На правах рукописи

Николаев Константин Сергеевич

МЕТОДЫ И АЛГОРИТМЫ ОБРАБОТКИ МАТЕМАТИЧЕСКОГО КОНТЕНТА НА ОСНОВЕ ТЕХНОЛОГИЙ СЕМАНТИЧЕСКОГО ВЕБА

Специальность 2.3.8 – «Информатика и информационные процессы»
(технические науки)

Автореферат диссертации
на соискание ученой степени
кандидата технических наук

Казань – 2024

Работа выполнена на кафедре системного анализа и информационных технологий Института вычислительной математики и информационных технологий ФГАОУ ВО «Казанский (Приволжский) федеральный университет».

Научный руководитель: **Невзорова Ольга Авенировна,**

кандидат технических наук, доцент кафедры информационных систем Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета

Официальные оппоненты: **Шалфеева Елена Арефьевна,**

доктор технических наук, ведущий научный сотрудник лаборатории Интеллектуальных систем Института автоматики и процессов управления Дальневосточного отделения РАН, г. Владивосток

Загорулько Юрий Алексеевич,

кандидат технических наук, заведующий лабораторией искусственного интеллекта Института систем информатики им. А.П. Ершова СО РАН, г. Новосибирск

Ведущая организация: Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр «Кольский научный центр Российской академии наук» (ФИЦ КНЦ РАН), Мурманская обл., г. Апатиты.

Защита состоится «17» мая 2024 года в 15:30 на заседании диссертационного совета КФУ.023.2 на базе ФГАОУ ВО «Казанский (Приволжский) федеральный университет» по адресу: 420008, Казань, ул. Кремлевская, 35, ауд. 1310.

С диссертацией можно ознакомиться в библиотеке ФГАОУ ВО «Казанский (Приволжский) федеральный университет» по адресу: 420008, Казань, ул. Кремлевская, 35.

Сведения о защите, автореферат и диссертация размещены на официальных сайтах Высшей аттестационной комиссии при Министерстве науки и высшего образования Российской Федерации (https://vak.minobrnauki.gov.ru/) и ФГАОУ ВО «Казанский (Приволжский) федеральный университет» (https://kpfu.ru/).

Автореферат разослан « » 2024 года.

Ученый секретарь диссертационного совета
КФУ.023.2, канд. физ.-мат. наук, доцент Липачев Евгений Константинович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Автоматическая обработка цифрового контента является востребованной задачей в реалиях современного мира, особенно в задачах поиска информации, обогащения семантики и интерпретации данных. Сфера применения цифрового математического контента весьма широка: в первую очередь электронные научные журналы и электронные математические библиотеки (Math-Net.Ru (А. Б. Жижченко, А. Д. Изаак), EuDML (J. Rákosník, P. Stanchev)), научные поисковые системы (MathSciNet, zbMATH), также различные системы дистанционного математического образования (GeoGebra, Maplesoft, Carnegie Learning). Обработка математического контента позволяет структурировать и определить семантику математических документов (как научных, так и образовательных), улучшить процесс поиска релевантных документов для исследователей, расширить возможности по управлению математическим контентом в системах дистанционного образования.

Технологии Semantic Web являются основным стеком технологий для представления связанных данных во Всемирной сети и открывают доступ к чётко структурированной информации для любых приложений, независимо от платформы и независимо от языков программирования. В стек технологий Semantic Web входят онтологии, а также приложения и пользовательские интерфейсы.

Технологии семантического веба часто используются при визуализации, связывании и представления математических документов. Под математическими документами здесь понимаются не только большие тексты, содержащие математическую информацию, но и отдельные формульные фрагменты. Основным подходом к формированию семантически обогащенных математических документов является выделение в них объектов и семантических связей. Для этого применяются технологии, основанные на Resource Description Framework (далее – RDF) – расширяемой модели отношений между объектами, разработанной для внедрения информации (семантики) во Всемирную сеть.

Математические данные по своей природе сложны для категоризации без дополнительной информации. Эта проблема свойственна многим видам данных сети, поэтому был разработан стандарт Web Ontology Language (далее – OWL), позволяющий выполнить процесс категоризации данных на определенные классы и подклассы. Технологии, основанные на этом стандарте, позволяют соотнести данные с онтологическими значениями: другими словами, сформировать множественные математические иерархии, подходящие по требованиям различных проектов и задач.

