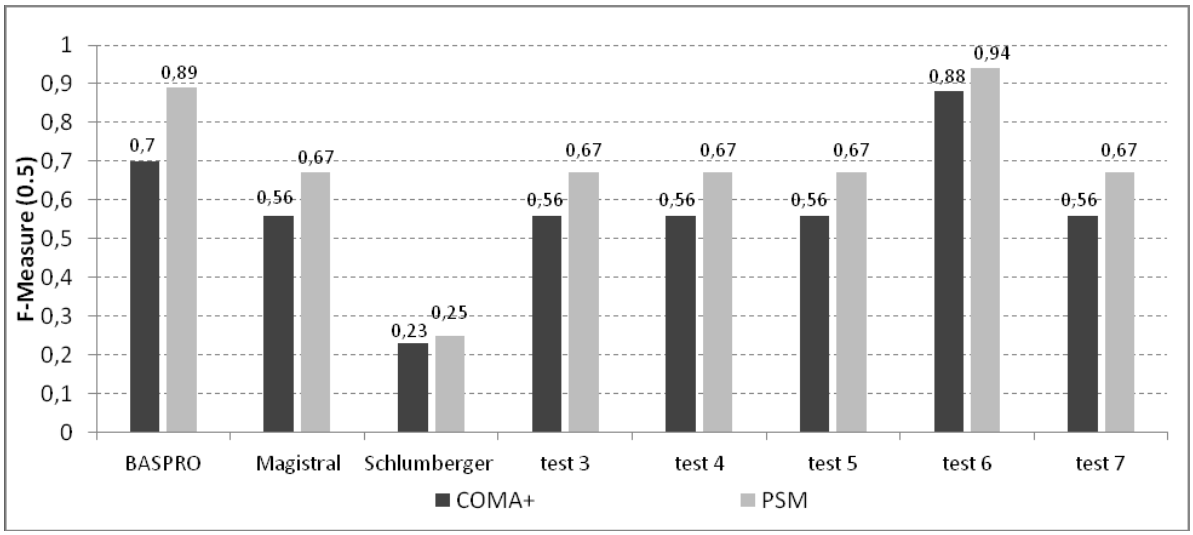

Рис. 5. *Результаты сравнения качества сопоставления на тестовых наборах*

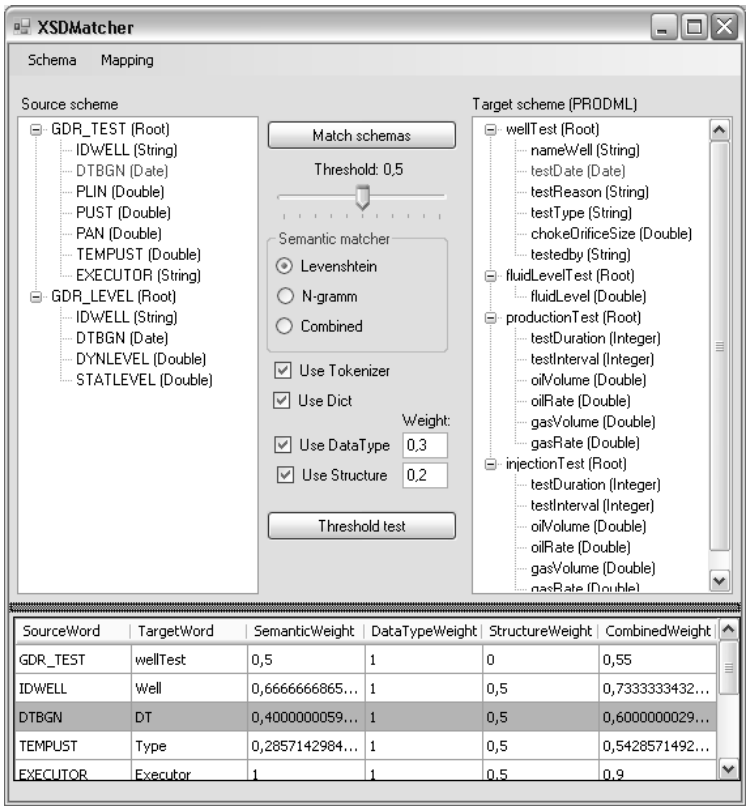

Рис. 6. *Интерфейс прототипа, реализующего алгоритм PSM*

горитмов из широко распространенной платформы COMA+ [7].

Для проведения этих экспериментов весовые коэффициенты для расчета комбинированного значения соответствия (строка 10 псевдокода) и пороговое значение (строка 11 псевдокода) получены эмпирически на наборе пар схем с заранее известными соответствиями элементов схем. Определение оптимального набора значений весовых коэффициентов и порогового значения, подходящих для решения более широкого круга задач, является темой для дальнейшего исследования.

Полученные результаты оценки качества сопоставления сравниваемых алгоритмов (рис. 5) показали, что алгоритм PSM имеет более высокие результаты на всех тестовых наборах. Оценка эффективности алгоритма PSM проведена с помощью разработанного прототипа (рис. 6).

Алгоритм PSM был использован при решении реальных интеграционных задач на предприятиях нефтегазовой отрасли, таких как ОАО «ВостокГазпром» и ОАО «Томскгазпром».

Выводы

Рассмотрена задача сопоставления схем данных и существующие подходы для ее решения. Предложен оригинальный алгоритм PSM для сопоставления схем данных информационных систем для нефтегазодобывающих предприятий, а также приведены результаты исследований его эффективности. Предложенный алгоритм реализован в составе платформы интеграции производственных данных нефтегазодобывающей компании и прошел апробацию при решении реальных производственных задач.

СПИСОК ЛИТЕРАТУРЫ

1. Вейбер В.В., Богдан С.А., Кудинов А.В., Марков Н.Г. Концепция построения платформы для интеграции производственных данных нефтегазодобывающей компании // Известия Томского политехнического университета. – 2011. – Т. 318. – № 5. – С. 126–131.
2. Rahm E., Bernstein P.A. A survey of approaches to automatic schema matching // VLDB Journal. – 2001. – № 10. – P. 334–350.
3. Do Hong-Hai, Rahm E. Matching large schemas: Approaches and evaluation // Information Systems. – 2007. – № 6. – P. 857–885.
4. Villanyi B., Martinek P., Szikoraa B. Framework for schema matcher composition // WSEAS transactions on computers. – 2010. – № 10. – P. 1235–1244.
5. Chen Y.P., Promparmote S., Maire F. MDSM: Microarray database schema matching using the Hungarian method // Information Sciences. – 2006. – № 8. – P. 2771–2790.
6. Jeong B., Lee D., Cho H., Lee J. A novel method for measuring semantic similarity for XML schema matching // Expert Systems with Applications. – 2008. – № 3. – P. 1651–1658.
7. Hai D.H. Schema matching and apping-based data integration: PhD thesis. – Leipzig, 2005. – 204 p.
8. PRODML Work Group. Reference Architecture PRODML 1.0 // Energistics: The Energy Standards Resource Center. 2011. URL: http://www.energistics.org/production (дата обращения: 27.07.2011).
9. Баспро Оптима // БАСПРО. 2011. URL: http://www.baspro.ru/programm (дата обращения: 27.07.2011).
10. Богдан С.А., Кудинов А.В., Марков Н.Г. Опыт внедрения MES «Магистраль-Восток» в нефтегазодобывающей компании // Автоматизация в промышленности. – 2010. – № 8. – С. 53–58.

Поступила 16.08.2011 г.