УДК 004.62 4

Создание интегрированной модели данных из разнородных источников, содержащих цифровые следы

П. К. Чернов[1], Е. А. Рабчевский[2]
[1]Пермский государственный национальный исследовательский университет; Пермь, Россия
e-mail: ch3rn0vpk@gmail.com
[2]ООО "СЕУСЛАБ"; Пермь, Россия
e-mail: e.rabchevskiy@seuslab.ru

Рассматривается подход к созданию интегрированных моделей данных на основе онтологической модели, а также приводится пример работы прототипа, который реализует интеграцию данных из нескольких разнородных источников и осуществляет логический вывод для реализации скоринговой системы оценки риска мошенничества.

Ключевые слова: *аналитическая система; модель данных; онтология.*

Поступила в редакцию 15.02.2022, принята к опубликованию 15.05.2022

Creation of an Integrated Data Model of the Heterogeneous Sources Containing Digital Footprints

P. K. Chernov[1], E. A. Rabchevskiy[2]
[1]Perm State University; Perm, Russia
e-mail: ch3rn0vpk@gmail.com
[2]"SEUSLAB" LLC; Perm, Russia
e-mail: e.rabchevskiy@seuslab.ru

The article revealed the creation of an integrated data models based on ontology. The prototype that implements the integration of data by several heterogeneous sources and performs a logical reasoning required by the scoring anti-fraud system is presented.

Keywords: *analytical system; data model; ontology.*

Received 15.02.2022, accepted 15.05.2022

DOI: 10.17072/1993-0550-2022-2-81-87

Введение

Современные аналитические системы сталкиваются с проблемой обработки большого количества противоречивых данных из множества источников, при этом количество источников данных постоянно растет.

В связи с этим актуальной является задача интеллектуальной интеграции данных, позволяющей представить данные из разнородных источников в виде единой модели и установить между ними логические связи.

В данной работе рассматривается подход к созданию интегрированных моделей данных на основе онтологической модели, а также приводится пример работы прототипа, который реализует интеграцию данных из нескольких разнородных источников и осуществляет логический вывод для реализации скоринговой системы оценки риска мошенничества.

1. Описание подхода

В данной работе рассматривается подход, который базируется на концепции, разработанной в компании ООО "СЕУСЛАБ".

Главным продуктом данной компании является аналитическая поисковая система "СЕУС", использующая данные о цифровых следах из открытых источников и разрабатываемая для нужд государственного сектора и силовых структур.

В основе концепции лежит сетевая модель данных, которая связывает между собой онтологии [1], относящиеся к различным уровням (данные, задачи обработки, знания) в соответствии с их предназначением (рис. 1).

При этом онтологии описываются с помощью языка OWL2 (Web Ontology Language) [2], реализуют дескрипционную логику и используют продукционные правила для получения новых знаний.

