УДК 004.021

Астаев Кирилл Алексеевич
Магистр, 2 курс
Факультет «Институт информационных технологий»
Кафедра «Кафедра математического обеспечения и стандартизации информационных технологий»
МИРЭА – Российский технологический университет
Россия, г. Москва

ПОДХОДЫ К ФОРМАЛИЗАЦИИ ПОЛЬЗОВАТЕЛЬСКОЙ СЕМАНТИКИ В ЗАДАЧАХ ИЕРАРХИЧЕСКОГО АГРЕГИРОВАНИЯ РЕЛЯЦИОННЫХ ДАННЫХ

Аннотация: *В статье рассматривается проблема формализации пользовательской семантики в задачах иерархического агрегирования и классификации реляционных структур данных. Показано, что традиционное представление реляционных данных фиксирует преимущественно структурные и логические связи, тогда как пользовательская интерпретация предметной области требует дополнительного семантического слоя. Рассмотрены основные подходы к представлению семантики: онтологические модели, метаданные, семантические правила, концептуальные иерархии и признаки пользовательской релевантности.*

Ключевые слова: *реляционная модель данных; иерархическое агрегирование; классификация данных; онтология; семантическая модель.*

Annotation: *relational data model; hierarchical aggregation; data classification; ontology; semantic model.*

Key words: *The article addresses the challenge of formalizing user semantics in tasks involving hierarchical aggregation and classification of relational data structures. It demonstrates that traditional representations of relational data*

primarily capture structural and logical relationships, whereas the user's interpretation of the subject domain requires an additional semantic layer. The main approaches to representing semantics are examined, including ontological models, metadata, semantic rules, conceptual hierarchies, and features related to user relevance.

Современные информационные системы накапливают значительные массивы данных, представленных в виде таблиц, связей между сущностями, атрибутов и ограничений целостности. Реляционная модель данных остаётся одной из базовых форм организации структурированной информации, поскольку позволяет формально описывать отношения между объектами предметной области, обеспечивать целостность данных и выполнять запросы над множествами записей. Вместе с тем при решении аналитических задач оказывается недостаточным рассматривать реляционные данные только как совокупность таблиц и связей. Для пользователя или эксперта предметной области данные обладают не только структурой, но и смыслом, который определяется задачей анализа, профессиональным контекстом, терминологией предметной области и целями принятия решений.

Проблема особенно заметна в задачах иерархического агрегирования и классификации. Агрегирование предполагает переход от отдельных записей к более крупным группам, категориям и уровням обобщения. Классификация, в свою очередь, требует отнесения объектов к определённым классам на основании значимых признаков. Если признаки выбираются только из физической или логической структуры базы данных, возникает риск формального объединения объектов без учёта их содержательной близости. Одинаковые по структуре записи могут иметь различную интерпретацию для разных пользователей, а разные по отдельным значениям записи могут рассматриваться как близкие в рамках конкретного пользовательского сценария. Следовательно, для повышения содержательной адекватности агрегирования и классификации требуется формализация пользовательской семантики как самостоятельного слоя описания данных.

Реляционная модель, основывается на представлении данных в виде отношений и на отделении логического описания данных от особенностей их физического хранения [1]. Данный подход обеспечил высокий уровень строгости в описании данных, однако сама реляционная структура не всегда содержит явное описание того, каким образом пользователь понимает смысл атрибутов, значений и связей. Например, поле income может быть представлено как числовой атрибут, но в прикладной задаче оно может интерпретироваться как показатель платёжеспособности, принадлежности к сегменту, уровня риска или потребительского потенциала. Формально значение остаётся одним и тем же, однако семантическая нагрузка меняется в зависимости от цели анализа.

Концептуальное моделирование данных частично решает указанную проблему, поскольку позволяет описывать сущности, связи и ограничения на уровне предметной области. Так, модель «сущность – связь» была предложена как средство более содержательного представления объектов реального мира и отношений между ними [2]. Однако даже концептуальная модель не исчерпывает проблему пользовательской семантики. Она фиксирует согласованное проектное представление предметной области, но не всегда отражает изменяемые пользовательские критерии группировки, экспертные правила интерпретации, веса признаков и контекстные основания классификации. В этом смысле пользовательская семантика может рассматриваться как дополнительный уровень описания, находящийся между формальной схемой данных и аналитической процедурой обработки.

