КОМПЬЮТЕРНЫЕ НАУКИ И ИНФОРМАТИКА

Научная статья
УДК 004.89
DOI: 10.17072/1993-0550-2025-1-145-159
https://elibrary.ru/lzqeqn

Методы и средства виртуальной семантической интеграции данных из распределенных разнородных источников

Светлана Игоревна Чуприна[1], Ксения Вадимовна Гимашева[2]

[1]Пермский государственный гуманитарно-педагогический университет, г. Пермь, Россия

[2]Пермский государственный национальный исследовательский университет, г. Пермь, Россия
[1]chuprinas@inbox.ru
[2]gimashevakv@mail.ru

Аннотация. Статья посвящена вопросам автоматизации обработки текстовых данных из распределенных разнородных источников на принципах их виртуальной семантической интеграции. Основная цель интеграции данных заключается в предоставлении пользователю унифицированного доступа к распределенным данным как к единому виртуальному хранилищу для выполнения запросов на естественном языке безотносительно формата хранения данных и их местоположения. В статье рассматриваются основные подходы, ориентированные на виртуальную семантическую интеграцию данных, и описана предлагаемая концепция построения онтологически управляемого инструментального окружения на базе технологии фабрик данных, что позволяет унифицировать и автоматизировать обработку данных за счет промежуточного слоя онтологий. Описывается реализация предложенной концепции в виде инструментального средства NuCoBoShell с сервисом запросов на естественном языке, который в отличие от поисковых сервисов интернет предоставляет возможность получения более полных ответов на запросы посредством автоматического извлечения необходимой информации из виртуальных источников, представляющих собой результат семантической интеграции не только разнородных веб-ресурсов, но и текстовых документов, хранящихся в доступных хранилищах данных и на локальном компьютере пользователя, без необходимости их физического копирования в единое хранилище.

Ключевые слова: *семантическая интеграции данных; виртуальная интеграция; онтология; онтологически управляемое решение; технология фабрик данных*

Для цитирования: *Чуприна С.И., Гимашева К.В.* Методы и средства виртуальной семантической интеграции данных из распределенных разнородных источников // Вестник Пермского университета. Математика. Механика. Информатика. 2025. Вып. 1(68). С. 145–159. DOI: 10.17072/1993-0550-2025-1-145-159. https://elibrary.ru/lzqeqn.

Статья поступила в редакцию 08.08.2024; одобрена после рецензирования 28.02.2025; принята к публикации 24.03.2025.

COMPUTER SCIENCE

Research article

Methods and Tools for Virtual Semantic Integration of Data from Distributed Heterogeneous Sources

Svetlana I. Chuprina[1], Kseniya V. Gimasheva[2]
[1]Perm State Humanitarian Pedagogical University, Perm, Russia
[2]Perm State University, Perm, Russia
[1]chuprinas@inbox.ru
[2]gimashevakv@mail.ru

Abstract. The article is devoted to the natural language processing from distributed heterogeneous sources based on the principles of their virtual semantic integration. The main purpose of data integration is to provide the user with unified access to distributed data as a single virtual storage for performing natural language queries, regardless of the data storage format and location. The article discusses the main approaches focused on virtual semantic data integration, and describes the proposed concept of building an ontology driven instrumental environment based on Data Fabric technology, which allows to automate data processing via intermediate layer of ontologies in a unified form. The article describes NuCoBoShell that is the instrumental environment implementing the proposed approach. NuCoBoShell uses ontology-driven semantic integration mechanism to provide the answering, which, unlike traditional Internet answering services, provides the opportunity to obtain more pertinent answers automatically extracting the necessary information from not only heterogeneous web resources, but also text documents stored in accessible data warehouses and user's local computer without the need to copy data to a single repository.

Keywords: *semantic data integration; virtual integration; ontology; ontology-driven development; data fabric technology*

For citation: Chuprina, S. I. and Gimasheva, K. V. (2025), "Methods and Tools for Virtual Semantic Integration of Data from Distributed Heterogeneous Sources", *Bulletin of Perm University. Mathematics. Mechanics. Computer Science*, no. 1(68), pp. 145-159. (In Russ.). DOI: 10.17072/1993-0550-2025-1-145-159. https://elibrary.ru/lzqeqn.

The article was submitted 08.08.2024; approved after reviewing 28.02.2025; accepted for publication 24.03.2025.

Введение

В условиях всеобщей цифровизации объем получаемой и производимой человеком информации непрерывно растет. В связи с этим проблема автоматизации и унификации обработки данных с целью адаптации к персональным предпочтениям пользователей является актуальным направлением развития информационных технологий. Под обработкой данных понимаются как процессы автоматического сбора и извлечения данных из различных ресурсов, так и выполнение виртуальной семантической интеграции данных.

Необходимость семантической интеграции данных связана с распределенностью и разнородностью информационных ресурсов и выполняется с целью предоставления пользователям возможности работать с данными из множества произвольных ресурсов как с единым целым. Данные из различных источников могут быть не только структурированными, неструктурированными или полуструктурированными и иметь различный формат представления, но и отличаться по местоположению: располагаться в локальных и/или корпоративных, в том числе облачных, хранилищах, в сети интернет или в виде файлов разного формата на локальных компьютерах пользователей.

В ситуации, когда отсутствуют источники, которые содержали бы полный ответ на поставленный пользователем вопрос, возникает потребность в ручном сопоставлении и смысловой интеграции данных из различных источников, так как отсутствуют ориентированные на конечного пользователя (неспециалиста в области ИТ) общедоступные высокоуровневые средства, позволяющие унифицированным образом выполнять автоматизированную обработку данных на принципах семантической интеграции.

Для решения этой проблемы используются разные методы, большинство из которых основано на консолидации данных. Консолидация предполагает создание единого хранилища, в котором происходит интеграция данных из разных источников за счет их автоматизированного извлечения, преобразования и загрузки (англ., ETL – Extract, Transform, Load) специальными средствами. На данный момент наиболее перспективные методы направлены на реализацию другого подхода – виртуальной интеграции, при которой не происходит физического перемещения данных в единое хранилище [1, 2]. Среди технологий, основанных на виртуальной интеграции данных и активно развивающихся в последнее время, можно выделить технологии построения интеллектуальных фабрик данных – нового поколения хранилищ данных [3]. Концепция построения фабрик данных предполагает установление между источниками данных и интерфейсом конечного пользователя промежуточного слоя в виде графа знаний (англ., Knowledge Graph), представленного, например, онтологиями [4], который описывает предметное содержание и метаданные различных распределенных разнородных источников данных, а также интерфейсы для доступа к необходимым программным сервисам их обработки.