Помимо задач визуализации и хранения математического контента, важными являются задачи автоматического выполнения и манипуляции математическими формулами в сети. Так как технологии всемирной сети оперируют унифицированными идентификаторами ресурса (далее – URI), возникает проблема создания новых URI в рамках конкретного документа/сервера/проекта, что приводит к изоляции кластеров математических данных в сети. Стандарт XML Query поддерживает простой поиск и извлечение данных из структуры документа, что делает схему XML более предпочтительной для представления простых математических документов. Но для представления данных со сложной структурой и использующих высокоуровневые и редкие математические отношения такой подход не оправдан, так как конвертация из RDF в XML пройдет с потерями важной информации.

Именно здесь возникает задача поиска математических документов по заданным понятиям и формулам, так как математические понятия и формулы оптимальнее хранить именно в формате RDF и онтологиях. Задача поиска математических документов по понятиям и формулам, и задача поиска научных статей, в частности, заметно отличается от задачи классического полнотекстового поиска, так как научные статьи содержат формулы, поиск по которым часто улучшает поисковую выдачу. Существуют работы, опирающиеся на анализ структуры математических формул в формате MathML и производящие поиск по заданной структуре формулы (Kohlhase). Данный подход хорош своей универсальностью, но не учитывает семантику документов, поэтому не подходит для поиска математических документов по названиям объектов и формул. Проблема состоит в том, что не всегда у формулы присутствует явное определение в тексте (например, в случае, когда формула общеизвестная или была определена ранее, или автор ссылается на формулу из другой своей работы). В таком случае необходимо определять семантику формулы через ее компоненты (функции, переменные). По такому представлению формулы через ее компоненты появляется возможность поиска формул по искомым переменным и функциям, и, как следствие, поиск документов по таким переменным и функциям. Именно поэтому требуются алгоритмы, способные с некоторой степенью точности определить название формулы, ее тип и, если это возможно, ее компоненты и их расположение в тексте.

Семантические технологии также применимы к образовательным задачам, а именно в генерации и визуализации тестовых заданий. Интеллектуализация систем дистанционного образования особенно актуальна в наше время в связи с заметным увеличением доли цифрового обучения в российской системе образования. Классическая (офлайн) методология проектирования учебных программ слабо применима к специфике дистанционных образовательных технологий. По этой причине необходима разработка новых образовательных методик, в которых используются семантические технологии, что позволяет решать задачу персонализации процесса обучения. Часть описанных проблем решена в рамках настоящей работы.

Разработанные методы и алгоритмы способствуют как повышению качества онтологий, так и формированию и представлению в технологиях Semantic Web документов для применения в образовательных и научных задачах.

Цели и задачи исследования. Целью работы является разработка методов и алгоритмов извлечения, хранения и обработки математического контента, для разработки комплекса сервисов, применяемых в образовательных и научных задачах.

Основная **задача**, решаемая в данной работе, заключается в создании комплекса сервисов на основе разработанных методов и алгоритмов анализа математического контента. Данная задача разбивается на следующее множество **подзадач**:

1. Провести исследование современного состояния семантических методов обработки математического контента.

2. Разработать архитектуру экосистемы OntoMathEdu, которая направлена на применение семантических технологий в решении научных и научно-образовательных задач, и включает взаимосвязанные сервисы для обработки математических документов (сервис выделения математических

понятий, сервис обогащения онтологии, сервис семантического аннотирования формул в PDF документах, сервис генерации тестовых заданий, сервис визуализации подграфов семантических сетей). Разработать методы и алгоритмы для извлечения, хранения и обработки математического контента и их реализацию в форме сервисов на основе семантических технологий и специализированных математических ресурсов.

3. Разработать метод автоматического аннотирования математических понятий в образовательных математических текстах на основе онтологии математического знания, в котором учитываются структурные свойства образовательных документов и применяются встроенные в онтологию дидактические отношения.

