УДК 004.822

**В. С. Заякин**[1,2], студент магистратуры[1], инженер данных[2],
e-mail: vszayakin@yandex.ru
**Л. Н. Лядова**[1], канд. физ.-мат. наук, доц., e-mail: lnlyadova@gmail.com
**Е. А. Рабчевский**[2], генеральный директор, e-mail: e.rabchevskiy@seuslab.ru
[1] Национальный исследовательский университет
«Высшая школа экономики», Пермский филиал, Пермь, Россия
[2] ООО «СЕУСЛАБ», Пермь, Россия

# Онтологический подход к интеграции информации в областях с интенсивным использованием данных

*Обсуждаются проблемы интеграции информации в областях с интенсивным использованием данных. Рассмотрены и проанализированы существующие подходы к интеграции. Предложен подход к интеграции, основанный на онтологической и продукционной моделях знаний, а также формальная теоретико-множественная модель, лежащая в его основе. Практические преимущества подхода продемонстрированы на примере концептуализации задачи выявления ключевых мостов из области анализа социальных сетей.*

***Ключевые слова:*** *интеграция, онтологии, базы знаний, базы данных, жизнеспособность систем, аналитические платформы, открытые данные, анализ данных, социальные сети, слабоструктурированные объекты.*

## Введение

Разработка и эксплуатация систем баз знаний в сфере анализа социальных сетей и медиа неизбежно сопровождается проблемами поддержки их *жизнеспособности* [1, 2]. Логические модели баз знаний (БЗ) и баз данных (БД), используемые интеллектуальной системой для моделирования анализируемых объектов (пользователей, публикаций и т. д.) и процессов (например, распространения деструктивного контента), а также для описания способов их физического представления, постоянно расширяются, модифицируются. В большой степени это обусловлено свойствами объектов и процессов, подвергаемых анализу, а также интенсивным использованием (сбором, обработкой, порождением и т. п.) данных в подобных областях.

Во-первых, в процессе эксплуатации аналитической системы *постоянно появляются новые источники данных*, доступные для сбора и аналитической обработки (новые онлайновые социальные сети, публикуемые базы данных и пр.). С одной стороны, за счёт интеграции данных существует возможность извлечения ранее недоступных фактов об анализируемых объектах. С другой, неполнота данных, логическая противоречивость и т. п. создают препятствия к логически согласованной интеграции новых источников с существующими массивами данных. Это в свою очередь создает необходимость неоднократной и по возможности быстрой адаптации (расширения, изменения, обобщения) существующих информационных моделей с учетом особенностей сематического контекста, возникающих связей с другими объектами, а также особенностей их физического хранения в БД.

Во-вторых, *извлечение знаний в предметных областях с интенсивным использованием открытых данных зачастую сопровождается применением интеллектуальных методов* автоматизированной обработки данных, собираемых и агрегируемых из множества источников, например, социальных сетей, новостных агрегаторов и т. п. В частности, для анализа социальных сетей применяются методы риск-анализа [3], эвристические методики [4], алгоритмы машинного обучения [5, 6]. При этом так же, как и источники данных, с течением времени для экспертов и аналитиков становятся доступными новые методы анализа, которые разрабатываются или осваиваются ими. Таким образом, появляется возможность применения новых методов к существующим данных, и наоборот, применения существующих методов к новым данным. В итоге естественным образом возникает потребность в интеграции предметных данных с результатами анализа, формулируемых как новые факты, для которых, как и для исходных данных, необходимо разрабатывать соответствующие модели.

В-третьих, *динамичность слабоструктурированных объектов, подвергаемых анализу, обуславливает постоянное изменение моделей*

*информационного представления этих объектов*. На практике это проявляется при интерпретации результатов анализа собираемых данных, т. е. в процессе концептуализации выявленных знаний. Коллективное представление экспертов и аналитиков о свойствах объекта и его связей с другими объектами предметной области может меняться как вследствие получения информации за счёт доступа к новым источникам и методам анализа данных, так и за счёт приобретения нового, порой субъективного аналитического или управленческого опыта. Кроме того, информация об одних и тех же объектах может трактоваться по-разному в зависимости от контекста. Следовательно, использование богатых с точки зрения выразительности языковых средств для описания независимых интерпретаций результатов анализа данных с учетом изменяемости их моделей является важным фактором поддержки актуальности, доступности и прозрачности информации для конечных её потребителей [7].

Таким образом, актуально создание средств для формирования расширяемых БЗ, ориентированных на использование открытых данных для автоматизированного получения новых фактов и их интерпретации согласно независимым моделям экспертных знаний в различных предметных областях. Для решения этой задачи данная работа предлагает концептуальное описание подхода к интеграции информации на основе онтологий, а также формальную модель процесса формирования онтологических баз знаний, лежащую в его основе. Предложенный подход помимо обеспечения расширяемости и независимости моделей интегрируемых источников учитывает их изменяемость с течением времени при эксплуатации интеллектуальной системы БЗ.

## 1. Анализ существующих подходов к интеграции информации

На основе вышеописанных проблем интеграции информации можно предъявить следующие требования к подходам и средствам интеграции:

1. *Расширяемость* моделей при включении описаний новых источников информации таких, как источники данных, модели предметных областей и задач с сохранением логической целостности интегрированной модели.

2. Возможность *независимой интерпретации* данных и результатов их обработки в виде закономерностей, описываемых в терминах различных предметных областей.

3. Обеспечение *прослеживаемости изменений* в метамоделях с целью поддержки актуальности семантического аннотирования хранимой информации.

Подавляющее большинство существующих подходов к интеграции информации использует в своей основе онтологии. Эти подходы условно можно разделить на мультимодельные и одномодельные [8]. Первые основаны на разделении онтологических описаний интегрируемых ресурсов, а также описаний отображений или трансформаций между онтологиями и источниками информации. В одномодельных подходах, напротив, для интеграции используется одна онтология, описывающая всю информацию в виде глобальной централизованной связной сети понятий.