В статье описан подход к разработке инструментального окружения, позволяющего автоматизировать и унифицировать семантическую интеграцию распределенных разнородных данных посредством построения виртуального хранилища данных. В качестве технологий виртуальной интеграции данных были выбраны технологии построения интеллектуальных фабрик данных, что не предъявляет повышенных требований к уровню ИТ-квалификации пользователя, ведь интерфейс конечного пользователя представляет собой простую строку для задания поискового запроса по аналогии с поисковыми сервисами интернет. При вводе поискового запроса, благодаря онтологически управляемым решениям, пользователь получает подсказки, учитывающие контекст запроса и специфику предметной области тех информационных ресурсов, которые были выбраны пользователем как доверительные. Отличительной особенностью реализации таких подсказок является управляемый онтологиями механизм семантической фильтрации.

В статье представлена реализация предложенного подхода в виде инструментального средства NuCoBoShell. В отличие от поисковых сервисов интернет, NuCoBoShell позволяет получать более полные с точки зрения содержания ответы на ЕЯ-запросы (запросы на естественном языке) посредством автоматического извлечения необходимой информации из виртуального источника, генерируемого в результате семантической интеграции не только разнородных веб-ресурсов, но и текстовых документов, хранящихся в доступных хранилищах данных и/или на локальном компьютере пользователя без необходимости их физического копирования в единое хранилище и использования больших предобученных языковых моделей (англ., LLM – Large Language Model).

1. Необходимость семантической интеграции данных из распределенных разнородных источников

Потребность в интегрированной обработке как структурированных, так и неструктурированных, а также полуструктурированных текстовых данных, приводит к необходимости унификации методов и средств их автоматизированной обработки, не зависящей ни от специфики предметной области, ни от степени структурированности и формата данных. При этом, несмотря на успехи в решении задач автоматической обработки текстов (АОТ или англ., NLP – Natural Language Processing) методами глубинного машинного обучения (МО), в частности, генеративного искусственного интеллекта (ИИ), когда используются упомянутые выше LLM со множеством параметров (иногда, миллиардами), они далеко не всегда подходят для решения конкретных задач, особенно в тех случаях, когда по тем или иным причинам, например по соображениям безопасности, текстовые данные никогда не выкладывались в общий доступ.

Кроме того, использование для решения задач АОТ методов глубинного МО не оправдано и в тех случаях, когда компания или небольшая исследовательская группа, не говоря уже об индивидуальных исследователях, не обладает достаточными финансовыми и вычислительными ресурсами для дообучения этих моделей, особенно при незначительных объемах исходных данных. Поэтому применение для решения задач АОТ методов, отличных от МО, например, основанных на правилах и лексико-синтаксических шаблонах, методах онтологического инжиниринга или других методах инженерии знаний, по-прежнему остается актуальным направлением развития ИИ. Во многих практических задачах АОТ требуется интеграция указанных методов и средств с методами и средствами МО.

Все вышесказанное напрямую относится и к решению задач семантической интеграции данных, когда в отдельных исходных ресурсах нет полного ответа на сложный комплексный запрос и традиционная поисковая система фактически ранжирует ответы не по степени их пертинентности и релевантности, а раздельно в порядке их релевантности каждой из составных частей запроса. Это можно продемонстрировать на примере такого комплексного запроса как "рецепты с содержанием витамина B12, кальция и магния", в ответ на который традиционный интернет-поисковик сначала выдает две страницы ссылок на одну часть комплексного запроса (см. рис. 1), а затем отдельно – на другую часть, причем в начале выдачи приводятся рецепты безотносительно указанных в запросе нутриентов. Проблема остается и в случае, если из текста запроса убрать кавычки.

Основная цель интеграции данных состоит в том, чтобы объединить информацию из различных источников и представить ее в едином формате, что позволяет пользователю получать доступ к множеству данных без необходимости учитывать внутреннюю структуру и содержание каждого ресурса. Задача семантической интеграции данных усложняется в случае, когда источники данных представляют собой разнородные неструктурированные ресурсы на естественном языке. АОТ представляет собой трудно формализуемую задачу [5], для качественной эффективной обработки текстовых данных необходимо учитывать синонимы и контекст при извлечении и интеграции разнородной информации.

Виртуальная семантическая интеграция данных позволяет не только привести данные к некоторому единому формату или единой структуре, но и повысить качество данных за счет установления смысловых связей между различными понятиями предметной области.