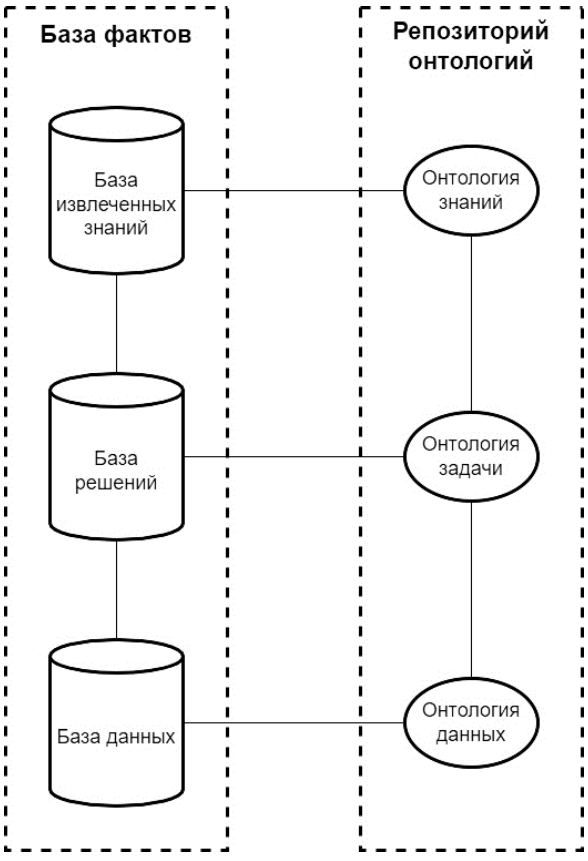

Рис. 1. *Связи между онтологиями разных уровней*

2. Сравнение с подходом на основе реляционной модели

Описанная концепция была сопоставлена с подходом на основе реляционной модели. Данный подход предполагает хранение данных за счет использования совокупности взаимосвязанных таблиц [3].

Подход к созданию систем с поддержкой логического вывода, при котором сведения хранятся в реляционной базе данных, имеет примеры использования на практике. Например, подобная реализация применялась при разработке проекта (ONTO)2 Agent [4].

В результате анализа были выявлены следующие достоинства и недостатки подхода на основе онтологической модели по сравнению с подходом на основе реляционной модели:

1. Меньшая производительность хранилища данных. Ввиду необходимости поддержания специфического формата представления данных ("тройки"), разработчики хранилищ "троек" не имеют возможности оптимизировать формат физического хранения данных на том же уровне, что и разработчики реляционных СУБД. По этой причине хранилища "троек" уступают в производительности реляционным базам данных.

2. Более простой, нативный и функциональный способ создания базы знаний – онтологии – уже включают в себя механизмы для формирования правил вывода новых данных на основе имеющихся (аксиомы классов и свойств в OWL, SWRL-правила).

3. Более наглядная модель данных итоговой базы знаний – граф онтологии – более нагляден и прост в интерпретации, при этом он позволяет визуализировать взаимосвязи между сущностями, а также характер этих связей (взаимоисключение, обратная функциональность и т.п.).

3. Сравнение с аналогичными подходами на базе онтологий

Помимо сравнения с подходом на основе реляционной модели для сравнения были рассмотрены несколько подходов на базе онтологий.

В работе "Distributed Query Processing on the Cloud: the Optique Point of View" [5] описывается способ организации доступа пользователей к хранилищу интегрированных данных из разнородных источников, при котором пользователи имеют возможность формулиро-

вать запросы в терминах знакомой концептуализации базовой области. Основным элементом архитектуры при этом является онтология, описывающая предметную область приложения в терминах словаря классов ("concepts") и отношений между ними ("roles"), а также набор отображений, которые связывают термины онтологии и схему данных источника.

Таким образом, одной из целей данного подхода является упрощение доступа пользователей к данным за счет формулирования запросов в терминологии предметной области исходных источников данных вне зависимости от их структуры и способов интеграции в более высокий уровень представления. Описанный же в разделе № 1 подход подразумевает наличие нескольких онтологий различного уровня, в том числе уровня задач, что предполагает формулирование пользовательских запросов в терминах предметной области решаемых задач, а не исходных источников данных, с помощью которых предполагается решать обозначенные задачи.

В работе "Ontology-Based Data Access: Ontop of Databases" описываются архитектурные решения и технологии, лежащие в основе системы OBDA Ontop, которая реализует хранение данных за счет использования реляционных баз данных [6]. Помимо этого, приводится анализ производительности Ontop в серии экспериментов и демонстрируется, что для стандартных онтологий, запросов и данных, хранящихся в реляционных базах данных, Ontop работает быстро и эффективно. В основе системы Ontop лежит модификация стандартной архитектуры системы OBDA, взаимодействующей с источниками данных в формате реляционных баз данных.

Данная архитектура, в отличие от описанного в разделе № 1 подхода, предполагает формулирование запросов – conjunctive queries [7] – в терминах единственной во всей архитектуре онтологии.

Для реализации логического вывода в рамках данного подхода предполагается применение механизмов, позволяющих создавать дополнительные логические связи между теми или иными данными – средства SQL (и их расширения), а также инструменты конкретных систем управления базами данных.

К таковым можно отнести ограничения целостности, триггеры и хранимые процедуры.