4. Разработать метод полуавтоматической оценки структурной полноты горизонтальных связей онтологии на основе анализа структурных свойств графа классов онтологии, включающий этап формирования рекомендаций для эксперта и последующую экспертную оценку по принятию рекомендаций о внедрении новых отношений между концептами онтологии, позволяющий повысить качество онтологии путем улучшения её горизонтальной связности.

5. Разработать метод автоматического извлечения и семантического аннотирования математических формул в PDF документах, в котором учитываются структурные свойства научных математических документов, производится связывание компонентов формул с математическими понятиями в тексте с целью формирования семантического представления формулы.

6. Разработать алгоритмы для визуализации и генерации образовательного контента, включая алгоритм визуализации подграфов семантических сетей для обеспечения совместной работы с методом аннотирования математических понятий в образовательных математических текстах, и алгоритм генерации тестовых заданий и заданий с вводом ответа на основе понятий и отношений онтологии $\text{OntoMath}^{\text{Edu}}$.

7. Разработать прототип комплекса семантических сервисов для обработки математического контента на основе разработанных методов и алгоритмов.

8. Реализовать встраивание и адаптацию разработанного комплекса семантических сервисов в дистанционный курс общеобразовательной геометрии на платформе дистанционного образования КФУ.

9. Провести экспериментальные исследования эффективности разработанных методов семантической обработки математического контента учебного курса и сформулировать рекомендации по дальнейшим исследованиям.

Основные положения, выносимые на защиту:

1. Предложена структура экосистемы OntoMathEdu, которая включает в своем составе онтологию общеобразовательной математики $\text{OntoMath}^{\text{Edu}}$ и взаимосвязанные сервисы для обработки математического научно-образовательного контента, разработанные на основе семантических технологий и специализированных математических ресурсов, что позволяет использовать преимущества семантических и онтологических технологий в обучении математике.

2. Предложен новый метод автоматического аннотирования математических понятий в образовательных математических текстах на основе онтологии математического знания, в котором

учитываются структурные свойства образовательных документов и применяются встроенные в онтологию дидактические отношения, позволяющий обогатить образовательные документы ссылками на результаты работы алгоритма визуализации подграфов семантических сетей.

3. Предложен новый метод полуавтоматической оценки структурной полноты горизонтальных связей онтологии, включающий этап автоматического формирования рекомендаций для эксперта и последующую экспертную оценку по принятию рекомендаций о внедрении новых отношений между концептами онтологии, позволяющий повысить качество онтологии путем улучшения её горизонтальной связности.

4. Предложен новый метод автоматического извлечения и семантического аннотирования математических формул в PDF документах, в котором учитываются структурные свойства научных математических документов, производится связывание компонентов формул с математическими понятиями в тексте с целью формирования семантического представления формулы, позволяющий сформировать размеченную коллекцию формул для семантического поиска документов.

Научная новизна работы заключается в разработке основанных на семантических технологиях методов обработки и использования научного математического контента в решении прикладных задач в науке и образовании.

1. Впервые представлена структура экосистемы OntoMathEdu, направленной на применение семантических технологий в задачах математического образования.

2. Разработан новый метод автоматического аннотирования математических понятий в образовательных математических текстах, в котором учитываются дидактические отношения между понятиями, встроенные в онтологию $\text{OntoMath}^{\text{Edu}}$ (пререквизиты, образовательные уровни, образовательные системы), и производится разметка понятий в соответствии с объемом изучаемых понятий образовательного уровня.

3. Впервые разработан метод полуавтоматической оценки структурной полноты горизонтальных связей онтологии, который позволяет определить качество онтологии на уровне связанности понятий, используя данные методических заданий по решению задач в предметной области.

4. Разработан новый метод автоматического извлечения и семантического аннотирования математических формул в PDF документах, в котором производится связывание компонентов формул с математическими понятиями в тексте с целью формирования семантического представления формулы для дальнейшего применения в семантическом поисковике, в отличие от методов выделения формул в тексте, основанных на известных математических нотациях (TeX, LaTeX, MathML).

Разработанные методы работы с математическим контентом применяются при разработке учебных курсов по математике в системе дистанционного образования, а также в поисковых сервисах электронных математических библиотек.