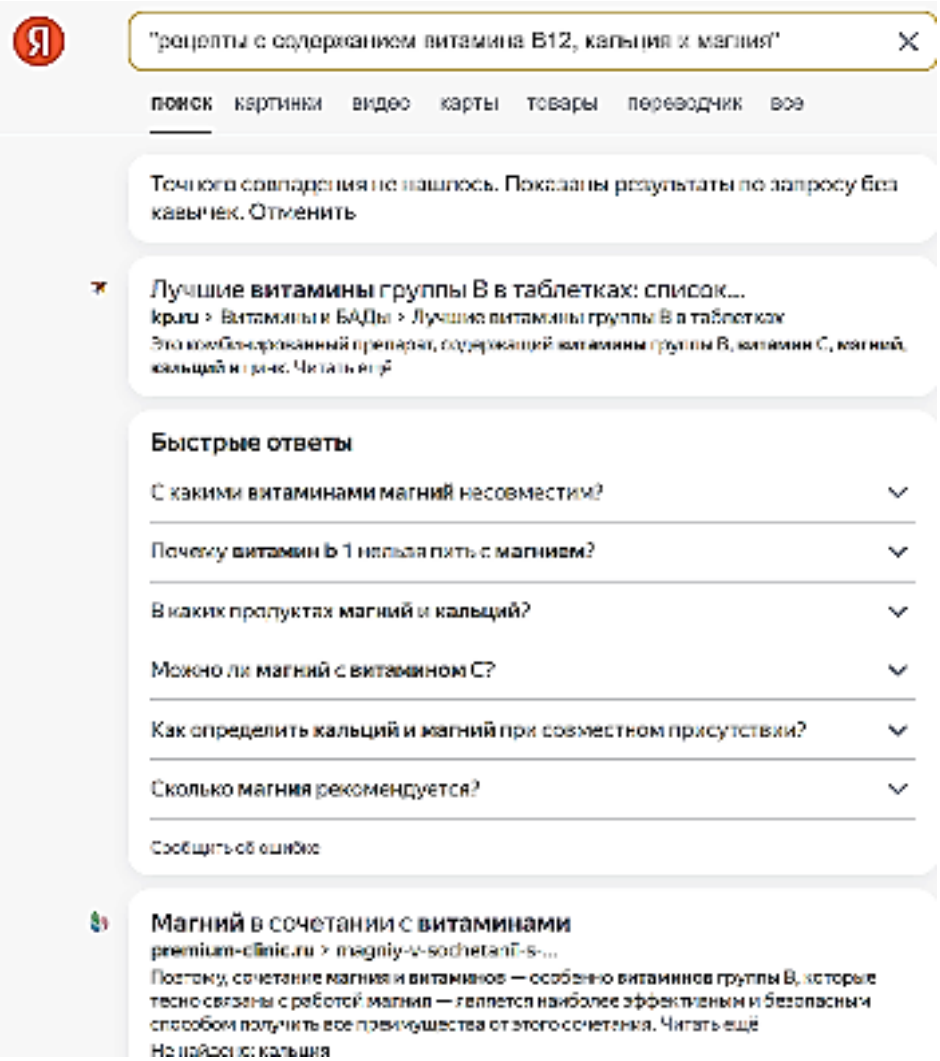

Рис. 1. *Пример нерелевантной интернет-выдачи информации в ответ на комплексный запрос*

В результате появляется возможность выполнять запросы на естественном языке к разнородному виртуальному пространству данных, скрывая от пользователя не только формат и структуру исходных данных, но и их местоположение, так как обработка данных выполняется независимо от того, расположен ли информационный ресурс в сети интернет, в корпоративном хранилище или на локальном компьютере.

2. Описание подходов к семантической интеграции данных

С точки зрения способов интеграции данных в настоящее время существует два основных подхода [1] – консолидация и федерирование. Консолидация, или материализованная интеграция, предполагает создание единого ETL-хранилища данных из интегрируемых информационных ресурсов, которое будет периодически пополняться за счет извлечения, преобразования и загрузки данных специальными средствами. Запросы задаются уже после интеграции данных в единое хранилище, что порождает проблему "свежести" данных.

Федерирование данных или виртуальная интеграция основана на использовании посредников и адаптеров, где посредник представляет собой центральный компонент системы интеграции, а адаптеры обеспечивают единообразное взаимодействие посредника с источниками данных. Ключевая идея подхода состоит в том, что центральный элемент системы интеграции (посредник), благодаря онтологическому слою метаданных, выполняет преобразование запросов, которые трансформируются в подзапросы к отдельным информационным ресурсам (источникам данных). Выполнение подзапросов поддерживается специальными модулями (адаптерами), которые позволяют оптимизировать взаимодействие посредника с источниками данных.

Основное отличие федерирования от консолидации заключается в том, что при виртуальной интеграции происходит извлечение данных из разнородных источников с последующим объединением и анализом в режиме реального времени без их физического перемещения в единое хранилище – местоположение данных не меняется, они остаются у владельцев, а результаты запросов формируются подобно так называемым удаленным представлениям (remote views) и выдаются по требованию. Благодаря виртуальной интеграции данных появляется возможность оперативно учитывать изменения в источниках данных и настраивать ограничения прав доступа как к отдельным распределенным разнородным ресурсам, так и ко всему виртуальному пространству данных.

На рис. 2 продемонстрирована упрощенная схема двух подходов к интеграции данных [6]: слева представлен традиционный подход, основанный на принципах консолидации; справа – подход к виртуальной интеграции данных на принципах федерализации.

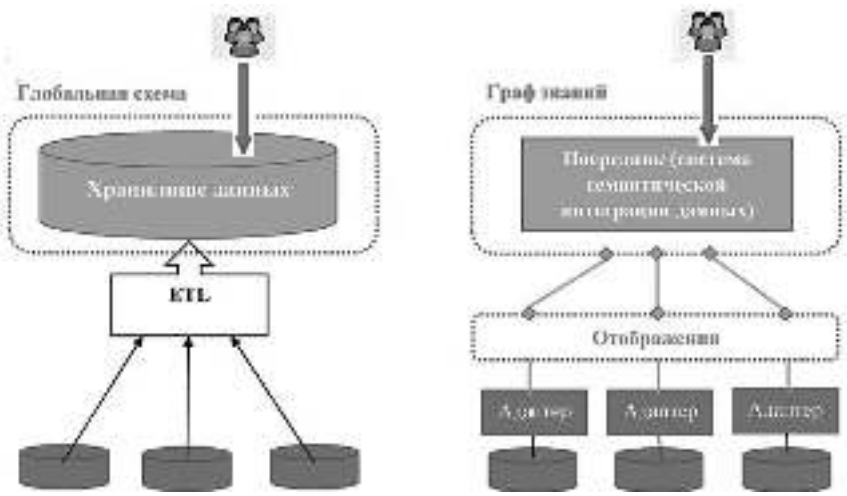

Рис. 2. *Упрощенная схема интеграции на принципах консолидации и федерирования*

Как отмечалось выше, одним из подходов к реализации семантической интеграции данных является применение методов машинного обучения [7], которые позволяют автоматизировать обработку текстов на естественном языке за счет обучения моделей для извлечения закономерностей и прогнозирования результатов на основе исходных данных. Использование методов глубинного МО, например LLM, позволяет генерировать связный и семантически корректный текст с учетом контекста ЕЯ-запроса и синонимов. Однако в случае загрузки данных в открытые общедоступные сервисы, основанные на

LLM, например OpenAI, нет достаточных гарантий защищенности и конфиденциальности исходных данных, а полученные результаты и весь контекст их получения (сам запрос, уточняющие запросы и промежуточные результаты) также становятся общедоступными, так как учитываются генеративным ИИ при поиске и генерации ответов для решения задач других пользователей. Альтернативой является использование локальных языковых моделей, но при этом, как уже отмечалось, для хранения большой предобученной языковой модели и ее дообучения требуются значительные вычислительные ресурсы [8], иначе качество генерируемых ЕЯ-ответов и синтезируемых текстов резко падает.