Под пользовательской семантикой в рамках настоящего исследования целесообразно понимать совокупность смысловых интерпретаций, правил, предпочтений, ограничений и категорий, с помощью которых пользователь или эксперт предметной области определяет значимость данных, устанавливает основания их группировки и задаёт критерии принадлежности объектов к классам. В отличие от системной семантики, связанной с типами данных, ключами, ограничениями целостности и логическими связями,

пользовательская семантика ориентирована на содержательный контекст использования данных. Она выражает не только то, как данные хранятся, но и то, как они должны пониматься в рамках конкретной аналитической или управленческой задачи.

Пользовательская семантика может проявляться в нескольких формах. Во-первых, она может быть представлена через терминологию предметной области, когда значения атрибутов соотносятся с профессиональными понятиями: «активный клиент», «проблемный объект», «высокий риск», «приоритетная заявка». Во-вторых, она может задаваться через экспертные правила, связывающие значения нескольких признаков с некоторой категорией. В-третьих, она может фиксироваться через иерархии понятий, например: город - регион - страна; товар - категория - товарная группа; операция - тип операции - класс операций. В-четвёртых, пользовательская семантика может выражаться через веса признаков, когда для конкретного пользователя одни характеристики являются определяющими, а другие имеют вспомогательное значение. Наконец, она может быть описана через онтологические отношения, позволяющие формализовать классы, свойства, связи и ограничения предметной области.

Иерархическое агрегирование реляционных данных представляет собой процесс построения уровней обобщения, при котором отдельные записи или объекты объединяются в более крупные смысловые единицы. В простейшем случае агрегирование может быть реализовано с помощью группировки по значениям атрибутов. Однако в задачах аналитической обработки этого недостаточно, поскольку уровни обобщения должны быть не только вычислимыми, но и интерпретируемыми. Например, группировка клиентов по региону является структурно простой операцией, но группировка по уровню лояльности, инвестиционной привлекательности или рисковому профилю требует дополнительной семантической интерпретации признаков.

В этом контексте иерархическое агрегирование можно рассматривать как переход от низкоуровневого представления данных к концептуальным

категориям, релевантным пользователю. На нижнем уровне находятся записи реляционных таблиц и значения атрибутов. На следующем уровне формируются объекты предметной области, получаемые в результате объединения связанных записей. Далее объекты могут агрегироваться в группы на основании признакового сходства, экспертных правил или принадлежности к определённым концептам. На верхнем уровне возникают классы и категории, используемые для принятия решений или последующей классификации. Таким образом, иерархия данных оказывается не только структурной, но и семантической.

Важное отличие семантически ориентированного агрегирования от обычной группировки состоит в том, что критерий объединения объектов определяется не только совпадением значений, но и смысловой близостью. Формализация пользовательской семантики необходима для того, чтобы сделать подобные интерпретации воспроизводимыми и пригодными для машинной обработки. Если пользовательская семантика остаётся неявной, агрегирование и классификация зависят от субъективных решений конкретного аналитика и плохо поддаются проверке. Если же она выражена в виде правил, метаданных, онтологических связей или весовых моделей, её можно включить в алгоритмическую обработку данных. Это позволяет обеспечить большую прозрачность процедуры классификации, повысить интерпретируемость результатов и связать аналитическую модель с терминологией предметной области.

Одним из наиболее простых способов формализации пользовательской семантики является использование расширенных метаданных. Метаданные могут описывать назначение атрибута, допустимые значения, единицы измерения, интерпретационные диапазоны, связь с бизнес-понятием или уровень значимости в конкретной задаче. Например, числовой атрибут возраста может быть дополнен диапазонами «молодой пользователь», «пользователь среднего возраста», «старшая возрастная группа». Атрибут частоты обращений может быть связан с категориями «низкая активность»,

«средняя активность», «высокая активность». В таком случае реляционные данные получают дополнительный слой смысловой разметки, пригодный для построения уровней агрегирования.