В *мультимодельных* подходах онтологии зачастую выступают в роли независимых метамоделей отдельных информационных источников. В работах [9, 10] совокупность таких онтологий обозначена как многоаспектная онтология. Она предназначена для *независимого моделирования и объединения различных аспектов* интеллектуальной системы, а именно структур данных (БД, журналов событий, текстов), предметных областей, задач обработки данных, предметно-ориентированных языковых средств и т. п. Такой подход также обеспечивает простое *расширение* модели интеграции, позволяя встроить новую онтологию, описав аксиоматические отношения между её элементами и элементами ранее сформированных онтологий.

Подобный подход реализован в [11] на примере геоинформационной системы органов исполнительной власти описывается интеграция нескольких пространственных БД с помощью онтологий метаданных, а также пространственных и атрибутивных данных, описываемых отдельно для каждой интегрируемой БД, что обеспечивает горизонтальную *расширяемость* модели

при подключении нового источника. Непосредственно же интеграция в представленном подходе задаётся связью элементов этих онтологий с элементами общей онтологии предметной области, которая является *независимой от локальных моделей* интегрируемых БД.

Похожим образом в работах [12, 13, 14] описываются методы интеграции реляционных БД, отличающиеся способами доступа к интегрированным данным. Именно, авторы [12] транслируют SPARQL-запрос в набор SQL-запросов к интегрированным БД с использованием парсинга входного запроса с помощью библиотеки Jena, а в [13] авторы применяют разработанное расширение для языка SPARQL, позволяющий изменить режим функционирования обработки запроса на стороне SPARQL-сервера. В [14] метод трансформации используется для интеллектуализации запросов к реляционным БД. Для этого запросы пользователей на естественном языке с помощью синтаксической и семантической предобработки преобразуются в SQL-запросы. Онтологии при таком подходе описывают терминологию, которую могут использовать пользователи при формулировании запросов. Описанные способы интеграции позволяют обращаться к табличным данным *независимо от схем БД*, аннотируя извлекаемые данные с использованием онтологии, которая описывается независимо от используемых в системе БД.

*Одномодельные* подходы в основном используются для автоматизированного построения общей онтологии предметной области на основе анализа предметных данных в совокупности интегрируемых баз или датасетов. В [15] описывается технология вывода онтологии с применением методов анализа формальных понятий (formal concept analysis). Разработанные средства требуют экспертной (ручной) идентификации и именования классов объектов (в англоязычной литературе используется термин *концепт*). В [16] описан прототип системы для генерации метаданных в виде общей онтологии, построенной по набору XSL-, XML- и RDF-документов. Оба подхода могут быть полезны для быстрого формулирования первичных выводов о структуре

информации в источниках по набору относительно небольших фрагментов данных. Однако в случае изменений в структурах первичных источников, целевая онтология должна быть перестроена *без возможности отследить изменения* и представить их в наглядной форме.

Существуют также примеры проектов, успешно сочетающих мультимодельный и одномодельный подходы. Так, в компании General Electric разработаны средства доступа к информации, интегрируемой из БД, находящихся под управлением разных СУБД (реляционные, NoSQL, HDFS и пр.). В работе [17] описывается архитектура подобной системы, основанной на описании глобального графа знаний (многоаспектной онтологии), содержащего всю информацию, необходимую для доступа к различным платформам хранения данных. Компонент NodeGroup данной системы используется для создания метаописания структуры данных, извлекаемых из интегрированных баз. В статье [18] описано его применение для формирования выборок при разработке моделей машинного обучения. Каждой структуре входных и выходных данных соответствует онтология, которая интегрируется, как с глобальным графом знаний, так и с программными компонентами, реализующими модели машинного обучения.

Таким образом, в этих проектах реализуется *возможность интеграции данных с результатами их обработки* на уровне метаданных. При этом на основе каждой онтологии может быть сгенерирован параметризуемый SPARQL-запрос, который при выполнении в системе так же, как и в [12], транслируется в набор подзапросов на языках, используемых в конкретных платформах хранения данных, обеспечивая *независимый от локальных схем БД доступ к данным*.

Похожие принципы реализуются в системе управления большими данными Data Civilizer [19], предназначенной для поиска, очистки и майнинга данных в распределённых хранилищах, интегрированных через общий граф связей. Процедуры поиска и обработки данных в системе могут быть

выстроены для выполнения в разном порядке. Стоит, впрочем, заметить, что ни результаты обработки данных, ни информация о процедурах обработки *не интегрированы с исходными источниками*. Это влечёт за собой *проблемы прослеживаемости данных* в случае изменения структуры метаданных, что, как упомянуто выше, неизбежно в случае анализа слабоструктурированных объектов.

Заметим, что большинство существующих подходов обеспечивают простоту расширения логической модели интегрируемых БД и БЗ, а также независимость описания отдельных информационных источников. Однако невозможность учёта изменений моделей источников в условиях интенсивного использования динамически изменяемых данных большого объёма является существенным ограничением. В частности, это может создавать необходимость повторного аннотирования данных или рефакторинга сервисов обработки запросов от пользователей с привлечением соответствующих специалистов: программных инженеров, инженеров знаний, администраторов БД и т. д.

Представленная в работе концепция интеграции нацелена на восполнение этого пробела за счёт однообразия интеграции как различных моделей источников, так и разных версий моделей одного и того же источника.

## 2. Концептуальное описание онтологического подхода к интеграции

В основе разрабатываемого подхода к интеграции лежат процессы применения методов автоматизированного анализа данных (например, методы Data Mining) и экспертной интерпретации результатов на основе продукционных правил (рис. 1).