В обоих случаях возникают проблемы с валидацией полученных результатов, так как корректность ответов в значительной степени зависит от правильно сформулированного контекста запроса. Кроме того, при извлечении закономерностей и связей встает вопрос вообще о целесообразности и необходимости решения этих задач на основе LLM в случае, когда эти закономерности или их часть уже известны пользователю, и он ищет простого легковесного способа их применения для решения своих задач и, возможно, интеграции соответствующих ресурсов в свои или сторонние системы. Поэтому перспективной представляется интеграция подходов на базе методов машинного обучения и инженерии знаний, в частности, онтологического инжиниринга, который играет основную роль в современных технологиях семантического веба (англ., Semantic Web) [3, 9].

Технологии Semantic Web позволяют унифицированным образом выполнять обработку текстов за счет использования единой онтологической модели представления данных. Благодаря единой модели, отвечающей принятым стандартам, а также разработанным стандартизированным языкам запросов, которые являются частью стека технологий Semantic Web, повышается качество реализации поисковых запросов. Однако Semantic Web создавалось как расширение WWW (англ., World Wide Web), которое позволяет представить опубликованную в интернете информацию и данные вместе с машинно-интерпретируемыми метаданными. И хотя концепция представления знаний о предметной области в виде графа знаний может быть использована и для представления содержания локальных текстовых ресурсов, сами технологии предназначены для обработки именно веб-ресурсов.

Данная работа предлагает унифицированный онтологически управляемый подход к смысловой интеграции данных из разнородных ресурсов вне зависимости от их формата, местоположения и контента. В онтологии хранятся знания не только о структуре и содержании разнородных ресурсов, но и метаданные и декларативные знания об их местоположении, необходимых сервисах для их интерпретации, пред- и пост-обработки. Отличительной особенностью нашего подхода является то, что унифицированные способы семантической интеграции распространяются также и на локальные текстовые ресурсы пользователя без их перемещения и/или выкладывания в общий доступ. Однако с увеличением количества ресурсов и, соответственно, объема и/или числа онтологий, описывающих метаданные об этих ресурсах, возникает необходимость в интеграции самих онтологий [9]. В связи с этим встает вопрос о корректности построенных онтологий, необходимости оценки полноты и качества разработанной онтологической модели и самих онтологий.

Для автоматизации этого процесса мы разрабатываем средства визуальной аналитики и обращаемся к сервисам генеративного ИИ. Таким образом, мы используем LLM не для генерации ЕЯ-ответов и синтеза текстов на ЕЯ, а для автоматизации построения онтологий.

3. Виртуальная интеграция данных с использованием технологии Data Fabric

Подход к виртуальной интеграции данных прежде всего направлен на решение проблемы вариативности данных. При обработке большого количества распределенных разнородных ресурсов при условии непрерывно поступающего потока запросов к данным централизованное единое хранилище данных перестает быть эффективным и требует значительных вложений в вычислительные ресурсы. В связи с этим в последнее время активно развивается подход к созданию виртуального хранилища данных на основе технологий построения интеллектуальных фабрик данных [3].

Фабрика данных представляет собой интегрированное виртуальное хранилище, архитектура и сервисы которого обеспечивают смысловую взаимосвязь между данными, хранящимися в разнородных источниках вне зависимости от формата их хранения и технологии создания систем – источников данных [6]. Преимущество технологии построения фабрик данных заключается в управлении процессом поиска и обработки данных из распределенных разнородных источников без их физического перемещения в единое хранилище посредством промежуточного слоя графа знаний, представляемого чаще всего в виде онтологий. Этот граф знаний располагается между интерфейсом конечного пользователя и источниками данных и описывает предметное содержание этих источников в терминах онтологии, а также метаданные об организации ресурсов, включая местоположение, формат и другое. Такой подход позволяет оперативно учитывать изменения в источниках, так как доступ к данным осуществляется по ЕЯ-запросу в терминах онтологии сразу ко всему виртуальному пространству данных, а семантическая интеграция выполняется в режиме реального времени. Для реализации произвольных ЕЯ-запросов используются NLP-сервисы, устанавливающие смысловое соответствие между терминами из ЕЯ-запроса и терминами графа онтологии с учетом синонимов в контексте, обобщающих и конкретизирующих понятий.

На рис. 3 приведена обобщенная архитектура системы виртуальной интеграции данных, представленных не только в различных текстовых форматах, но и в форматах структурированных и полуструктурированных данных.

Центральным компонентом системы интеграции на принципах федерализации является так называемый посредник (англ., mediator), который интегрирует данные, полученные от адаптеров (англ., wrappers), обеспечивающих единообразное взаимодействие посредника с источниками данных в терминах единой модели глобального графа знаний, описывающего метаданные и смысловое содержание ресурсов. Посредник осуществляет интеграцию через сопоставление глобальных и локальных моделей данных.

Пользовательский запрос, сформулированный в терминах предметной области, описанной графом знаний, автоматически декомпозируется на множество подзапросов, адресованных к нужным источникам данных, в том числе локальным. На основе результатов их обработки генерируется более полный ответ на запрос.