4. Концепция

В соответствии с выбранным подходом была сформулирована концепция построения интегрированных моделей, которая показана на рис 2.

На уровне источников данных (DS) онтологии описывают информационные сущности, содержащиеся в источниках данных.

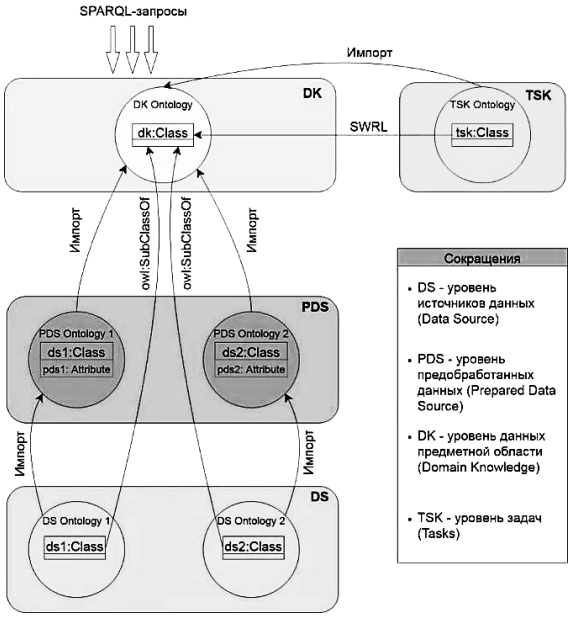

Рис. 2. *Схема построения интегрированной модели*

Развитием принципов, заложенных в упомянутой концепции, стало выделение дополнительного уровня – уровня предобработанных данных (PDS). На нем информация с предыдущего уровня дополняется данными, которых нет в источниках в явном виде, но которые могут быть получены путем определенной обработки исходных данных. Уровень предобработанных данных по своей сути является специализированным уровнем данных и потому может быть в него включен.

Дополнение информации с предыдущего уровня реализуется за счет добавления классам (owl:Class) из онтологий уровня источников данных новых свойств (DataProperty и ObjectProperty) [8].

Например, в источнике могут содержаться данные о номере мобильного телефона. При этом в источнике могут отсутствовать в явном виде данные о его принадлежности к региону или оператору.

Эти данные могут быть определена путем обработки номера.

Онтологии уровня предметной области (DK) объединяют в себе информационные сущности из всех источников данных. Между классами из онтологий уровня источников данных и уровня предметной области устанавливается связь наследования (owl:SubClassOf).

За счет этого экземпляр класса уровня источников данных будет автоматически отнесен к соответствующему классу уровня предметной области. В онтологиях уровня задач (TSK) формулируются понятия, необходимые для решения конкретных задач и отсутствующие на других уровнях.

Например, если данные анализируются для выявления мошенничества, то на уровне задач будут присутствовать такие информационные сущности как "мошенник", "подозрительная активность" и т. п. Информационные сущности с уровня предметной области отображаются в соответствующие сущности уровня задач за счет продукционных правил Semantic Web Rule Language [9].

Логический вывод в данной модели обеспечивается за счет механизмов языка OWL (таких как, например, аксиомы классов), продукционных правил (Semantic Web Rule Language), а также SPARQL-запросов.

5. Скоринговая система

Для того чтобы убедиться в работоспособности модели, был создан прототип, использующий структуру нескольких открытых источников данных для реализации скоринговой системы оценки риска мошенничества.

Главная задача системы – определить, отнести ли профиль человека к категории "потенциальный мошенник" или нет. Для принятия этого решения реализуется скоринговая система. Это означает, что профиль человека будет отнесен к категории "потенциальный мошенник" в том случае, если он имеет итоговый балл в системе выше определенного значения.

Итоговый балл определяется по итогу проверки нескольких условий. Если условие выполняется, то профилю человека начисляется определенное количество баллов.

Также для сравнения были рассмотрены несколько уже имеющихся подходов к созданию систем выявления мошенничества с применением онтологий.

