УДК 004.67

СЕМАНТИЧЕСКОЕ СЛИЯНИЕ ДАННЫХ НА ОСНОВЕ ГРАФОВ ЗНАНИЙ: ИНТЕГРАЦИЯ СТРУКТУРИРОВАННЫХ И НЕСТРУКТУРИРОВАННЫХ ДАННЫХ ДЛЯ РАСКРЫТИЯ КОНТЕКСТНОГО ИНТЕЛЛЕКТА

Е.А. Алуев
Инженер-исследователь ATEK,
Бакалавр технических наук
alooeff@atek.dev

М.А. Булычева
Инженер ATEK,
Магистр математических наук
mbulytcheva@gmail.com

Е.А. Алуев
Бакалавр компьютерных наук, факультет ЭВМиС Брестского государственного технического университета. Научные интересы: облачные технологии, машинное обучение и мультиагентное моделирование.

М.А. Булычева
Магистр математических, кафедра механики и математики Государственного университета им. Ломоносова, Москва. Научные интересы: машинное обучение, крупномасштабная оптимизация и прогнозирование.

Аннотация. Компании сектора профессиональных услуг работают с разнородными данными, где структурированные системы сосуществуют с большими массивами неструктурированных документов. Их раздельная обработка ограничивает получение контекстно-зависимых и межисточниковых инсайтов, необходимых для таких задач, как выявление конфликтов, поиск экспертизы и оценка рисков. В работе

предлагается архитектура слияния данных на основе графа знаний, объединяющая структурированные и неструктурированные корпоративные данные в единое семантическое представление с сохранением происхождения данных. Для извлечения сущностей и отношений из документов используются методы обработки естественного языка, а разрешение сущностей между источниками выполняется с помощью гибридной стратегии, сочетающей лексическое, семантическое и реляционное сопоставление. Полученный размеченный граф свойств поддерживает логический вывод и построение признаков для машинного обучения. Результаты показывают, что такой подход повышает качество сопоставления сущностей, улучшает эффективность последующей аналитики и обеспечивает более высокую интерпретируемость. Семантическая интеграция на основе графов знаний тем самым выступает масштабируемой и эффективной основой для корпоративной аналитики в гетерогенных средах данных.

Ключевые слова: граф знаний; интеграция данных; неструктурированные данные; сопоставление сущностей; семантический анализ; корпоративная аналитика.

Введение. Рост объемов структурированных и неструктурированных данных в корпоративных средах создает серьезные вызовы для аналитики и систем поддержки принятия решений. В компаниях сектора профессиональных услуг данные распределены по разрозненным источникам, включая CRM-системы, workflow-платформы и массивы текстовых документов, таких как договоры, отчеты и переписка. Их раздельная обработка ограничивает возможность получения межисточниковых инсайтов и устранения семантической неоднозначности.

Графы знаний представляют собой эффективный подход к интеграции гетерогенных данных в единое семантическое пространство, где сущности и их отношения могут анализироваться совместно. Такое представление позволяет учитывать контекст и поддерживать более сложные аналитические сценарии, включая выявление конфликтов, поиск экспертизы и профилирование рисков.

В данной статье предлагается масштабируемая архитектура построения корпоративного графа знаний, объединяющая структурированные и неструктурированные данные из нескольких источников для поддержки аналитики и задач машинного обучения. Показано, что семантическая интеграция повышает полноту данных, контекстную релевантность и объяснимость результатов.

Основной вклад работы состоит в следующем:

- предложен масштабируемый конвейер приема и семантического слияния данных из нескольких источников;
- разработаны методы извлечения и сопоставления сущностей в гетерогенных данных;
- продемонстрированы преимущества объединенных представлений по сравнению с раздельной обработкой источников.

Предпосылки и обзор литературы.

Графы знаний и интеграция данных. Граф знаний формально определяется как графовая репрезентация объектов реального мира и их семантических отношений, позволяющая интегрировать различные типы данных через узлы, представляющие сущности, и ребра, представляющие отношения. Предыдущие исследования показали, что графы знаний являются фундаментальной структурой интеграции гетерогенных данных, включая как структурированные, так и неструктурированные источники. Tamašauskaitė и соавт. определяют графы знаний как системы, аккумулирующие и передающие знания о реальном мире путем представления сущностей и их связей в семантически насыщенном формате, подчеркивая роль графов в преодолении разрозненности данных [1].