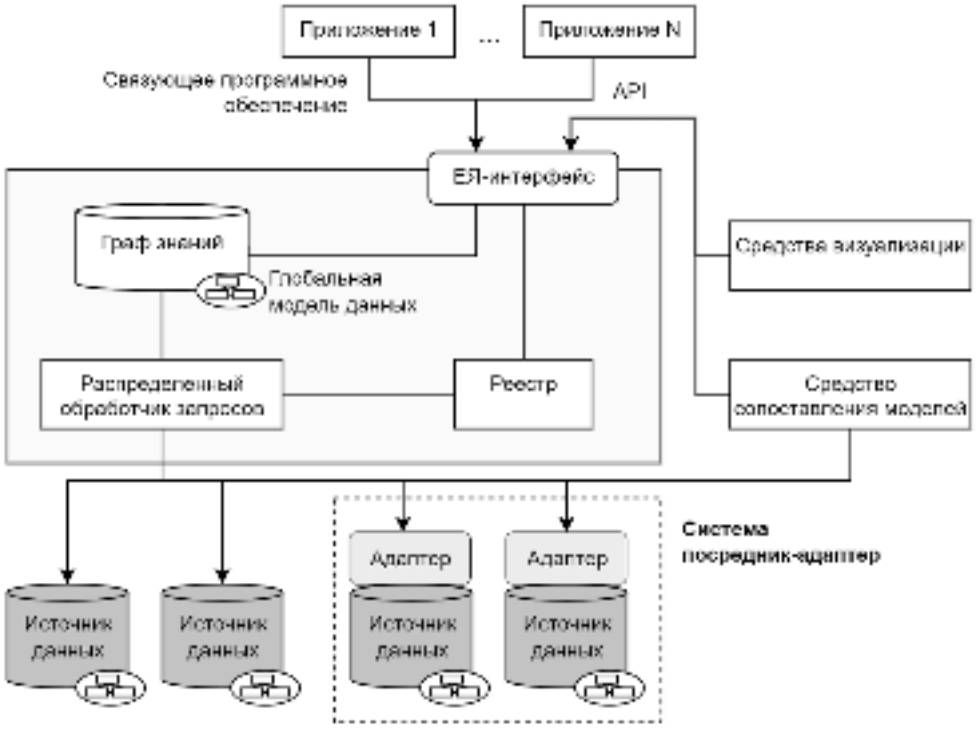

Рис. 3. *Обобщенная архитектура системы виртуальной интеграции данных*

4. Онтологически управляемое инструментальное средство виртуальной семантической интеграции данных

Концепция предлагаемого подхода к разработке инструментального окружения, позволяющего автоматизировать и унифицировать виртуальную семантическую интеграцию распределенных разнородных текстовых ресурсов, основана на использовании описанных выше принципов построения интеллектуальных фабрик данных. Реализация указанных принципов выполнена в среде программной платформы NuCoBoShell, которая является онтологически управляемым инструментальным окружением и использует технологию веб-сервисов для автоматического извлечения необходимых данных из коллекций неструктурированных документов и их семантической интеграции.

Для более качественного решения этих задач дополнительно используются сторонние лингвистические ресурсы и унаследованные базы данных. В результате разработанное в соответствии с предлагаемым подходом инструментальное окружение позволяет получать более полные ответы на пользовательские ЕЯ-запросы за счет автоматически выполняемой предварительной виртуальной интеграции данных. Задача автоматического определения необходимости в указанной предварительной интеграции информационных ресурсов решается специальными интеллектуальными программными средствами, которые анализируют доступные ресурсы на предмет того, имеются ли отдельные источники, содержащие достаточно полный ответ на запрос, или требуется интеграция нескольких ресурсов. Описание этих программных средств выходит за рамки данной работы и является предметом обсуждения в [11].

На рисунке 4 представлена упрощенная схема работы инструментального NuCoBoShell, ориентированного на поиск пертинентной информации посредством виртуальной смысловой интеграции данных из распределенных разнородных ресурсов и позволяющего адаптироваться к новым источникам данных и/или к новым предметным областям за счет пополнения общего репозитория онтологий, управляющих функционированием системы, без необходимости внесения изменений в исходный программный код. Представленный на рис. 4 репозиторий онтологий является частью Реестра, указанного на рис. 3, при этом в нашем подходе к реализации виртуальной интеграции распределенных ресурсов доступ к репозиторию онтологий организован унифицированным образом по аналогии с доступом к другим ресурсам системы, включая доступ к исходным текстовым документам.

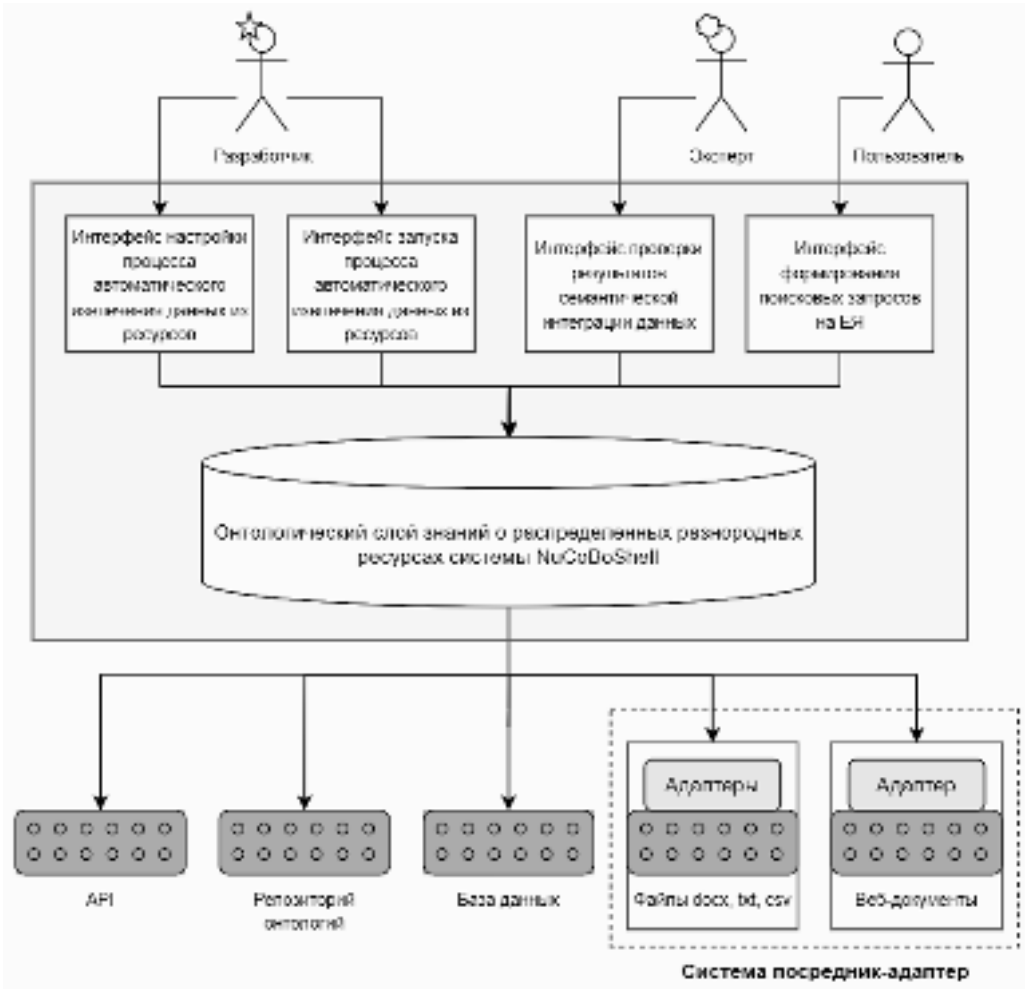

Рис. 4. *Обобщенная архитектура NuCoBoShell*

Для реализации подхода к виртуальной интеграции на основе предлагаемой концепции достаточно использования так называемых "легковесных" (англ., lightweight) онтологий, которые состоят из описания понятий предметной области с указанием их свойств и взаимосвязей с другими понятиями, но не включают спецификацию аксиом. Помимо того, что "тяжеловесная" онтология (англ., heavyweight) включает контент, аналогичный "легко-