Основная идея подхода заключается в независимом онтологическом моделировании различных информационных источников (с использованием соответствующей терминологии) и сопоставлении элементов интегрируемых онтологий (классов, отношений, атрибутов и т. д.) с использованием продукционных правил логического вывода. Условно данный подход можно назвать *интерпретирующим*. Предметные данные, аннотированные в

соответствии с классами и свойствами онтологии источника данных, при использовании для решения аналитических задач интерпретируются в терминах онтологии, моделирующей концептуальную схему этой задачи, согласно правилам логического вывода. Аналогично результаты анализа, аннотированные классами и свойствами онтологии задачи, интерпретируются в контексте онтологии предметной области с помощью, правил, моделирующих определённые экспертами закономерности.

Процесс анализа и интерпретации данных состоит из следующих этапов:

1. *Аналитик данных*, исходя из решаемой аналитической задачи определяет метод её решения.

2. *Программный инженер* на основе выбранного метода анализа предметных данных реализует соответствующие алгоритмы анализа.

3. *Эксперт* в зависимости от возможных результатов (исходов) применения методов анализа совместно с *инженером знаний* формулируют правила интерпретации результатов на основе продукционной модели.

4. После интерпретации результатов с помощью составленных правил *эксперт* сформулирует новые знания о предметной области.

5. На основе извлеченных знаний соответствующие модели могут быть уточнены описанием новых закономерностей *инженером знаний*.

При этом БЗ можно структурно разделить по виду знаний (рис. 2), моделируемых онтологиями:

1. Знания о данных и их источниках. Включают информацию о типах данных, форматах хранения данных в источнике, ограничения на арность и значения атрибутов и т. д.

2. Знания о задачах обработки данных, использующих предметные данные для вывода новых фактов и знаний. Включают информацию о структуре входных и выходных данных для процедур обработки данных (в т. ч. алгоритмов интеллектуального анализа), ссылки на внешне выполняемые скрипты, последовательность выполнения процедур и т. д.

3. Знания о предметных областях. Могут включать классы и свойства объектов предметной области, а также аксиоматические утверждения, моделирующие ограничения предметной области и субъективный опыт экспертов.