Cudré-Mauroux и соавт. описывают конвейер интеграции данных, демонстрирующий, каким образом полуструктурированный и неструктурированный контент может быть интегрирован с использованием графов знаний, включая предварительную обработку, преобразование и семантическое связывание элементов данных [2].

Сопоставление сущностей и слияние знаний. Сопоставление сущностей означает установление соответствия между семантически идентичными сущностями в разных наборах данных или графах знаний, что является критически важным для многоканального

слияния знаний. Современные обзоры подчеркивают, что сопоставление сущностей - это ключевая технология, обеспечивающая интеграцию разрозненных данных в единые графовые структуры. Методы обучения представлений для сопоставления сущностей демонстрируют более высокую эффективность по сравнению с традиционными признаковыми подходами, особенно в крупных и гетерогенных графовых средах [3].

Исследования в области многоканального слияния знаний подчеркивают, что эффективное объединение требует не только связывания сущностей, но и разрешения конфликтов, выравнивания атрибутов и вывода новых отношений. Именно эти процессы во многом определяют качество и практическую ценность итогового графа знаний для аналитики и логического вывода [4].

Построение графов знаний из структурированных и неструктурированных данных. Исследования, посвященные построению корпоративных графов знаний, подчеркивают сложность интеграции структурированных записей с сущностями и отношениями, извлеченными из текста. Yan и соавт. предлагают подход к объединению структурированных и неструктурированных данных в интегрированный граф знаний, рассматривая предварительную обработку текста, извлечение информации и стратегии хранения [5]. Более ранняя работа Masoud и соавт. представляет обзор методов автоматического построения графов знаний как из структурированных, так и из текстовых источников данных, выделяя точки интеграции до, во время и после построения графа [6]. В совокупности эти исследования формируют основу для понимания слияния данных на основе графов знаний как самостоятельного исследовательского и инженерного направления, демонстрируя необходимость семантического выравнивания, разрешения сущностей и мультимодальной интеграции для корпоративных аналитических приложений.

Постановка задачи и сценарии использования.

Гетерогенные корпоративные данные. Компании обычно хранят структурированные данные в реляционных базах данных, CRM-системах и системах учета финансовых транзакций, тогда как неструктурированные данные, такие как электронные письма, отчеты, юридические заключения и договоры, размещаются в документных хранилищах. Такое разнообразие усложняет аналитические задачи, требующие межисточникового контекста, например выявление всех упоминаний клиентской сущности в документах разных типов или сопоставление структурированных транзакционных историй с их описанием в повествовательной форме.

Аналитические сценарии использования. Объединенные представления знаний поддерживают ряд высокоценных сценариев применения:

– выявление конфликтов: обнаружение пересечений или несовместимостей в клиентских взаимодействиях между подразделениями;

– поиск экспертизы: профилирование внутренней предметной экспертизы путем связывания структурированных кадровых баз с вкладом сотрудников в документы;

– оценка рисков: агрегирование факторов риска из структурированных журналов комплаенса и текстовых оценок рисков для формирования более полной риск-оценки.

Иллюстративный сценарий из сферы коммерческой недвижимости. Чтобы показать архитектуру на конкретном примере, рассмотрим процесс инвестирования в коммерческую недвижимость. Структурированные CRM-системы отслеживают инвестиционные возможности, брокеров, показатели андеррайтинга и стадии воронки. Параллельно неструктурированные информационные меморандумы о продаже объекта (Offering Memoranda, OM) содержат описания характеристик объектов, финансовых допущений, состава арендаторов, долговых ковенант и раскрытия рисков.

Например, запись в CRM может содержать:

- идентификатор возможности #7421;
- целевой объект: “Orchard Lofts”;

- ожидаемый NOI;
-назначенного брокера.

В то же время документ OM, например, содержит текстовые утверждения, такие как:

- заявленная ставка капитализации (5,2%);
- долговой ковенант (DSCR ≥ 1,25);
- экологическая проверка Phase I находится в процессе;
- сведения о концентрации арендаторов.

Изолированные системы не могут ответить на такие запросы, как:

- «Показать 10 лучших возможностей по ожидаемому NOI, исключив сделки, где в OM указана заполняемость ниже 90%»;
- «Выявить возможности, где одна и та же организация выступает и брокером, и управляющей компанией объекта».