Другим способом является правило-ориентированное представление семантики. В этом случае пользовательские интерпретации задаются в виде условий и выводов. Подобные правила могут использоваться как для предварительного агрегирования, так и для присвоения классов. Их преимущество состоит в понятности для эксперта, а ограничение – в необходимости регулярной актуализации и согласования с фактическими изменениями данных.

Онтологический подход предоставляет более развитые средства формализации семантики. В онтологии предметная область описывается через классы, отношения, свойства, ограничения и экземпляры. Возможно рассматривать онтологию как спецификацию концептуализации, то есть как формальное описание понятий и отношений некоторой области знания [1]. Для задач агрегирования реляционных данных онтология может выполнять функцию связующего слоя между схемой базы данных и пользовательскими категориями. Атрибуты и таблицы соотносятся с понятиями предметной области, а отношения между понятиями позволяют строить более содержательные уровни обобщения.

Семантические технологии в целом направлены на то, чтобы сделать данные понятными не только человеку, но и программным средствам обработки. В работах, посвящённых семантической сети, подчёркивается значение формального представления смысла для автоматизированной обработки информации [3]. Для рассматриваемой задачи это означает, что пользовательская семантика должна быть представлена не в виде произвольных комментариев, а в виде структур, пригодных для вычисления: концептов, отношений, правил, ограничений, иерархий и весов. Только в этом случае она может быть включена в процедуру агрегирования и классификации наравне с традиционными структурными признаками.

С практической точки зрения формализацию пользовательской семантики можно представить как построение отображения между реляционной структурой и семантическим слоем. Пусть реляционная структура данных задаётся множеством отношений, атрибутов и ограничений. Семантический слой дополняет её множеством пользовательских понятий, правил интерпретации, иерархий обобщения и признаковых весов. Тогда аналитическая процедура работает не только с исходными таблицами, но и с результатом их смыслового преобразования. В общем виде такую модель можно представить следующим образом:

$$D = \langle R, A, K, M, S, H, C \rangle,$$

где R – множество реляционных отношений, A – множество атрибутов, K – ключи и ограничения целостности, M – метаданные, S – пользовательская семантика, H – иерархия агрегирования, C – множество классов или категорий. В данной схеме пользовательская семантика выступает не внешним пояснением к данным, а одним из компонентов модели, влияющим на построение иерархии и определение классов.

Включение пользовательской семантики в процедуру агрегирования может быть реализовано поэтапно. На первом этапе выполняется извлечение структурных признаков из реляционной базы данных. На втором этапе признаки сопоставляются с пользовательскими понятиями и метаданными. На третьем этапе применяются правила интерпретации, позволяющие преобразовать числовые и категориальные значения в смысловые признаки. На четвёртом этапе строится иерархия агрегирования, где уровни определяются не только структурой данных, но и пользовательскими основаниями обобщения. На пятом этапе агрегированные объекты используются для классификации.

Таблица 1.

Подходы к формализации пользовательской семантики в задачах агрегирования реляционных данных

Подход	Сущность подхода	Роль в агрегировании и классификации

Метаданные	Описание атрибутов, значений, единиц измерения, смысловых диапазонов	Уточняют интерпретацию признаков и позволяют формировать смысловые группы
Экспертные правила	Условия отнесения объектов к категориям	Задают явные критерии классификации и построения агрегатов
Иерархии понятий	Представление уровней обобщения предметной области	Позволяют переходить от частных значений к более общим категориям
Онтологии	Формальное описание классов, свойств и отношений	Связывают схему данных с концептами предметной области
Весовые модели признаков	Задание значимости признаков с точки зрения пользователя	Позволяют учитывать приоритеты пользователя при сравнении и классификации объектов

Представленные подходы не являются взаимоисключающими. Напротив, в прикладных системах они могут использоваться совместно. Метаданные обеспечивают базовое описание атрибутов, экспертные правила задают условия интерпретации, иерархии понятий формируют уровни обобщения, онтологии связывают элементы данных с концептами предметной области, а весовые модели позволяют учитывать различную значимость признаков. Комбинация этих средств обеспечивает более гибкое и содержательное агрегирование по сравнению с чисто структурным подходом.