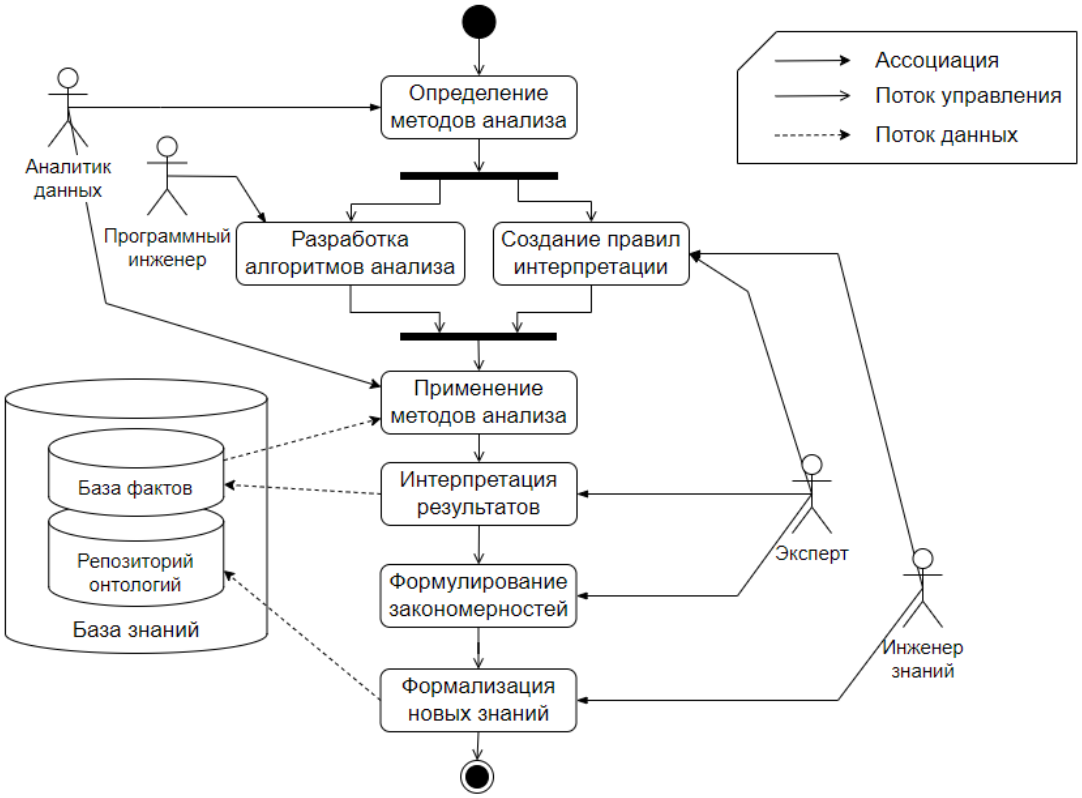


Рис. 1. Модель процесса анализа и интерпретации данных

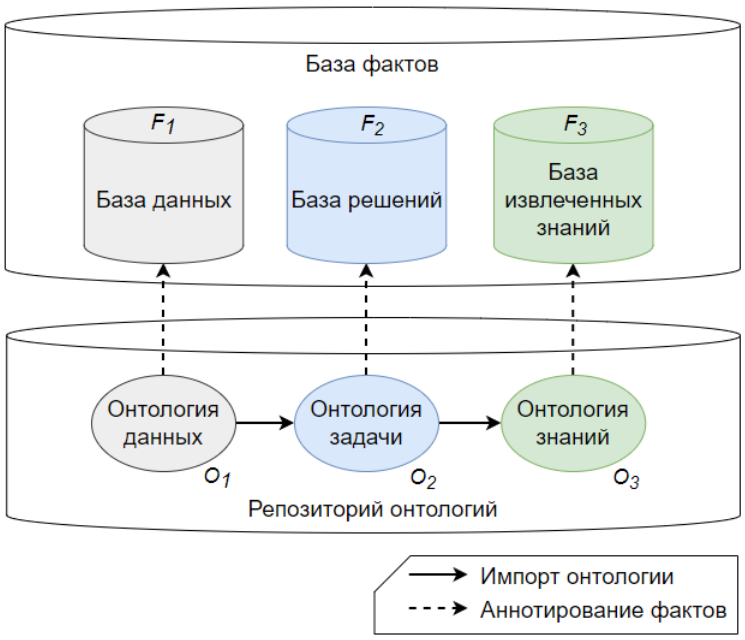


Рис. 2. Структура базы знаний

На практике предложенный подход позволяет избежать дублирования данных с использованием разных онтологий. Их интерпретация определяется в соответствии с контекстом, определяемым конкретной онтологией. При этом можно отследить, какие факты и на каком этапе обработки появилась. Факты, извлечённые из источника (например, конкретной социальной сети) будут аннотированы онтологией, описывающей этот источник. Результат обработки предметных данных (например, факты выявления дубликатов) будет аннотирован одной из онтологий задач и т. д. Это повышает прозрачность данных и позволяет представить одну и ту же физически хранимую информацию разным участникам проекта или пользователям системы в соответствии со знакомой им терминологией.

## 3. Формальная модель формирования баз знаний

Определим формально онтологию как шестёрку $O = <C, R, P, D, A, U>$, где $C$ – множество классов (концептов), $R$ – множество отношений (объектных свойств), $P$ – множество атрибутов классов (свойств типов данных), $D$ – множество типов данных, $A$ – множество аксиом, $U$ – множество экземпляров (объектов классов).

Множество *отношений* $R$ определяется как множество бинарных отношений между классами

$$\forall R_1 \in R \; \exists C_1, C_2 : R_1 \subseteq C_1 \times C_2.$$

Множество *атрибутов* $P$ определяется как множество бинарных отношений между классами и типами данных

$$\forall P_1 \in P \; \exists C_1 \in C, D_1 \in D : P_1 \subseteq C_1 \times D_1.$$

Экземпляры *онтологий* $U$ определим как некоторое подмножество экземпляров, каждому из которых соответствует некоторый (хотя бы один) класс онтологии

$$U \subseteq \bigcup\nolimits_{C_i \in C} C_i, \; \forall a \in U \; \exists C_1 \in C : a \in C_1.$$

Множество *аксиом* $A$ определим как множество высказываний

относительно экземпляров онтологии трёх видов:

1. Экземпляр $a$ принадлежит классу $C_1$.

2. Экземпляры $a$ и $b$ состоят в отношении $R_1$.

3. Экземпляр $a$ имеет значение свойства $P_1$, равное $d$ типа данных $D_1$.

Для удобства дальнейших определений введем обозначения для данных видов аксиом в соответствии со следующими формальными определениями:

$$C_1(a) \Leftrightarrow (a \in U, C_1 \in C, a \in C_1),$$

$$R_1(a,b) \Leftrightarrow (a,b \in U, R_1 \in R, aR_1b, (\exists C_1, C_2 \in C : R_1 \subseteq C_1 \times C_2)),$$

$$P_1(a, D_1(d)) \Leftrightarrow (a \in U, d \in D_1, P_1 \in P, aP_1d, (\exists C_1 \in C, D_1 \in D : P_1 \subseteq C_1 \times D_1)).$$

Любой факт, аннотируемый с помощью онтологии $O$, представляет собой совокупность аксиом. При разработке онтологий обычно известно небольшое число экземпляров и фактов относительно них, которые могут быть заранее включены в онтологию. Основная часть фактов формулируется на основе предметных данных, которые поступают в базу фактов в процессе сбора, предобработки, анализа и интерпретации информации из различных источников.

Таким образом, множество $A$ составляет лишь небольшую часть из множества всех аксиом $A^*$, которые теоретически могут быть аннотированы с помощью онтологии $O$. *Базу фактов*, описываемую онтологией $O$, обозначим $F$ и определим формально как подмножество множества $A^*$. В итоге имеем

$$A^* = \bigcup_{C_i \in C} \{C_i(a) \mid a \in U\} + \bigcup_{R_j \in R} \{R_j(a,b) \mid a,b \in U\} +$$
$$+ \bigcup_{P_k \in P,\, D_l \in D} \{P_k(a, D_l(d)) \mid a \in U, d \in D\},$$

$$A, F \subseteq A^*, |F| \square |A|.$$

Наконец, базу знаний $K$ можно формально определить как двойку

$$K = <O, F>.$$

Совокупность продукционных правил, задающих связь элементов онтологии $O_1$ с элементами онтологии $O_2$ для определения интерпретации данных при изменении контекста предметной области, назовём

*онтологическим отображением* $f$ и будем обозначать $f : O_1 \rightsquigarrow O_2$. На его основе определим соответствие, которое бы обозначало некоторую формальную процедуру интерпретации аксиом из терминов одной онтологии в термины другой онтологии. Назовём такое отображение *интерпретирующим* и будем обозначать $I(f)$.

Пусть $q$ – *запрос* к базе фактов, соответствующей онтологии $O$. Он определяет подмножество аксиом, описываемых онтологией $O$, которое извлекается из соответствующей базы фактов. Далее, пусть $p$ обозначает процедуру обработки предметных данных, которая на основе подмножества предметных данных (аксиом), описываемых некоторой онтологией, ставит в соответствие ему новое подмножество аксиом, описываемых той же онтологией. Наконец, пусть $O_1$, $O_2$, $O_3$ – онтологии, $f : O_1 \rightsquigarrow O_2$, $g : O_2 \rightsquigarrow O_3$ – онтологические отображения. Тогда определение процесса пополнения базы знаний новыми фактами при применении процедур обработки и интерпретации предметных данных можно формально определить в виде следующей композиции

$$[I(g)] \circ p \circ [I(f)] \circ q.$$

**Примечание.** В соответствии с нотацией, используемой в теории множеств, отображения последовательно применяются справа налево. Здесь $q$ представляет собой процедуру формирования входных данных, структура которых описывается онтологией $O_1$. Далее интерпретирующее отображение $I(f)$ используется для структурирования этих данных в соответствии с онтологией $O_2$, которая декларативно описывает задачу обработки этих данных с использованием процедуры $p$. Наконец, интерпретирующее отображение $I(g)$ представляет собой интерпретацию результатов применения процедуры обработки данных для структурирования в соответствии с терминами онтологии $O_3$.