Для ответа на такие вопросы требуется объединение структурированных CRM-данных с фактами, извлеченными из неструктурированных OM-документов (рисунок 1).

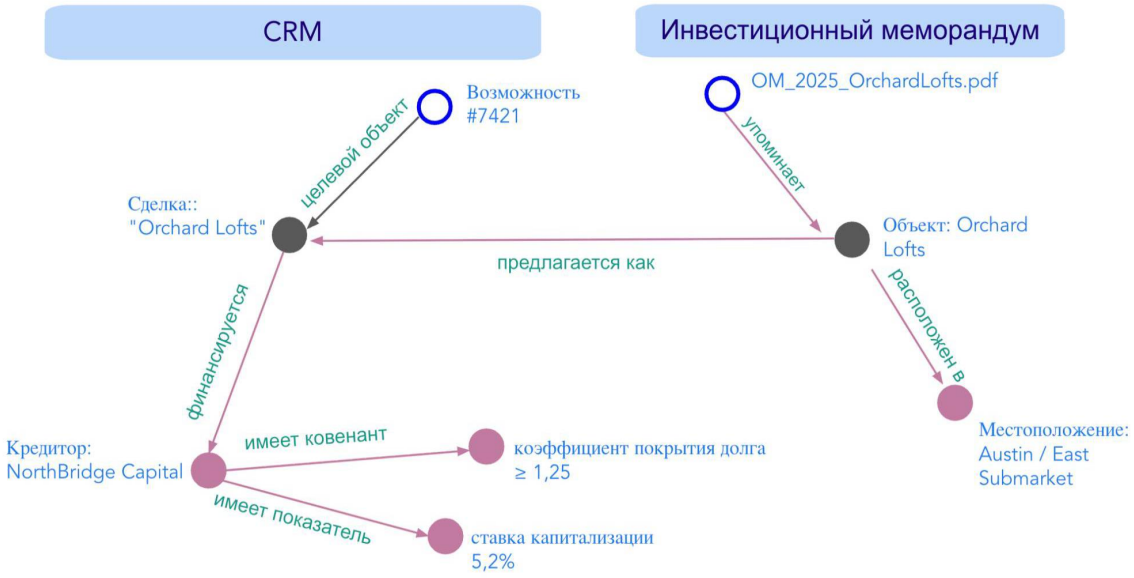

Рисунок 1. Пример

Помимо простого извлечения данных, такие запросы требуют интеграции разнородных свидетельств из разных модальностей и логического вывода по отношениям, охватывающим несколько систем. Например, чтобы определить, превышает ли концентрация арендаторов заданный порог, может потребоваться объединение структурированных данных андеррайтинга со списками арендаторов, извлеченными из нарративных разделов OM. Аналогично, выявление организационных конфликтов может требовать разрешения разных вариантов имен сущностей в CRM-записях и упоминаниях в документах, а затем анализа ролевых связей, например брокера, управляющего объектом или кредитора, в рамках одной и той же сделки. Эти задачи по своей природе реляционны и не могут быть сведены к изолированному поиску по ключевым словам или простым соединениям таблиц.

Архитектура системы. Предлагаемая архитектура представляет собой модульный многоэтапный конвейер, в рамках которого гетерогенные корпоративные данные последовательно преобразуются в единый граф знаний с сохранением происхождения данных. Вместо того чтобы рассматривать структурированные и неструктурированные данные как параллельные аналитические потоки, архитектура изначально проектируется

так, чтобы сводить их в единый семантический слой, поддерживающий логический вывод, аналитику и машинное обучение.

На высоком уровне система разделяет задачи приема данных, семантической интерпретации, выравнивания идентичности, построения графа и последующей аналитики. Каждый этап может развиваться и масштабироваться независимо, что позволяет архитектуре адаптироваться к новым источникам данных, обновленным моделям извлечения и меняющимся требованиям к схеме без нарушения работы всей системы. Такая модульность особенно важна для корпоративных сред, где форматы данных, бизнес-процессы и регуляторные ограничения меняются со временем.