Научная и практическая значимость работы. В данной работе разработан прототип комплекса сервисов по автоматической обработке математических документов с использованием семантических технологий. Разработанный комплекс сервисов может быть использован в качестве модулей в составе научных библиотечных систем, а также в системах дистанционного обучения. В данный момент сервис визуализации подграфов семантических сетей и сервис автоматической раз-

метки математических документов активно используются при подготовке цифровых образовательных ресурсов по общеобразовательной геометрии. Исследования, отраженные в диссертационной работе, проведены в рамках научно-исследовательских работ при поддержке Российского научного фонда (РНФ): проект №21-11-00105 «Lobachevskii-DML как составляющая Всемирной цифровой математической библиотеки» (Николаев К.С. – основной исполнитель) и в рамках государственного задания по теме FNEF-2022-0014.

Соответствие паспорту специальности. Работа выполнена в рамках направлений области исследований «Техническое обеспечение информационных систем и процессов, в том числе новые технические средства сбора, хранения, передачи и представления информации. Комплексы технических средств, обеспечивающих функционирование информационных систем и процессов, накопления и оптимального использования информационных ресурсов» и «Лингвистическое обеспечение информационных систем и процессов. Методы и средства проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и иных лексических комплексов. Методы семантического, синтаксического и прагматического анализа текстовой информации для представления в базах данных и организации интерфейсов информационных систем с пользователями» паспорта специальности 2.3.8 — Информатика и информационные процессы.

Достоверность. Степень достоверности полученных в работе результатов обеспечивается строгостью постановки задач и применением научной методологии разработки, соответствием разработанных программных средств установленным стандартам программной инженерии, использованием системного подхода к построению программного комплекса, верификацией разработанных программных изделий и проведенными экспериментами.

Эксперименты по обработке математического контента для образовательных приложений проводились на основе разработанных лекций по предметам математического цикла Института математики и механики Казанского (Приволжского) федерального университета, с применением математических онтологий: онтологии профессиональной математики $\text{OntoMath}^{\text{Pro}}$ и образовательной математической онтологии $\text{OntoMath}^{\text{Edu}}$, а также коллекции научных статей журнала «Известия высших учебных заведений. Математика».

Апробация результатов работы. Основные результаты работы докладывались на следующих международных конференциях в период с 2018 по 2023 гг.: Международная научная конференция «Электронная Казань» (Казань, 2018; Казань, 2019), Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL) (Москва, 2018; Казань, 2019; Воронеж, 2020; Москва, 2023), Международный форум по математическому образованию – IFME (Казань, 2021; Казань, 2022), International Technology, Education and Development Conference (Валенсия, 2020; Валенсия, 2021), International Symposium on Computer Science, Digital Economy and Intelligent Systems (Москва, 2019).

Публикации. Основные результаты по теме работы изложены в 14 печатных изданиях, 6 – в журналах, индексируемых Scopus, 8 — в прочих изданиях, и получено 2 свидетельства о государственной регистрации программы для ЭВМ.

Личный вклад. Все выносимые на защиту результаты получены автором данной работы лично. В опубликованных совместных работах постановка задачи осуществлялась совместными усилиями соавторов при непосредственном участии автора данной работы.

Объем и структура диссертации. Диссертация состоит из введения, 3 глав, заключения и приложений. Полный объем работы составляет 118 страниц, включая 38 рисунков, 13 таблиц, 3 приложения.

Содержание работы

Во **введении** обоснована актуальность исследований, проведенных в рамках диссертационного исследования, приводится обзор научной литературы по изучаемой проблеме, сформулированы цели, поставлены задачи работы, обоснованы научная новизна и практическая значимость работы.

Первая глава посвящена обзору существующих методов и технологий обработки математического контента. Основной акцент сделан на технологиях, позволяющих производить машинную обработку веб-документов. Дано описание технологии URI (унифицированный идентификатор ресурса), позволяющей расширить содержимое обычного веб-документа ссылками на объекты реального мира и на абстрактные понятия. Также дано описание языка RDF, используемого в качестве основного способа представления семантических данных.