Описанная модель позволяет определить подход к обеспечению прослеживаемости изменений. Именно, между разными версиями онтологий с

использованием продукционных правил можно определить соответствующие онтологические отображения $F_{ij} : O^i \rightsquigarrow O^j$, где $i$, $j$ – версии онтологий. На рисунке 3 представлена схема потоков данных в базе знаний и связей между отображениями, которые интегрируют информацию в виде онтологий.

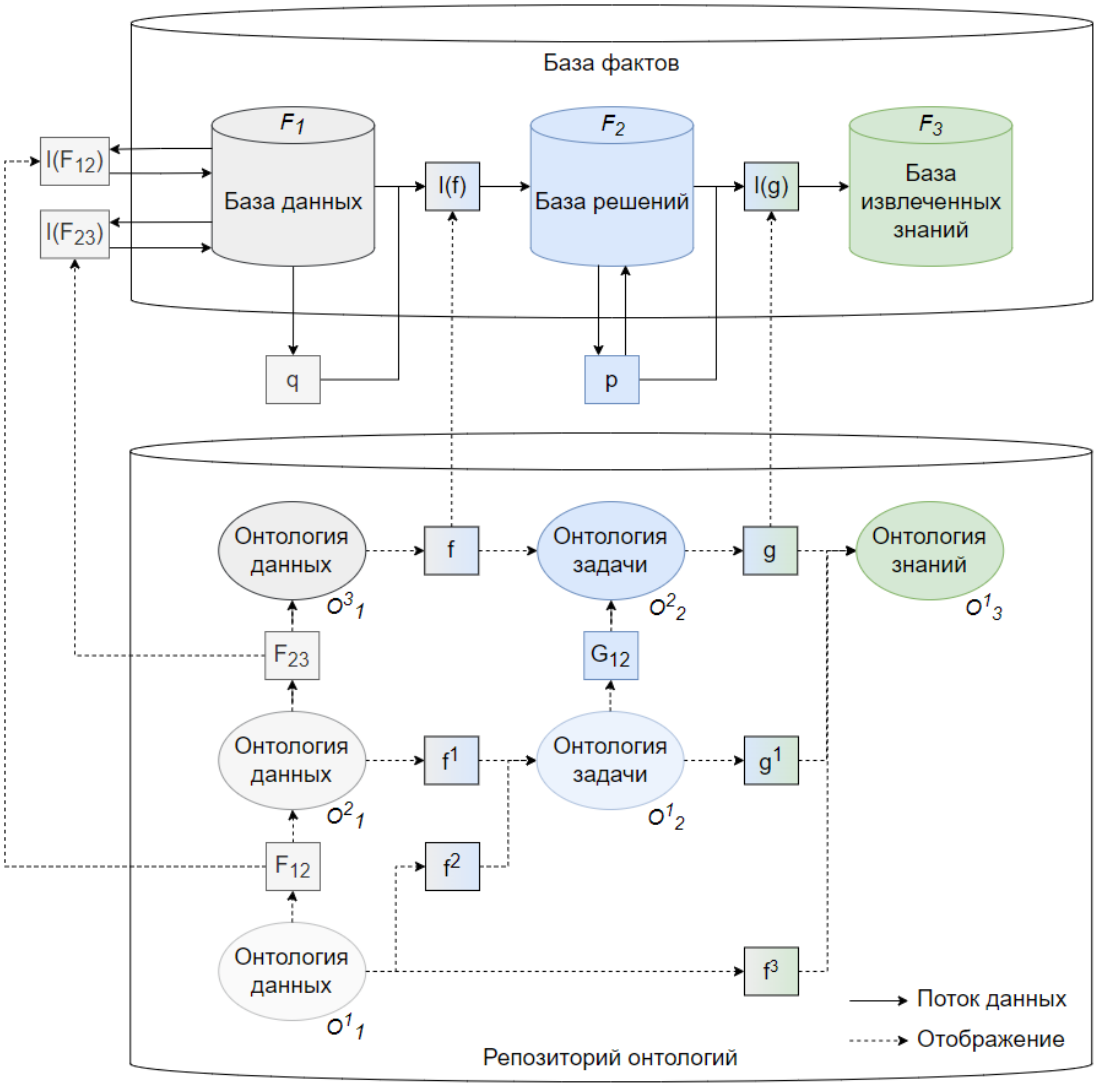


Рис. 3. Схема потоков данных и связей между отображениями в базе знаний

## 4. Демонстрация подхода на примере задачи выявления ключевых мостов

В рамках разработки аналитических платформ, предназначенных для анализа социальных сетей, зачастую требуется решение задачи определения ключевых пользователей. Продемонстрируем описанный подход к интеграции с использованием онтологий на примере частного случая этой задачи – определения мостов, методика решения которой подробно описана в [20] и для

чего может быть использована соответствующая программная реализация [21]. Для описания онтологий используется язык OWL2, а для их интеграции – механизмы импорта и продукционные правила на языке SWRL, который позволяет описать совокупность правил, представляющих онтологическое отображение, напрямую в файле онтологии.