Центральным принципом проектирования является рассмотрение отношений как объектов первого класса. Отношения не редуцируются к неявным соединениям по внешним ключам и не уплощаются на этапе построения признаков; напротив, они явно моделируются, типизируются, версионируются и доступны для запросов внутри графа. Такой подход сохраняет структурные зависимости, придающие корпоративным данным смысл, включая договорные связи, консультационные отношения, финансовые зависимости и ссылки на документы, что делает возможными многоскачковый логический вывод и контекстную агрегацию.

Не менее важно и то, что происхождение данных и их временная семантика сохраняются на всех этапах конвейера. Каждая сущность и каждое отношение сопровождаются метаданными об источнике, времени извлечения, интервале действия и уровне уверенности. Это позволяет рассматривать результирующий граф знаний не просто как артефакт интеграции, а как управляемое и аудируемое представление, пригодное для аналитических процессов с высокими требованиями к надежности.

Рисунок 2. Архитектура системы

Прием и предварительная обработка данных. Корпоративные данные поступают из гетерогенных источников, различающихся по схемам, форматам и частоте обновления. Структурированные данные обычно включают записи из CRM-систем, баз workflow и систем учета финансовых транзакций, тогда как неструктурированные данные состоят из текстовых документов, таких как договоры, отчеты, электронные письма и внутренние меморандумы.

На этапе приема структурированные данные отображаются в каноническую схему с помощью процедур выравнивания схем и нормализации. Этот шаг устраняет

синтаксические несоответствия, например различия в соглашениях об именовании и типах данных, и обеспечивает согласованное представление основных идентификаторов, таких как клиенты, дела, организации и физические лица, во всех источниках.

Неструктурированные документы проходят предварительную обработку, включающую извлечение текста, определение языка, сегментацию на предложения и токенизацию. Метаданные, такие как тип документа, время создания, авторство и права доступа, сохраняются и связываются с представлением документа, чтобы позднее использоваться для контекстной фильтрации.

Семантическое извлечение из неструктурированных данных. Чтобы интегрировать неструктурированный текст в граф знаний, необходимо извлечь из него семантическую информацию в структурированной форме. Это достигается с помощью конвейера обработки естественного языка, который выявляет сущности, отношения и релевантные атрибуты в документах.

Модели распознавания именованных сущностей (Named Entity Recognition, NER) используются для обнаружения специфичных для предметной области типов сущностей, таких как физические лица, организации, клиенты, юрисдикции и договорные понятия. Методы извлечения отношений определяют семантические связи между сущностями, например связи между людьми и организациями или отсылки между документами и бизнес-сущностями.

Извлеченные сущности не рассматриваются как изолированные упоминания; каждая сущность обогащается контекстными сигналами, полученными из окружающего текста, метаданных документа и лингвистических признаков. Эти сигналы впоследствии используются для дизамбигуации и сопоставления сущностей между источниками.

Разрешение и сопоставление сущностей между источниками.

Постановка задачи. Разрешение сущностей между источниками направлено на выявление случаев, когда сущности, происходящие из структурированных систем E_s, и сущности, извлеченные из неструктурированного текста E_u, обозначают один и тот же объект реального мира.

Пусть: $E_S = \{e_1^s, ..., e_n^s\},\ E_u = \{e_1^u, ..., e_m^u\}$

Тогда целью является построение отображения: $\Phi : E_S \bigcup E_u \to E_c$

где E_c обозначает канонические сущности в объединенном графе.

Генерация кандидатов. Для обеспечения масштабируемости сопоставление выполняется в два этапа. Сначала формируются пары-кандидаты с высоким recall с использованием эвристик блокировки:

- нормализованного сравнения строк;
- общих идентификаторов, таких как домены электронной почты и адреса;
- совместной встречаемости в одном документе или сделке;
- пересекающихся структурированных атрибутов.

Это позволяет сократить квадратичное пространство сравнений до вычислительно приемлемого подмножества.

Построение гибридных признаков. Для каждой пары-кандидата (ei,ej) вычисляется вектор признаков x_ij, основанный на трех независимых группах сигналов.

Лексическое сходство включает:

- нормализованное расстояние редактирования;
- пересечение токенов;
- раскрытие аббревиатур.

Семантическое сходство включает:

- контекстные эмбеддинги, полученные из упоминаний сущностей;
- косинусное сходство между векторами эмбеддингов;
- агрегированные контекстные представления, полученные из разделов документов.