весной", она включает в себя также явную спецификацию аксиом на основе некоторой дескриптивной модели, например декларацию всякого рода правил и ограничений. Это позволяет не только получить более богатую семантическую модель онтологии, но и выводить из имеющейся онтологии новые знания с использованием стандартных средств логического вывода, например, языка SPARQL (подробнее см. [12, 13]). Однако, как показывает наш опыт (см., например [11, 14–17] и библиографию к ним), использование легковесных онтологий позволяет унифицированным образом расширять возможности системы и адаптировать их под новые требования и/или изменения в предметной области для решения широкого круга задач из разных областей, если визуальный редактор онтологий позволяет, как в нашем случае [6], хранить в вершинах и дугах онтологии атрибуты с описанием нужных метаданных.

Согласно Дэвису [13] и др. исследователям, из-за того, что легковесные онтологии не включают описание аксиоматики и обычно состоят из иерархии понятий и набора отношений, существующих между этими понятиями, их легче понимать, адаптировать, управлять, обновлять и использовать. Как следствие, создание легковесной онтологии требует меньше времени и усилий, чем создание тяжеловесной онтологии, что делает их более доступными как для широкого круга ИТ-разработчиков, так и экспертов в других предметных областях, не являющихся ИТ-специалистами, например в области биомедицины. Как уже отмечалось, для автоматизации построения легковесных онтологий мы на этапе создания онтологий используем также сервисы генеративного ИИ [11], но по известным причинам не применяем их для интеграции разнородных ресурсов (см. раздел 2 настоящей статьи) и не встраиваем их в инфраструктуру NuCoBoShell.

Подчеркнем, что наш подход предполагает при необходимости использование вместо одной онтологии целого семейства взаимосвязанных онтологий, причем все они имеют общую модель и поддерживают один и тот же набор типов связей между понятиями. В частности, для интерпретации взаимосвязей типа "is_a" (обратное отношение к "класс–подкласс") и "a_part_of" ("часть–целое") используется одна и та же функция для разных онтологий. Это позволяет применять один интерпретатор для разных онтологий (подробнее см. [6]).

В соответствии с технологией фабрик данных инструментальное окружение NuCoBoShell реализуется на принципах микросервисной архитектуры. Использование микросервисов позволяет унифицированным образом адаптировать систему к различным предметным областям и распределенным разнородным ресурсам, а также конфигурировать и подключать независимые программные модули для обработки данных внутри системы. На рис. 5 представлена диаграмма компонентов инструментального окружения NuCoBoShell.

Благодаря микросервисной архитектуре и онтологически управляемому решению, добавление новых сервисов в систему унифицировано и не требует внесения изменений в программный код других компонентов: достаточно пополнить репозиторий онтологий новыми метаданными о подключаемых сервисах и ресурсах. Это позволит в перспективе обогатить NuCoBoShell дополнительными средствами, в первую очередь, продвинутыми средствами визуальной аналитики. Создание онтологий и просмотр с целью валидации результатов генерации онтологий средствами генеративного ИИ выполняются в среде разработанного в ПГНИУ визуального редактора онтологий ОНТОЛИС 2.0 (см. [6] и библиографию к этой статье).