Связь пользовательской семантики с классификацией проявляется в том, что классификация всегда предполагает выбор существенных признаков и определение границ между классами. В традиционных задачах интеллектуального анализа данных классификация рассматривается как один из методов обнаружения закономерностей и построения моделей на основании данных [3]. Однако применительно к реляционным структурам данных выбор признаков часто зависит от того, какие таблицы и атрибуты считаются релевантными для пользователя. Если эта релевантность не формализована, модель классификации может учитывать статистически доступные, но содержательно второстепенные признаки.

Классические методы анализа данных, включая классификацию и кластеризацию, предполагают предварительную подготовку признакового пространства. Для реляционных данных эта подготовка может быть сложной, поскольку объект классификации нередко распределён по нескольким связанным таблицам. Например, характеристика клиента может включать данные из таблицы профиля, истории заказов, платежей, обращений в поддержку и маркетинговых взаимодействий. Само решение о том, какие из этих данных важны для классификации, относится уже не только к технической задаче извлечения признаков, но и к пользовательской семантике.

Таким образом, пользовательская семантика влияет на классификацию как минимум в четырёх аспектах. Во-первых, она определяет, какие признаки должны быть извлечены из реляционной структуры. Во-вторых, она задаёт правила преобразования исходных значений в смысловые категории. В-третьих, она участвует в определении уровней агрегирования, на которых будет выполняться классификация. В-четвёртых, она влияет на интерпретацию результата, поскольку один и тот же класс может иметь разный смысл в различных пользовательских сценариях.

Особое значение имеет различие между структурной, статистической и семантической близостью объектов. Структурная близость определяется совпадением или сходством значений атрибутов. Статистическая близость выражается через меры расстояния, корреляции или вероятностные зависимости. Семантическая близость определяется принадлежностью объектов к близким понятиям предметной области. В реляционных структурах эти виды близости могут расходиться. Пользовательская семантика позволяет преодолеть это расхождение за счёт введения дополнительных правил интерпретации.

Формализация пользовательской семантики также важна для построения многоуровневых классификационных схем. В таких схемах объект сначала относится к укрупнённой категории, затем – к более частному подклассу. Реляционная база может содержать только отдельные признаки,

тогда как иерархия классов формируется на основании пользовательского понимания предметной области. Следовательно, без семантического слоя невозможно корректно описать переход от значений атрибутов к уровням классификации.

Отдельного внимания заслуживает вопрос о соотношении пользовательской семантики и нормализации данных. Нормализация направлена на устранение избыточности и аномалий обновления в реляционной структуре [4]. Однако нормализованная схема не всегда удобна для аналитического агрегирования, поскольку смысловой объект может быть распределён по нескольким таблицам. Пользователь, напротив, часто мыслит объектами и категориями более высокого уровня. Поэтому при построении агрегатов необходимо выполнять семантическую реконструкцию объекта из связанных отношений. Такая реконструкция требует знания того, какие связи существенны для конкретной задачи, а какие могут быть проигнорированы.

В базах данных различают концептуальный, логический и физический уровни представления информации [5]. Пользовательская семантика может быть соотнесена прежде всего с концептуальным уровнем, однако она не полностью совпадает с ним. Концептуальная модель обычно создаётся на этапе проектирования системы и отражает устойчивую структуру предметной области. Пользовательская семантика может быть более динамичной, поскольку зависит от аналитического сценария. Один и тот же набор данных может использоваться для финансовой оценки, маркетинговой сегментации, контроля рисков или прогнозирования поведения. В каждом случае основания агрегирования и классификации будут различаться.

Следовательно, формализация пользовательской семантики может рассматриваться как промежуточный этап между проектированием схемы данных и применением методов классификации. На этом этапе исходные реляционные структуры преобразуются в признаки и категории, имеющие содержательную интерпретацию. Такой подход позволяет объединить преимущества реляционной модели, такие как строгость,

структурированность и целостность, с преимуществами семантического моделирования, ориентированного на понятия предметной области [6]. В результате агрегирование становится не только технической операцией, но и способом построения смысловой иерархии данных.