В работе "Knowledge Base Ontology Building For Fraud Detection Using Topic Modeling" [10] описывается метод формирования базы знаний на основе онтологии для выявления мошеннических действий в цифровой среде.

В рамках данного метода текстовые документы подвергаются обработке, результатом которой являются наборы ключевых слов и тем. На их основе генерируется единая онтология, в которой между ключевыми словами и темами устанавливаются логические связи (например, связь типа "equivalentTo" между синонимами).

Этот подход предполагает агрегацию данных из множества различных источников в целях построения системы выявления мошенничества, но подразумевает создание единственной монолитной онтологии.

По этой причине данный подход не может обладать преимуществами, которые предоставляет построение системы из нескольких онтологий различного уровня.

В работе "Credit Card Fraud Detection Based on Ontology Graph" [11] описывается подход к разработке системы выявления мошеннических транзакций по банковским картам с применением онтологических графов.

В рамках данного подхода используется онтология, содержащая данные о шаблонах мошеннических транзакций. Для каждой поступающей на вход системы транзакции строится ее онтологический граф, на основе которого выбираются похожие шаблоны мошеннических транзакций для запуска на них алгоритма.

Алгоритм определяет расстояние между входной транзакцией и шаблоном мошеннической транзакции, если это расстояние оказывается больше порогового, то входная транзакция определяется как мошенническая.

В рамках данного подхода онтология предназначена для решения конкретной задачи путем реализации конкретного алгоритма (вычисление близости конкретной транзакции к шаблону на графе).

По этой причине архитектура не обладает модульностью, позволяющей использовать уже имеющиеся данные для решения тех же задач иными методами или же для решения иных задач.

6. Прототип

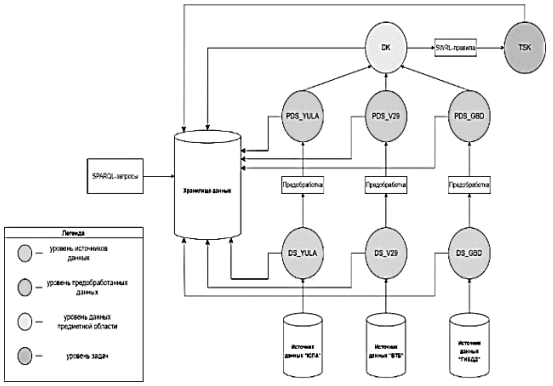

Рис. 3. *Потоки данных между различными уровнями*

Для реализации, описанной скоринговой системы был разработан ряд онтологий (рис. 3) в соответствии с представленной концепцией.

Далее приводится схема потоков данных между различными уровнями.

7. Пример логического вывода

Далее будет рассмотрен пример логического вывода (рис. 4), осуществляемого разработанным прототипом.

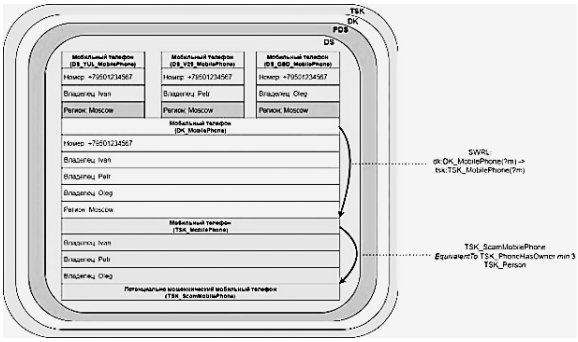

Рис. 4. *Пример логического вывода*

На уровне источников данных имеется три онтологии, описывающих соответственно три различных источника, в которых содержатся данные об одном и том же номере мобильного телефона, но при этом имеются противоречивые данные о владельце этого номера. На уровне предобработанных данных у класса "мобильный телефон" появляется дополнительный атрибут – "регион". Данные о принадлежности номера к региону отсутствуют в источниках в явном виде, но могут быть получены путем обработки номера.