Реляционная согласованность включает:

- общих соседей в частичной структуре графа;
- совместимые типы отношений;
- метрики сходства окрестностей.

Реляционный сигнал формализует интуицию о том, что сущности с похожими реляционными окрестностями с высокой вероятностью являются идентичными.

Сопоставление на основе уверенности. Пары-кандидаты оцениваются с помощью обучаемой или эвристической функции:

$s(e\_i,e\_j)=f(x\_ij)$

Далее применяются пороговые правила принятия решений:

- высокая уверенность - автоматическое слияние;
- средняя уверенность - мягкая связь, при которой сущности сохраняются раздельно, но соединяются;
- низкая уверенность - отсутствие слияния.

Такой механизм позволяет сбалансировать цену ложных объединений и пропущенных совпадений.

Сохранение происхождения данных. При слиянии сущностей исходные узлы источников сохраняются как ссылки на происхождение. Каждая каноническая сущность хранит:

- исходные записи;
- уровень уверенности сопоставления;
- временные метки извлечения.

Это обеспечивает аудируемость и обратимость.

Схема графа знаний и его построение. Граф знаний строится как размеченный граф свойств, в котором узлы представляют сущности, а ребра – типизированные отношения. Схема задается через легковесную онтологию, охватывающую ключевые предметные понятия и их связи, оставаясь при этом достаточно гибкой для учета развивающихся источников данных.

Узлы сущностей хранят атрибуты, происходящие как из структурированных систем, так и из текстовой информации, извлеченной из документов. Ребра отношений кодируют как явные связи, например отношения между клиентом и проектом, так и выведенные отношения, полученные из текста или графового вывода.

Построение графа выполняется инкрементально, что позволяет добавлять новые данные без необходимости полного пересчета. Механизмы версионирования обеспечивают отслеживание изменений сущностей и отношений во времени, поддерживая аудируемость и воспроизводимость аналитических результатов.

Граф следует модели размеченного графа свойств.

Типы узлов (иллюстративный сценарий для рынка недвижимости). Граф знаний реализован как размеченный граф свойств, в котором узлы представляют канонизированные сущности предметной области, охватывающие транзакционные записи, организационных участников, документы, финансовые атрибуты и индикаторы риска. Ребра представляют типизированные семантические отношения, отражающие структурные, договорные и контекстные зависимости между сущностями.

В единой схеме представлены как отношения из структурированных систем, так и ассоциации, извлеченные из текста.

Каждое ребро хранит метаданные о происхождении, временной валидности и уровне уверенности. Это позволяет графу поддерживать версионирование, анализ с учетом

неопределенности и воспроизводимую аналитику, одновременно сохраняя связь с исходными структурированными записями и документными источниками.

Рисунок 3 показывает, каким образом гетерогенные элементы, происходящие из структурированных CRM-систем и неструктурированных информационных меморандумов, объединяются в рамках единой реляционной модели.

Транзакционные записи, финансовые метрики, ссылки на документы и раскрытия рисков представлены как взаимосвязанные сущности, а не как изолированные артефакты данных. Такая конструкция позволяет явно выражать и единообразно анализировать такие отношения, как финансовые зависимости, консультативные роли, связи с арендаторами или ковенантные ограничения.

Важно отметить, что граф не просто воспроизводит структуры исходных данных; он реифицирует семантические отношения как объекты первого класса с собственными метаданными. За счет присоединения к каждому отношению атрибутов происхождения, временной валидности и уверенности граф поддерживает аналитику с учетом неопределенности и принятие решений с возможностью трассировки. В результате последующие запросы и модели машинного обучения работают с представлением, которое сохраняет как структурные зависимости, так и доказательную основу.

Типы узлов:

- объект недвижимости
- сделка / листинг
- организация (продавец, покупатель, кредитор, брокер)
- физическое лицо (команда сделки, подписанты)
- местоположение (адрес, город, субрынок)
- договор аренды / арендное соглашение
- арендатор (может быть подтипом)
- финансовый показатель (NOI, ставка капитализации, DSCR, IRR)
- документ (инвестиционный меморандум, договор аренды, отчет об оценке)
- фактор риска (экологический риск, риск вакантности, долговые ковенанты)

Связи:

- расположен в
- предлагается как
- ринадлежит / продается
- представлен
- управляется
- сдан в аренду
- имеет договор аренды
- имеет показатель
- финансируется
- подпадает под ковенант
- имеет фактор риска
- ссылается на *(для подтверждения источника)*

Рисунок 3. Структура графа знаний

Слой графового вывода и аналитики. Граф знаний поддерживает два взаимодополняющих режима анализа:

- детерминированные многоскачковые графовые запросы;
- статистическое обучение с использованием представлений, производных от графа.