Описан проект формирования связанных данных Linked Data (https://www.w3.org/DesignIssues/LinkedData), представлены основные принципы формирования связанных данных. Дано описание онтологий, выступающих в роли словарей понятий разных предметных областей, а также в качестве моделей для отображения отношений между объектами в рамках некоторой информационной сферы. Дано описание семантического веба и открытых связанных данных, перечислены проекты, направленные на применение открытых связанных данных в сфере образования. Представлены основные принципы обработки математических документов. Описаны особенности математических документов, а именно наличие структуры, общей для некоторой коллекции документов (например, в рамках электронного научного журнала, или в пределах одного образовательного курса). Описана онтология Mocassin, разработанная В.Д. Соловьевым и Н. Г. Жильцовым, отражающая основные структурные элементы научных математических документов. Описаны языки представления математических знаний с возможностями автоматического разрешения математических утверждений разной степени формализованности: MathML, OpenMath, OMDoc, MathLang.

Вторая глава посвящена разработке методов для автоматического извлечения, обработки, хранения и использования математических понятий. Дано описание экосистемы OntoMathEdu, в рамках которой производится разработка методов.

Экосистема OntoMathEdu разработана в Казанском федеральном университете и включает набор цифровых компонентов, в том числе онтологию общеобразовательной математики OntoMath^{Edu} и специализированные сервисы, применяемые для персонализации процесса дистанционного обучения общеобразовательной математике. Экосистема включает в своем составе онтологию общеобразовательной математики OntoMath^{Edu}, содержащую иерархию понятий предметной области, иерархию материализованных отношений, образовательные уровни, образовательные системы. Структура экосистемы OntoMathEdu приведена на Рис.1. Компоненты экосистемы используют пре-

имущества образовательных онтологий (дидактические отношения, систему образовательных уровней и образовательных систем). Все компоненты экосистемы взаимосвязаны. Так, сервис визуализации подграфов семантических сетей внедряется в сервис выделения математических понятий в документах посредством гиперссылок, сервис выделения математических понятий в документах применяется при распознавании понятий в сервисе семантического аннотирования формул в PDF документах, сервис генерации тестовых заданий использует записи из хранилища формул и применяется для пополнения коллекции тестовых заданий.

В качестве технологической поддержки экосистемы выступает компонент «Интеллектуальная цифровая образовательная платформа для школьной математики», который обеспечивает доступ к сервисам экосистемы.