Далее за счет механизма наследования (owl:SubClassOf) данные из всех источников автоматически относятся к соответствующему классу из онтологии уровня предметной области. Благодаря SWRL-правилу ("dk:DK_MobilePhone(?m) -> tsk:TSK_MobilePhone(?m) ") данные отображаются на уровень задач. За счет аксиомы класса ("TSK_ScamMobilePhone EquivalentTo TSK_PhoneHasOwner min 3 TSK_Person") проверяется условие того, что у телефона не менее трех владельцев, и делается логический вывод о том, что номер должен быть отнесен к классу "потенциально мошеннический мобильный телефон".

8. Преимущества подхода

Благодаря тому, что экземпляр класса уровня источников данных автоматически относится к соответствующему классу уровня предметной области, отсутствует дублирование информационных сущностей на уровне предметной области.

В примере, продемонстрированном на рис. 4, это выражается в том, что три различных класса уровня источников отображаются в единый класс уровня предметной области. Благодаря тому, что экземпляр класса уровня источников данных автоматически относится к соответствующему классу уровня предметной области, появляется возможность проследить, из каких источников получены те или иные данные. В примере видно, из какого источника получены данные о том или ином владельце.

Наличие уровня задач позволяет представлять исходные данные в новом контексте и делает их более понятными с точки зрения решаемой задачи.

В примере на уровне задач вводится новое понятие "потенциально мошеннический мобильный телефон", которое отсутствует на всех остальных уровнях и позволяет сделать более простым и наглядным решение поставленной задачи (в данном случае – оценки риска мошенничества).

Описанный подход позволяет обеспечить определенную степень независимости уровней задач и предметной области.

Благодаря этому возникает возможность решать новые задачи на уже имеющихся данных, при этом изменяя только этот уровень и не внося изменений на остальных уровнях.

В приведенном примере может быть разработана иная онтология уровня задач, позволяющая, к примеру, распределять владельцев телефонов по регионам. При этом не будет необходимости в том, чтобы вносить изменения на остальных уровнях.

Заключение

В результате проведенной работы была сформулирована концепция построения интегрированной модели данных, были выделены ее преимущества, а также создан прототип, демонстрирующий способность модели к логическому выводу.

Развитием данной модели может быть выделение новых уровней или изменение характера связей между существующими. Это может привести к появлению новых преимуществ модели или же может позволить подстроить модель под какие-либо дополнительные ограничения.

Иным вариантом развития представленной работы может стать исследование возможности интеграции онтологий с подключаемыми средствами автоматической обработки данных (скриптами, сервисами и т.п.) для возможности описания не только структуры данных или знаний, но и процедур их обработки (в т. ч. аналитической). Это связано с тем, что над созданием подобных аналитических систем работают не только специалисты в программировании, но также аналитики и эксперты.

В данном случае существует потребность в минимизации требований к знанию программирования со стороны подобных специалистов.

Список литературы

1. *Лапшин В.А.* Онтологии в информационных системах. Современный подход. М., 2009.
2. *Спецификация* языка OWL 2. URL: https://www.w3.org/TR/owl2-overview/.
3. *Кара-Ушанов В.Ю.* Реляционная модель данных. Екатеринбург: Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, 2017.
4. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб.: Питер, 2001. 384 с.: ил.
5. *Herald Kllapi, Dimitris Bilidas, Ian Horrocks, Yannis Ioannidis, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Manolis Koubarakis, Dmitriy Zheleznyakov.* Distributed Query Processing on the Cloud: the Optique Point of View [электронный ресурс]: https://www.researchgate.net/profile/Herald-Kllapi/publication/270960545_Distributed_Query_Processng_on_the_Cloud_the_Optique_Point_of_View_Short_Paper/links/54bb00d80cf253b50e2d07a2/Distributed-Query-Processing-on-the-Cloud-the-Optique-Point-of-View-Short-Paper.pdf.
6. *Mariano Rodrıguez-Muro, Roman Kontchakov, Michael Zakharyaschev*. Ontology-Based Data Access: Ontop of Databases. ISWC 2013: The Semantic Web – ISWC (2013) 558–573.
7. *Birte Glimm, Ian Horrocks, Carsten Lutz, Uli Sattler.* Conjunctive Query Answering for the Description Logic SHIQ. IJCAI-07 399-404.
8. *Matthew* Horridge Practical Guide To Building OWL Ontologies Using Prot´eg´e 4 and COODE Tools. The University Of Manchester, 2011.
9. *Спецификация* языка Semantic Web Rule Language. URL: https://www.w3.org/Submission/SWRL/.
10. *Girija Attigeri, Manohara Pai M. M., Radhika M. Pai, Rahul Kulkarni.* Knowledge Base Ontology Building For Fraud Detection Using Topic Modeling. Procedia Computer Science. №135 (2018). P. 369–376.
11. *Ali Ahmadian Ramaki, Reza Asgari, Reza Ebrahimi Atani*. Credit Card Fraud Detection Based on Ontology Graph // International Journal of Security, Privacy and Trust Management (IJSPTM). Vol. 1, № 5. October, 2012.