Обход графа и запросы на сопоставление шаблонов позволяют выявлять косвенные отношения, агрегировать свидетельства из документов и отслеживать взаимодействия сущностей во времени.

Параллельно структура графа преобразуется в численные представления посредством:

- локальных структурных статистик, таких как степень вершины и количество соседей разных типов;
- признаков, основанных на путях и реляционном контексте;

- представлений на основе эмбеддингов, обученных методами графовых эмбеддингов или графовых нейронных сетей.

Эти признаки используются в последующих задачах классификации, ранжирования и предсказания связей.

Графовые запросы и символический вывод. Граф поддерживает многоскачковые запросы и структурное сопоставление шаблонов, что позволяет решать задачи логического вывода, такие как обнаружение косвенных связей, агрегирование свидетельств из документов или отслеживание взаимодействий сущностей во времени. Эти запросы выполняются непосредственно над топологией графа и семантикой ребер, обеспечивая детерминированные и объяснимые результаты.

Генерация признаков на основе графа. Для поддержки моделей машинного обучения структура графа преобразуется в численные представления. Для каждого узла-сущности его признаки могут включать:

- локальные структурные статистики, такие как степень и состав окрестности;
- признаки на основе путей, отражающие реляционный контекст;
- векторы эмбеддингов, обученные на структуре графа.

Графовые эмбеддинги обучаются методами, сохраняющими близость и реляционное сходство, что позволяет сущностям с похожими ролями или контекстами располагаться близко друг к другу в пространстве эмбеддингов. Такие эмбеддинги служат компактными представлениями сложной реляционной информации.

Последующие задачи машинного обучения. Графовые признаки интегрируются в модели машинного обучения с учителем и без учителя для решения таких задач, как классификация, ранжирование и обнаружение аномалий. Важно, что использование графовых признаков часто повышает обобщающую способность моделей за счет учета реляционного контекста, отсутствующего в плоских признаковых представлениях.

Предсказания, формируемые этими моделями, могут быть соотнесены с лежащими в их основе структурами графа, что позволяет проводить post hoc объяснение путем анализа влиятельных узлов, ребер или путей. Это особенно важно для компаний сектора профессиональных услуг, где требуются прозрачность и аудируемость.

Оценка. Мы оцениваем предлагаемый подход, сравнивая эффективность аналитики при использовании и без использования слияния на основе графа знаний. В качестве метрик рассматриваются precision и recall для разрешения сущностей, точность классификации в задаче поиска экспертизы, а также контекстная релевантность, оцениваемая экспертами-аннотаторами. Предварительные результаты показывают заметные улучшения при использовании объединенных графов, подтверждая ценность семантической интеграции.

Цели оценки. Оценка направлена на анализ эффективности слияния данных на основе графа знаний по трем направлениям:

- качество разрешения сущностей;
- влияние на последующие аналитические задачи;
- качественные улучшения в контекстном логическом выводе.

Вместо концентрации на абсолютных значениях метрик акцент делается на сравнительных улучшениях относительно базовых подходов, обрабатывающих структурированные и неструктурированные данные независимо.

Оценка разрешения сущностей. Качество сопоставления сущностей оценивается с использованием вручную подготовленных эталонных соответствий для репрезентативного подмножества сущностей. Используются стандартные метрики:

- precision - для измерения числа ошибочных слияний;
- recall - для измерения числа пропущенных соответствий;
- F1-мера - для сбалансированной оценки обоих аспектов.

Результаты показывают, что включение семантических и реляционных признаков существенно улучшает качество сопоставления по сравнению с базовыми подходами,

использующими только лексические признаки, особенно в случаях неоднозначных или сокращенных упоминаний сущностей.

Оценка последующей аналитики. Для оценки влияния слияния данных на аналитические задачи графовые признаки используются в моделях машинного обучения с учителем для репрезентативных задач, таких как классификация экспертизы и предсказание отношений.