**Постановка задачи.** Задача определения ключевых мостов заключается в анализе топологических свойств графа, построенного на основе связей между пользователями (дружба, подписки). Технически задача сводится к вычислению значения специально разработанной для этого метрики *центральности по взвешенному вкладу*, которая помимо структуры анализируемого графа учитывает публикационную активность пользователей, входящих в граф в качестве узлов. Уровень публикационной активности (*рейтинг* узла) определяется числом публикаций пользователя, относящихся к определённой тематике. Тематика может быть определена по набору ключевых слов, содержащихся в тексте публикации.

В данной предметной области *мостом считается любой узел, значение центральности которого отлично от нуля*. При превышении установленного порогового значения $t$ метрики мост считается одним из ключевых участников процессов распространения информации в социальных сетях и обладает высоким уровнем информационного влияния. Отличительная особенность данного типа пользователей заключается в том, что в процессе распространения информации мосты подключают к этому процессу *кластеры* пользователей, которые проявляют публикационную активность и связаны в рассматриваемой части сети только с этим мостом (см. рис. 4). При этом можно считать, что *пользователь входит в некоторый кластер тогда и только тогда, когда значение упомянутой метрики центральности равно нулю*.

Данные для построения графов и определения рейтинга его узлов могут собираться из нескольких социальных сетей, описываемых разными онтологиями. При этом пользователи могут представлять собой одни и те же

узлы. Для простоты будем считать, что узлы совпадают при совпадении номеров телефонов, указанных в профилях пользователей.

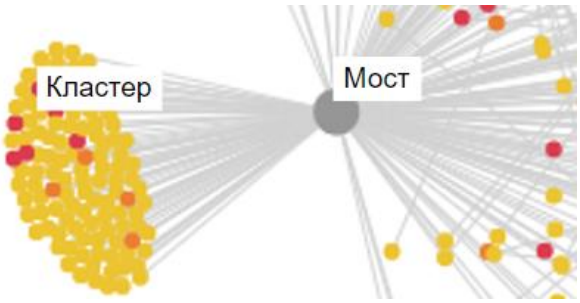


Рис. 4. Пример моста и соответствующего ему кластера

В итоге фрагменты онтологий интегрируемой информации в соответствии с описанным подходом могут быть сопоставлены с использованием продукционных правил, описывающих онтологические отображения, и их реализаций на языке SWRL (см. табл.). Графическая схема интеграции онтологий представлена на рисунке 5.

Предположим, что для анализа используются данные социальных сетей, в частности ВКонтакте и Facebook, которые описываются отдельными онтологиями (vk и fb). Из ВКонтакте собираются данные о связях дружбы между пользователями, факты совершения различных видов публикаций (текстовые посты, видеозаписи и т. п.) по исследуемой тематике и некоторую информацию из профиля пользователя. Из Facebook – факты о подписке пользователей друг на друга и информацию из профиля. Очевидно, что для решения указанной задачи могут быть использованы данные из прочих социальных сетей, что демонстрирует расширяемость подхода. Для включения нового источника в модель БЗ необходимо с помощью онтологических отображений связать соответствующую онтологию данных с онтологией задачи (task), описывающей концептуальную схему решаемой задачи.

При этом результаты решения задачи могут быть интерпретированы по-разному. Для одних экспертов может быть важным, является ли пользователь мостом (онтология bridges), для других – какие пользователи входят в кластеры в пределах рассматриваемой сети (онтология clusters). В зависимости от

нужного контекста указанные точки зрения могут быть описаны независимо друг от друга разными онтологиями и соответствующими правилами онтологического отображения, обеспечивая независимость экспертной интерпретации данных и результатов их анализа.

Таблица. Перечень правил онтологических отображений

| **Формулировка правила** | **Правила на языке SWRL** |
|---|---|
| Любой *Пользователь* является *Узлом графа* (*fb* ⇝ *task*; *vk* ⇝ *task*) | *fb*:Пользователь(?*x*) → *task*:Узел графа(?*x*)<br>*vk*:Пользователь(?*x*) → *task*:Узел графа(?*x*) |
| Если *x подписан на y*, то *x связан с y* (*fb* ⇝ *task*) | *fb*:подписан на(?*x*,?*y*) → *task*:связан с(?*x*,?*y*) |
| Если *x состоит в списке друзей y*, то *x связан с y* (*vk* ⇝ *task*) | *vk*:состоит в списке друзей(?*x*,?*y*) → *task*:связан с(?*x*,?*y*) |
| Если *x* имеет значение *центральности*, отличное от нуля, то *x – Мост* (*task* ⇝ *bridges*) | *swrlb*:greaterThan(*task*:центральность(?*x*),0) → *bridges*:Мост(?*x*) |
| Если *x* имеет значение *центральности*, большее порогового значения *t*, то *x – Ключевой мост* (*task* ⇝ *bridges*), *t* – некоторое число от 0 до 1 | *swrlb*:greaterThan(*task*:центральность(?*x*),*t*) → *bridges*:Ключевой мост(?*x*) |
| Если *x* имеет значение *центральности*, равное нулю, то *x – Узел кластера* (*task* ⇝ *clusters*) | *swrlb*:equal(*task*:центральность(?*x*),0) → *clusters*:Узел кластера(?*x*) |
| Любой *Узел кластера* в онтологии *clusters* является *Узлом кластера* в онтологии *clusters2* (*clusters* ⇝ *clusters2*) | *clusters*:Узел кластера(?*x*) → *clusters2*:Узел кластера(?*x*) |
| Любой *Мост* в онтологии *bridges* является *Мостом* в онтологии *clusters2* (*bridges* ⇝ *clusters2*) | *bridges*:Мост(?*x*) → *clusters2*: Мост (?*x*) |
| Если *Мост x связан с Узлом кластера y*, то *x присоединяет* некоторый *Кластер z*, и *y входит в Кластер z* (*bridges*, *clusters* ⇝ *clusters2*) | *brisges*:Мост(?*x*) ^ *clusters*:Узел кластера(?*y*) ^ *task*:связан с(?*x*,?*y*) → *clusters2*:присоединяет(?*x*,?*z*) ^ *clusters2*:входит в(?*y*,?*z*) |
| Здесь *x*, *y*, *z* обозначают экземпляры онтологий, *swrlb*, *vk*, *fb* и т. п. обозначают XML-префиксы пространств имен языка SWRL и онтологий | |