References

1. *Lapshin V.A.* Ontologii v informacionnyh sistemah. Sovremennyj podhod. M., 2009.
2. *OWL 2* Web Ontology Language Document Overview (Second Edition): https://www.w3.org/TR/owl2-overview/.
3. *Kara-Ushanov V.YU.* Relyacionnaya model' dan-nyh. Ekaterinburg: Ural'skij federal'nyj universitet imeni pervogo Prezidenta Rossii B.N. El'cina, 2017.
4. *Gavrilova T.A., Horoshevskij V.F.* Bazy znanij intellektual'nyh sistem. SPb.: Piter, 2001. 384 s.: il
5. *Herald Kllapi, Dimitris Bilidas, Ian Horrocks, Yannis Ioannidis, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Manolis Koubarakis, Dmitriy Zheleznyakov*. Distributed Query Processing on the Cloud: the Optique Point of View. URL: https://www.researchgate.net/profile/Herald-Kllapi/publication/270960545_. Distributed_Query_Processing_on_the_ Cloud_the_Optique_Point_of_View_Short_Pa

per/links/54bb00d80cf253b50e2d07a2/Distributed-Query-Processing-on-the-Cloud-the-Optique-Point-of-View-Short-Paper.pdf.
6. *Mariano Rodrıguez-Muro, Roman Kontchakov, Michael Zakharyaschev*. Ontology-Based Data Access: Ontop of Databases. ISWC 2013: The Semantic Web – ISWC (2013) 558–573
7. *Birte Glimm, Ian Horrocks, Carsten Lutz, Uli Sattler*. Conjunctive Query Answering for the Description Logic SHIQ. IJCAI-07 399-404.
8. *Matthew Horridge.* Practical Guide To Building OWL Ontologies Using Prot´eg´e 4 andCOODE Tools. The University Of Manchester, 2011.
9. *SWRL* A Semantic Web Rule Language Combining OWL and Rule. ML W3C Member Submission 21 May 2004. URL: https://www.w3.org/Submission/SWRL/.
10. *Girija Attigeri, Manohara Pai M. M., Radhika M. Pai, Rahul Kulkarni*. Knowledge Base Ontology Build-ing For Fraud Detection Using Topic Modeling. Procedia Computer Science. № 135 (2018). P. 369–376.
11. *Ali Ahmadian Ramaki, Reza Asgari, Reza Ebrahimi Atani*. Credit Card Fraud Detection Based on Ontology Graph // International Journal of Security, Privacy and Trust Management (IJSPTM)**.** Vol. 1, № 5. October**,** 2012**.**

Просьба ссылаться на эту статью:

Чернов П.К., Рабчевский Е.А. Создание интегрированной модели данных из разнородных источников, содержащих цифровые следы // Вестник Пермского университета. Математика. Механика. Информатика. 2022. Вып. 2(57). С. 81–87. DOI: 10.17072/1993-0550-2022-2-81-87.

Please cite this article as:

Chernov P.K., Rabchevskiy E.A. Creation of an Integrated Data Model of the Heterogeneous Sources Containing Digital Footprints // Bulletin of Perm University. Mathematics. Mechanics. Computer Science. 2022. Issue 2(57). P. 81–87. DOI: 10.17072/1993-0550-2022-2-81-87.