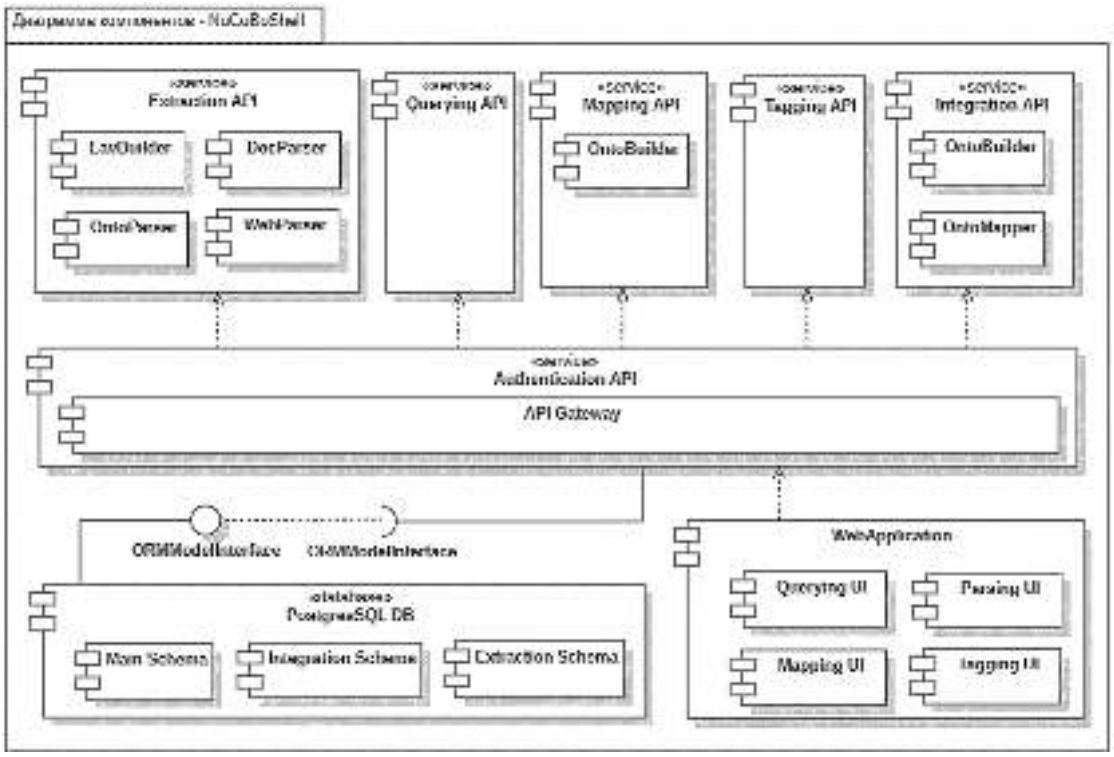

Рис. 5. *Диаграмма компонентов NuCoBoShell*

Заключение

В статье на основе сравнения разных подходов к интеграции распределенных разнородных ресурсов делается обоснованный вывод в пользу интеграции на принципах федерализации и построения виртуальных хранилищ с использованием технологий интеллектуальных фабрик данных без физического копирования данных. Описан новый управляемый легковесными онтологиями подход к реализации виртуальных хранилищ на принципах федерализации для семантической интеграции данных из различных ресурсов, содержащих коллекции документов как в текстовых, так и в веб-форматах, безотносительно их местоположения: в традиционных хранилищах данных, в сети Интернет или на локальном компьютере пользователя. Этот подход позволяет в ситуации, когда в отдельных ресурсах нет полного ответа на поставленный вопрос, а использование современных средств генеративного ИИ затруднено или невозможно по соображениям безопасности данных или другим описанным в статье причинам, генерировать более полные, по сравнению с традиционными поисковыми системами, пертинентные ответы на ЕЯ-запросы пользователей.

Реализация предложенного подхода в рамках инструментального окружения NuCoBoShell на принципах микросервисной архитектуры и модельно-ориентированного подхода с использованием онтологической модели позволяет унифицированным образом адаптировать систему к семантической обработке данных из разных предметных областей и добавлять новые функциональные возможности без внесения изменений в исходный программный код.

Данная работа входит в серию авторских статей, посвященных вопросам онтологически управляемой виртуальной семантической интеграции текстовых данных.

Помимо данной работы, по указанной тематике нами опубликована работа [11], освящающая вопросы применения методов визуального анализа данных для автоматизированного выявления потребности в семантической интеграции данных и готовится к публикации статья, посвященная более подробному описанию разрабатываемых высокоуровневых средств визуального анализа данных для автоматизации работ по подбору наиболее адекватных специфике конкретной предметной области метрик семантической близости понятий и комплексной проверке качества построенных онтологий.

Список источников

1. *Тузовский А.Ф., Ямпольский В.З.* Интеграция информации с использованием технологий semantic web // Проблемы информатики. 2011. № 2. С. 51–58.
2. *Ballard C.* IBM Informix: Integration through data federation / C. Ballard, N. Davies, M. Gavazzi, J. Stephani, M. Lurie // IBM International Technical Support Organizat, 2003. 270 p. URL: http://www.iiug.org/library/ids/technical/sg247032.pdf (дата обращения: 30.06.2024).
3. *Patel A., Debnath, N.C., Bhushan, B. (Eds.).* Semantic Web Technologies: Research and Applications (1st ed.). CRC Press. 2022. 404 p. DOI: 10.1201/9781003309420.
4. *Gruber T.R.* A Translation approach to portable ontology specifications // Knowledge Acquisition. 1993. Vol. 5, № 2. P. 199–220.
5. *Большакова Е.И.* Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Е.И. Большакова, К.В. Воронцов, Н.Э. Ефремова, Э.С. Клышинский, Н.В. Лукашевич, А.С. Сапин М.: НИУ ВШЭ, 2017. 269 с.
6. *Chuprina S.I.* Using Data Fabric Architecture to Create Personalized Visual Analytics Systems in the Field of Digital Medicine // Scientific visualization. 2023. Vol. 15(5). P. 50–63. DOI: 10.26583/sv.15.5.05.
7. *Найденова, К.А., Невзорова О.А.* Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований // Учен. зап. Казан. ун-та. Серия Физико-матем. науки. 2008. № 4. С. 5–24.
8. *Нурутдинов А.Р., Латыпов Р.Х.* Перспективы биоинспирированного подхода в разработке систем искусственного интеллекта (обзор тенденций) // Учен. зап. Казан. ун-та. Сер. Физико-матем. науки. 2022. Т. 164, кн. 2–3. С. 244–265. DOI: 10.26907/2541-7746.2022.2-3.244-265.
9. Semantic Web W3C. URL: https://www.w3.org/standards/ (дата обращения: 30.06.2024).
10. *Calvanese D., De Giacomo G., Lenzerini M.* Ontology of integration and integration of ontologies // Proc. of the 14th Int. Workshop on Description Logics (DL 2001). 1-3 August 2001, Stanford, CA, USA. Vol. 49. P. 10–19.
11. *Чуприна С.И., Гимашева К.В.* Применение методов визуального анализа данных для выявления потребности в семантической интеграции данных // Труды междунар. конф. по компьютерной графике и машинному зрению "Графикон 2024".17–19 сентября 2024, Омск. С. 389–402. DOI: 10.25206/978-5-8149-3873-2-2024-389-402.
12. *Gomes-Perez A., Fernandez-Lopez M., Corcho O.* Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web (1st ed.). Springer-Verlag, London. 2004. 403 p.
13. *Davies J.* Lightweight Ontologies // Theory and Applications of Ontology: Computer Applications. 2010. P. 197-229. DOI: 10.1007/978-90-481-8847-5_9.

14. *Ryabinin K., Chuprina S.* Development of ontology-based multiplatform adaptive scientific visualization system // Journal of Computational Science. Elsevier. 2015. Vol. 10. P. 370–381. DOI: 10.1016/j.jocs.2015.03.003.
15. *Ryabinin K., Chuprina S., Belousov K.* Ontology-Driven Automation of IoT-Based Human-Machine Interfaces Development // Computational Science – ICCS 2019 / Edit by J. M. F. Rodrigues. – Cham: Springer International Publishing, 2019. P. 110–124.
16. *Chuprina S.I.* Ontology-Driven Visual Analytics Software Development / S. Chuprina, K. Ryabinin, K. Matkin, D. Koznov// Programming and Computer Software. 2022. T. 48, № 3. P. 208–214. DOI: https://doi.org/10.1134/ S0361768822030033.
17. *Ryabinin K., Chuprina S., Labutin I.* Tackling IoT Interoperability Problems with Ontology-Driven Smart Approach // Science and Global Challenges of the 21st Century - Science and Technology / Edit by A. Rocha, E. Isaeva. Cham: Springer International Publishing. 2022. P. 77–91.