Далее, предположим, что возникла потребность в модификации онтологии анализа кластеров *clusters* для отражения в ней связи кластеров пользователей с соответствующими им мостами. Для этого может быть определена новая версия данной онтологии *clusters2*, расширяющая модель и дополняющая её новыми понятиями (*Кластер*, *присоединять*, *входить в*). Подход на основе правил

позволяет без замены и разрыва связей между ранее определенными онтологиями переопределить (доопределить) интерпретацию существующих фактов. Этим обеспечивается *прослеживаемость изменений* без необходимости модификации предыдущих версий онтологий, которые помимо прочего могут использоваться программными компонентами системы, т. н. решателями.

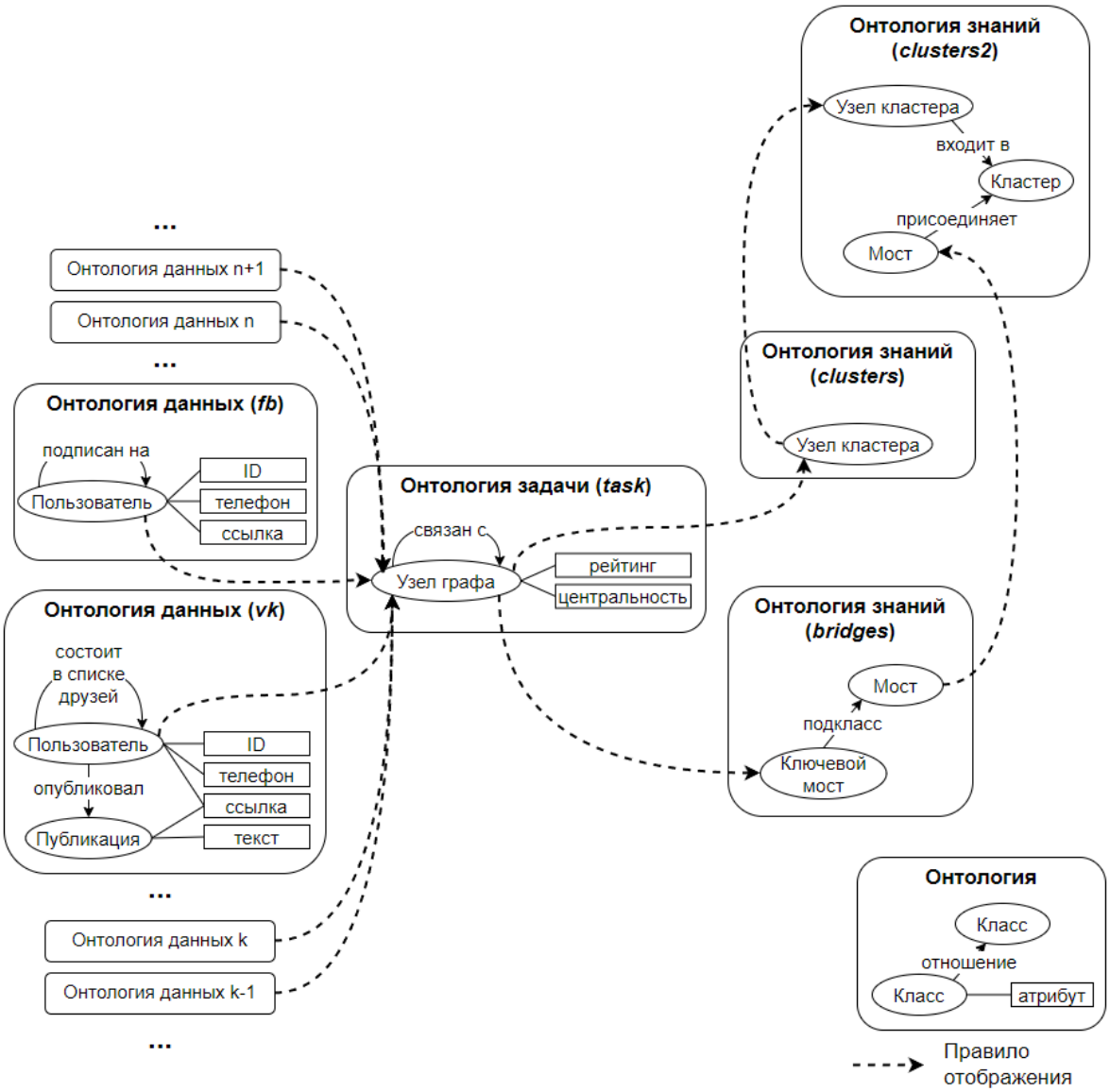


Рис. 5. Схема интеграции онтологий

## Заключение

Предложенный в работе метод интеграции информации целесообразно применять при разработке интеллектуальных систем баз знаний в областях с интенсивным использованием данных. Интеграция онтологий задач потенциально позволяет обеспечить систематизированный способ

комбинирования различных методов анализа предметных данных для извлечения знаний, что может быть эффективным средством формирования специализированных методик анализа данных. Комбинирование механизма продукционных правил с онтологиями при этом может обеспечить автоматизированную поддержку актуальности используемой терминологии для аннотирования информации. В совокупности с независимой интерпретацией результатов анализа данных правила являются средством интеллектуальной агрегации исходных данных для представления этой информации в виде, доступном для более простого восприятия экспертами предметных областей. Указанные особенности подхода обеспечивают его практическую значимость, выраженную в расширяемости модели БЗ, независимости моделей интегрируемых источников и прослеживаемости вносимых в модели изменений при эксплуатации интеллектуальной системы.