Одновременно необходимо учитывать ограничения рассматриваемого подхода. Во-первых, пользовательская семантика может быть неоднозначной. Разные пользователи способны задавать различные критерии значимости и разные правила отнесения объектов к классам. Во-вторых, семантические правила могут устаревать, если меняются процессы предметной области. В-третьих, чрезмерно сложная семантическая модель может затруднить сопровождение системы. В-четвёртых, при наличии нескольких групп пользователей возможны конфликты интерпретаций. Поэтому формализация пользовательской семантики должна сопровождаться процедурами согласования, документирования и проверки правил [3].

Несмотря на указанные ограничения, включение пользовательской семантики в процесс иерархического агрегирования реляционных данных имеет существенные преимущества. Оно повышает прозрачность аналитической процедуры, делает классификацию более объяснимой, обеспечивает связь между структурой данных и понятиями предметной области, а также позволяет адаптировать модель обработки к конкретным пользовательским задачам. В отличие от подхода, основанного только на таблицах и атрибутах, семантически ориентированный подход учитывает, что данные используются в контексте деятельности пользователя и потому должны интерпретироваться с учётом этого контекста.

Таким образом, формализация пользовательской семантики представляет собой необходимое условие содержательного иерархического агрегирования и классификации реляционных структур данных. Реляционная модель обеспечивает строгую структурную основу, но не всегда выражает пользовательские критерии значимости и смысловой близости объектов. Для восполнения этого ограничения могут использоваться метаданные,

экспертные правила, иерархии понятий, онтологии и весовые модели признаков. Их применение позволяет перейти от обработки отдельных записей к построению интерпретируемых агрегатов и классов, соответствующих задачам пользователя. Полученные положения могут быть использованы как теоретическая основа для разработки модели, в которой пользовательская семантика выступает самостоятельным компонентом процесса агрегирования и классификации данных.

Использованные источники:

1. Дородных Н. О., Юрин А. Ю. Подход к автоматизированному наполнению графов знаний сущностями на основе анализа таблиц // Онтология проектирования. 2022. №3 (45). URL: https://cyberleninka.ru/article/n/podhod-k-avtomatizirovannomu-napolneniyu-grafov-znaniy-suschnostyami-na-osnove-analiza-tablits (дата обращения: 15.05.2026).
2. Видия А. В., Дородных Н. О., Юрин А. Ю. Подход к созданию онтологий на основе электронных таблиц с произвольной структурой // Онтология проектирования. 2021. №2 (40). URL: https://cyberleninka.ru/article/n/podhod-k-sozdaniyu-ontologiy-na-osnove-elektronnyh-tablits-s-proizvolnoy-strukturoy (дата обращения: 15.05.2026).
3. Лукичев Руслан Владимирович эволюция технологий семантического веба: проблемы и перспективы // Программные системы и вычислительные методы. 2024. №3. URL: https://cyberleninka.ru/article/n/evolyutsiya-tehnologiy-semanticheskogo-veba-problemy-i-perspektivy (дата обращения: 15.05.2026).
4. Светлана Игоревна Чуприна, Ксения Вадимовна Гимашева Методы и средства виртуальной семантической интеграции данных из распределенных разнородных источников // Вестник Пермского университета. Серия: Математика. Механика. Информатика. 2025. №1 (68). URL: https://cyberleninka.ru/article/n/metody-i-sredstva-virtualnoy-semanticheskoy-integratsii-dannyh-iz-raspredelennyh-raznorodnyh-istochnikov (дата обращения: 15.05.2026).

5. С. А. Дерябин, И. О. Темкин Онтологическое моделирование и управление цифровой трансформацией архитектуры горно-добывающих предприятий // Записки Горного института. 2025. №275. URL: https://cyberleninka.ru/article/n/ontologicheskoe-modelirovanie-i-upravlenie-tsifrovoy-transformatsiey-arhitektury-gorno-dobyvayuschih-predpriyatiy (дата обращения: 15.05.2026).
6. Дородных Никита Олегович, Юрин Александр Юрьевич Разработка предметных графов знаний на основе семантического аннотирования табличных данных // Онтология проектирования. 2024. №4 (54). URL: https://cyberleninka.ru/article/n/razrabotka-predmetnyh-grafov-znaniy-na-osnove-semanticheskogo-annotirovaniya-tablichnyh-dannyh (дата обращения: 15.05.2026).