Качество измеряется с помощью метрик, соответствующих конкретной задаче, включая точность классификации и качество ранжирования. Модели, использующие граф-обогащенные представления, последовательно превосходят базовые модели, обученные только на структурированных или только на неструктурированных данных, что демонстрирует ценность интегрированной контекстной информации.

Качественный анализ. Помимо количественных метрик, проводится качественный анализ путем изучения путей логического вывода внутри графа знаний.

Эти анализы показывают, что объединенные представления позволяют осуществлять многоскачковый вывод через документы и структурированные записи, поддерживая сложные запросы, которые невозможно реализовать в изолированных конвейерах обработки данных. Такая интерпретируемость особенно важна в среде профессиональных услуг, где аналитики и предметные эксперты должны иметь возможность проверять и обоснованно доверять результатам системы.

Ограничения. Проведенная оценка имеет ряд ограничений.

Разметка ground truth для сопоставления доступна только для части сущностей, а показатели эффективности могут варьироваться в предметных областях с иными соглашениями об именовании или другой структурой документов. Тем не менее наблюдаемые тенденции последовательно свидетельствуют в пользу графового слияния по сравнению с базовыми подходами.

Заключение и направления дальнейшей работы. Мы представили архитектуру слияния данных на основе графа знаний, интегрирующую структурированные и неструктурированные корпоративные данные для обеспечения более глубокой аналитики и контекстного интеллекта.

В дальнейшем планируется исследовать интеграцию данных в режиме реального времени, федеративные графы знаний, охватывающие несколько организаций, а также более тесную интеграцию с большими языковыми моделями для интерактивного логического вывода.

Список литературы

[1] G. Tamašauskaitė, P. Groth. Defining a Knowledge Graph Development Process Through a Systematic Review (February 2023).

[2] Philippe Cudré-Mauroux. Leveraging Knowledge Graphs for Big Data Integration: the XI Pipeline (January 2020).

[3] Beibei Zhu, Ruolin Wang, Junyi Wang, Fei Shao, Kerun Wang. A survey: knowledge graph entity alignment research based on graph embedding (August 2024).

[4] Xiaojuan Zhao, Yan Jia, Aiping Li, Rong Jiang, Yichen Song. Multi-source knowledge fusion: a survey (April 2020).

[5] Chenwei Yan, Xinyue Fang, Xiaotong Huang, Chenyi Guo, Ji Wu. A solution and practice for combining multi-source heterogeneous data to construct enterprise knowledge graph (September 2023).

[6] Maraim Masoud, Bianca Pereira, John McCrae, Paul Buitelaar. Automatic Construction of Knowledge Graphs.

Авторский вклад

Е.А. Алуев – анализ существующих решений и общее руководство проектом.

М.А. Булычева – архитектурный анализ и разработка.

SEMANTIC DATA FUSION BASED ON KNOWLEDGE GRAPHS: INTEGRATION OF STRUCTURED AND UNSTRUCTURED DATA TO UNLEASH CONTEXTUAL INTELLIGENCE

Eugene Alooeff
R&D Engineer ATEK,
Bachelor of Computer Science
alooeff@atek.dev

Mariia Bulycheva
Engineer ATEK,
Master of Mathematics
mbulytcheva@gmail.com

Abstract. Professional services companies work with heterogeneous data, where structured systems coexist with large volumes of unstructured documents. Processing them separately limits the acquisition of context-sensitive and cross-source insights needed for tasks such as conflict detection, expertise search, and risk assessment. This paper proposes a knowledge graph-based data fusion architecture that combines structured and unstructured enterprise data into a single semantic representation while preserving data lineage. Natural language processing methods are used to extract entities and relationships from documents, and entity resolution across sources is accomplished using a hybrid strategy combining lexical, semantic, and relational matching. The resulting annotated feature graph supports logical inference and feature generation for machine learning. Results demonstrate that this approach improves the quality of entity matching, enhances the efficiency of subsequent analytics, and ensures greater interpretability. Knowledge graph-based semantic integration thus provides a scalable and efficient foundation for enterprise analytics in heterogeneous data environments.

Keywords: knowledge graph; data integration; unstructured data; entity mapping; semantic analysis; enterprise analytics.