**Список литературы**


1. **Грибова В. В.** Создание жизнеспособных интеллектуальных систем с управляемыми декларативными компонентами / **В. В. Грибова, Ф. М. Москаленко, В. А. Тимченко, Е. А. Шалфеева** // Информационные и математические технологии в науке и управлении. – 2018. – №. 3 (11). – С. 6-17.
2. **Грибова В. В.** Методы и средства разработки жизнеспособных интеллектуальных сервисов / **В. В. Грибова, А. С. Клещев, Ф. М. Москаленко, В. А. Тимченко, Л. А. Федорищев, Е. А. Шалфеева** // Вестник Дальневосточного отделения Российской академии наук. – 2016. – №. 4 (188). – С. 133-141.
3. **Остапенко А. Г.** Моделирование целенаправленных атак социальных информационных сетей / **А. Г. Остапенко, Е. А. Шварцкопф, Е. С. Соколова** // Информация и безопасность. – 2015. – Т. 18. – №. 2. – С. 298-301.
4. **Проноза А. А.** Методика выявления каналов распространения информации в социальных сетях / **А. А. Проноза, Л. А. Виткова, А. А. Чечулин, И. В. Котенко, Д. В. Сахаров** // Вестник Санкт-Петербургского университета. Серия 10. Прикладная математика. Информатика. Процессы управления. – 2018. – №. 4. – С. 362-377.
5. **Malmi E.** You are What Apps You Use: Demographic Prediction Based on User's Apps / **E. Malmi, I. Weber** // Proceedings of the International AAAI Conference on Web and Social Media. – 2016. – Т. 10. – №. 1.
6. **Xiang L.** Demographic Attribute Inference from Social Multimedia Behaviors: a Cross-OSN Approach / **L. Xiang, J. Sang, C. Xu** // International Conference on Multimedia Modeling. – Springer, Cham, 2017. – С. 515-526.
7. **Yu-Cheng T.** Transparency in Software Engineering // A Thesis Submitted in Fulfillment of the Requirements of Doctor of Philosophy in Electrical and Electronic Engineering. The University of Auckland. New Zealand. – 2014. – 337 с.

8. **Alizadeh M.** Ontology Based Information Integration: a Survey / **M. Alizadeh, M. H. Shahrezaei, F. Tahernezhad-Javazm** // arXiv preprint arXiv:1909.13762. – 2019.
9. **Лядова Л. Н.** Формирование событийных рядов с использованием многоаспектных онтологий / **Л. Н. Лядова, В. С. Заякин, М. А. Смирнов** // Технологии разработки информационных систем ТРИС-2020. – 2020. – С. 297-303.
10. **Лядова Л. Н.** Архитектура DSM-платформы, основанной на знаниях / **Л. Н. Лядова, Н. М. Суворов, В. А. Василюк** // Технологии разработки информационных систем ТРИС-2020. – 2020. – С. 304-311.
11. **Павлов С. В.** Онтологическая модель интеграции разнородных по структуре и тематике пространственных баз данных в единую региональную базу данных / **С. В. Павлов, О. А. Ефремова** // Онтология проектирования. – 2017. – Т. 7. – №. 3 (25). – С. 323-333.
12. **Asfand-E-Yar M.** Semantic Integration of Heterogeneous Databases of Same Domain Using Ontology / **M. Asfand-E-Yar, R. Ali** // IEEE Access. – 2020. – Т. 8. – С. 77903-77919.
13. **Xiao G.** Efficient Ontology-Based Data Integration with Canonical IRIs / **G. Xiao, D. Hovland, D. Bilidas, M. Rezk, M. Giese, D. Calvanese** // European Semantic Web Conference. – Springer, Cham, 2018. – С. 697-713.
14. **Чуприна С. И.** Концепция обогащения унаследованных информационных систем сервисом запросов на естественном языке / **С. И. Чуприна, И. С. Постаногов** // Вестник Пермского университета. Математика. Механика. Информатика. – 2015. – №. 2. – С. 78-86.
15. **Fu G.** FCA Based Ontology Development for Data Integration // Information Processing & Management. – 2016. – Т. 52. – №. 5. – С. 765-782.
16. **Mansukhlal G. P.** A Novel Approach for Semantic Integration of Data Using Ontology / **G. P. Mansukhlal, C. Malathy, U. U. Pratheebha** // Indian Journal of Science and Technology. – 2016. – Т. 9. – №. 48. – С. 1-6.
17. **McHugh J.** Integrated Access to Big Data Polystores Through a Knowledge-Driven Framework / **J. McHugh, P. E. Cuddihy, J. W. Williams, K. S. Aggour, V. S. Kumar, V. Mulwad** // 2017 IEEE International Conference on Big Data (Big Data). – IEEE, 2017. – С. 1494-1503.
18. **Kumar V. S.** NodeGroup: A Knowledge-Driven Data Management Abstraction for Industrial Machine Learning / **V. S. Kumar, P. Cuddihy, K. S. Aggour** // Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning. – 2019. – С. 1-4.
19. **Deng D.** The Data Civilizer System / **D. Deng, R. C. Fernandez, Z. Abedjan, S. Wang, M. Stonebraker, A. Elmagarmid, I. F. Ilyas, S. Madden, M. Ouzzani, N. Tang** // Proceedings of CIDR. – 2017. – С. 1-12.
20. **Заякин В. С.** Выявление мостов в кластерных сетях и оценка уровня их информационного влияния / **В. С. Заякин, А. Н. Рабчевский, Е. А. Рабчевский** // Информационные системы и технологии. – 2021. – Т. 5. – №. 127. – С. 21-30.
21. **Рабчевский А. Н.** Программа вычисления мостов в кластерных сетях: свидетельство о государственной регистрации программ для ЭВМ № 2021616086 / **А. Н. Рабчевский, В. С. Заякин** // № 2021615157; заявл. 13.04.2021; опубл. 16.04.2021.