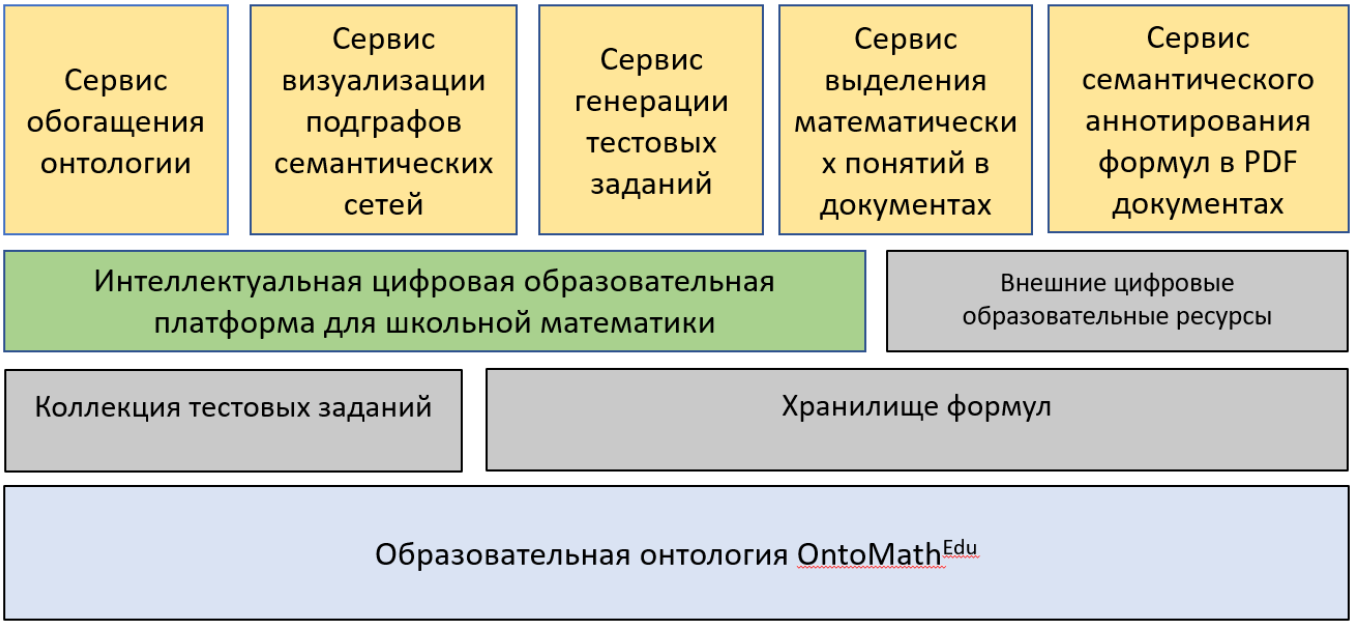

Рис.1. Структура экосистемы OntoMath^{Edu}

Далее приведено описание коллекции разработанных методов, интегрируемых в сервисы экосистемы, а именно:

1. Метод автоматического аннотирования математических понятий в образовательных математических текстах на основе онтологии математического знания.

В данном методе онтология используется в качестве источника понятий. Произвольного списка математических терминов недостаточно для задачи распознавания объектов в образовательных документах, так как результаты разметки в большинстве случаев будут избыточными. В методе учитываются дидактические отношения между понятиями, встроенные в онтологию OntoMath^{Edu} (пререквизиты, образовательные уровни, образовательные системы), и производится разметка понятий в соответствии с объемом изучаемых понятий образовательного уровня. Кроме того, в методе производится разделение документа на структурные элементы на основе разметки HTML и выделение понятий в таких элементах. Такой подход позволяет выделять понятия в более важных элементах текста (например, в постановке задачи). Разметка понятиями из онтологии позволяет отобразить страницу с информацией о понятии, сформированную на основе отношений между концептами в онтологии (онтологические отношения, пререквизиты, связанные внешние ресурсы).

Метод производит нормализацию предложений в математических текстах и названиях концептов онтологии, выделяет из нормализованных предложений цепочки значимых слов различных длин и сравнивает их с названиями понятий в онтологии, определяя наиболее вероятного кандидата из списка понятий онтологии для каждого слова в предложении. Для сравнения множеств слов применена мера Жаккара. Произведена оценка числа порождаемых цепочек для введения ограничений на количество значимых слов в предложении [1, 8] и количество цепочек [1, 255]. Ограничение введено с целью оптимизации работы метода. В случае такого ограничения, скорость работы алгоритма заметно повышается без критичной потери в корректности распознавания.

2. Метод полуавтоматической оценки структурной полноты горизонтальных связей онтологии.

Метод полуавтоматической оценки структурной полноты горизонтальных связей онтологии позволяет оценить горизонтальную связность онтологии на основе анализа текстов заданий. Структурная полнота онтологии по горизонтальным связям (не иерархическим) – это степень горизонтальной связности ее концептов. Метод определяет наличие/отсутствие горизонтальных связей в онтологии $\text{OntoMath}^{\text{Edu}}$ на основе учебных заданий по предметной области (планиметрия). Целью метода является обогащение дидактических отношений, проверка выводимости решения через горизонтальную цепочку понятий для конструирования новых задач. Выявленная цепочка позволяет внести новое элементы в аннотирование концептов онтологии.

Оценка структурной полноты горизонтальных связей дается экспертом на основе построенных с помощью метода цепочек между концептами онтологии. Построение цепочек осуществляется на основе методических заданий по решению задач в заданной предметной области, состоящих из условия (из которых выделяются начальные концепты цепочки) и постановки задачи (из которых выделяются целевые концепты цепочки). Цепочки могут содержать только иерархические отношения, либо комбинацию из иерархических и не иерархических отношений. Экспертной оценке подвергаются пары концептов, для которых существует только иерархическая цепочка концептов, что позволяет оценить необходимость введения новых горизонтальных связей между концептами в цепочке. Количество введенных горизонтальных связей зависит от объема данных (методического материала), применяемого для оценивания.

Метод применим для любой предметной онтологии, при замене файла онтологии и подготовке набора практических заданий по предметной области.

3. Метод автоматического извлечения и семантического аннотирования математических формул в PDF документах.