References

1. Tuzovskiy, A. F. and Yampolskiy, V. Z. (2011), "Integration of information using semantic web technologies", *Problemy informatiki*, no. 2, pp. 51-58.
2. Ballard, C., Davies, N., Gavazzi, M., Stephani, J.and Lurie, M. (2003), *IBM Informix: Integration through data federation*, IBM International Technical Support Organizat, USA, available at: http://www.iiug.org/library/ids/technical/sg247032.pdf (Accessed: 30.06.2024).
3. Patel, A., Debnath, N. C. and Bhushan, B. (2022), *Semantic Web Technologies: Research and Applications*, 1st ed., CRC Press, USA, 404 p. DOI: 10.1201/9781003309420.
4. Gruber, T. R. (1993), "A Translation approach to portable ontology specifications", *Knowledge Acquisition*, vol. 5, no. 2, pp. 199-220.
5. Bolshakova, E. I., Vorontsov, K. V., Efremova, N. E., Klyshinskiy, E. S., Lukashevich, N. V. and Sapin, A. S. (2017), *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh* [Automatic text processing in natural language and data analysis], HSE University, Moscow, Russia.
6. Chuprina, S. I. (2023), "Using Data Fabric Architecture to Create Personalized Visual Analytics Systems in the Field of Digital Medicine", *Scientific visualization*, vol. 15(5), pp. 50-63. DOI: 10.26583/sv.15.5.05.
7. Naidenova, X. A. and Nevzorova, O. A. (2008), "Machine Learning for Natural Language Processing: Contemporary State", *Uchehye zapiski Kazanskogo universiteta. Seriya Fiziko-matematicheskie nauki*, no. 4, pp. 5-24.
8. Nurutdinov, A. R. and Latypov, R. Kh. (2022), "Potentials of the bio-inspired approach in the development of artificial intelligence systems (trends review)", *Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki*, vol. 164, no. 2–3, pp. 244-265. DOI:10.26907/2541-7746.2022.2-3.244-265.
9. Semantic Web W3C. URL: https://www.w3.org/standards/ (Accessed: 30.06.2024).
10. Calvanese, D., De Giacomo, G. and Lenzerini, M. (2001), "Ontology of integration and integration of ontologies", *Proceedings of the 14th Int. Workshop on Description Logics (DL 2001)*, Stanford, CA, USA, 1-3 August 2001, vol. 49, pp. 10-19.
11. Chuprina, S. I. and Gimasheva, K. V. (2024), "Using visual data analysis methods to identify the need for semantic data integration", *Trudy Mezhdunarodnoy konferentsii po komputernoy grafike b mashinnomu zreniyu "GraphiCon"*, Omsk, Russia, 17–19 September 2024, pp. 389–402. DOI: 10.25206/978-5-8149-3873-2-2024-389-402.

12. Gomes-Perez, A., Fernandez-Lopez, M. and Corcho, O. (2004), *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web,* 1st ed., Springer-Verlag, London. 403 p.
13. Davies, J. (2010), "Lightweight Ontologies", *Theory and Applications of Ontology: Computer Applications*, pp. 197–229. DOI: 10.1007/978-90-481-8847-5_9.
14. Ryabinin, K. and Chuprina, S. (2015), "Development of ontology-based multiplatform adaptive scientific visualization system", *Journal of Computational Scienc*e, Elsevier, vol. 10, pp. 370-381. DOI: 10.1016/j.jocs.2015.03.003.
15. Ryabinin, K., Chuprina, S. and Belousov, K. (2019), "Ontology-Driven Automation of IoT-Based Human-Machine Interfaces Development", in Rodrigues, J. (eds), *Computational Science – ICCS 2019*, ICCS 2019, Lecture Notes in Computer Science, vol. 11540, Springer, Cham, pp. 110-124. DOI: https://doi.org/10.1007/978-3-030-22750-0_9.
16. Chuprina, S., Ryabinin, K., Matkin, K. and Koznov, D. (2022), "Ontology-Driven Visual Analytics Software Development", *Programming and Computer Software*, vol. 48, no. 3, pp. 208-214. DOI: https://doi.org/10.1134/S0361768822030033.
17. Ryabinin, K., Chuprina, S. and Labutin, I. (2022), "Tackling IoT Interoperability Problems with Ontology-Driven Smart Approach", in Rocha, A., Isaeva, E. (eds), *Science and Global Challenges of the 21st Century - Science and Technology*, Perm Forum 2021, Lecture Notes in Networks and Systems, vol. 342, Springer, Cham, pp. 77-91. DOI: https://doi.org/10.1007/978-3-030-89477-1_9.

Информация об авторах:

С. И. Чуприна – кандидат физико-математических наук; до сентября 2024 – профессор кафедры математического обеспечения вычислительных систем Пермского государственного национального исследовательского университета (614068, Россия, г. Пермь, ул. Букирева, 15); с ноября 2024 – доцент кафедры прикладной информатики, информационных систем и технологий Пермского государственного гуманитарно-педагогического университета (614990, Россия , г. Пермь, ул. Сибирская, 24); почетный работник высшего профессионального образования РФ, член-корреспондент Международной академии информатизации, AuthorID: 11124;

К. В. Гимашева – магистрант кафедры математического обеспечения вычислительных систем Пермского государственного национального исследовательского университета (614068, Россия, г. Пермь, ул. Букирева, 15), AuthorID: 1178188.

Information about the authors:

S. I. Chuprina – PhD in Physics and Mathematics; before September 2024 – Prof. at the Dept. of Computer Science, Perm State University (15, Bukireva St., Perm, Russia, 614068); since November 2024 – Associate Professor at the Dept. of Applied Mathematics, Information Systems and Technologies, Perm State Humanitarian Pedagogical University (24, Sibirskaya St., Perm, Russia, 614990); Honorary Worker of Higher Professional Education of the Russian Federation, Corresponding Member of the International Academy of Informatization, AuthorID: 11124;

K. V. Gimasheva – Master's student of Computer Science Dept. at Perm State University (15, Bukireva St., Perm, Russia, 614068), AuthorID: 1178188.