В данном методе учитывается тот факт, что в математических документах присутствует достаточно четко определенная структура (как минимум в пределах одного научного журнала статьи формируются в едином формате). Такая структурированность позволяет применять улучшенные методы анализа разметки документа. В методе упор делается на связь формулы с ее компонентами, что позволяет формировать ее семантическое представление. Без определения таких связей это невозможно. Следовательно, метод дополняет стандартный этап формирования связи между формулой в тексте и понятия онтологии этапом связывания компонентов формул с самими формулами.

Семантическое аннотирование формулы заключается в выделении из текста математической статьи формулы, отвечающей специальным требованиям, с последующим анализом ее структурных элементов и связыванием выделенных переменных формулы с легендами, данными в текстовом контексте формулы. Введены понятия текстового блока, главной формулы и локальной переменной. Под текстовым блоком понимается абзац, имеющий текстовое и формульное содержание. Под главной формулой понимается формула, выделенная в отдельный абзац. Под локальной переменной понимаются формулы, расположенные в текстовых блоках.

Метод производит выделение блоков в тексте документа, определяет тип блока (текстовый или встроенная формула), производит поиск локальных переменных и понятий онтологии OntoMath^{Pro} в текстовых блоках, производит связывание локальных переменных и встроенных формул, таким образом формируя семантическое представление встроенных формул.

Задачи определения типа блока и поиска локальных переменных выполнялись с помощью анализа разметки документа и посимвольного сканирования текстовых блоков с целью исключения текстовых символов. Связывание локальных переменных и главных формул производилось с помощью выделения обозначений формул с учетом нижних и верхних индексов (например, $f(x)$, $f_0(x,y)$) и поиска совпадающих формул в главной формуле и локальных переменных, находящихся в пределах некоторого текстового окна до и после формулы. Учет расстояния между переменной и формулой был введен по причине того, что некоторые переменные могут переопределяться автором в различных частях документа.

Полученная семантическая разметка документа позволит создать коллекцию документов, пригодных для сервиса семантического поиска формул, являющегося частью набора сервисов цифровой библиотеки Lobachevskii–DML (https://lobachevskii-dml.ru). Сервис семантического поиска формул позволяет находить формулы, содержащие переменную, обозначающую заданное математическое понятие (например, найти все формулы, содержащие понятие «Кольцо»). В поле запроса вводится математическое понятие из онтологии, в выдаче отображаются все вхождения понятия с указанием его обозначения в тексте, формулы, включающей это понятие, а также ссылку на документ, содержащий формулу.

4. **Алгоритм визуализации подграфов семантических сетей.** Данный алгоритм собирает всю доступную информацию о запрашиваемом понятии и представляет в виде интерактивной веб-страницы. Процесс можно разделить на 3 основных этапа: обработка GET-запроса и извлечение информации о запрашиваемом концепте; сбор данных, отображаемых на странице понятия из онтологии OntoMath^{Edu} и базы данных с определениями концептов (определение понятия из учебников по геометрии, значение поля «Уровень образования», список концептов выше и ниже по иерархии понятий в онтологии, список онтологически связанных понятий, список пререквизитов понятия); формирование HTML-страницы на основе собранных данных.

5. **Алгоритм генерации тестовых заданий и заданий с вводом ответа на основе понятий и отношений онтологии OntoMath^{Edu}.** При генерации первого типа заданий используются формулы из корпуса математических формул в формате OpenMath, для формирования задач на вычисление значения неизвестной переменной. Определяется древовидная структура математического утверждения в формате OpenMath, связываются понятия в формуле с концептами онтологии,

производится выбор искомой переменной случайным образом и генерируется текстовое представление задания. В таблице 1 приведены пример записи из корпуса математических формул.

Таблица 1. Пример записи из корпуса формул

Выражение на естественном языке	Формализированное представление выражения	Формализированное представление выражения (LaTeX)	Формализация выражения в OpenMath
Сумма двух острых углов прямоугольного треугольника равна 90°	∠ABC + ∠BAC = 90°, где ABC – прямоугольный треугольник; ∠BCA прямой угол этого треугольника.	$\angle ABC+\angle BAC=90^{\circ}$, где ABC - прямоугольный треугольник; $\angle BCA$ прямой угол этого треугольника.	∀ $ABC_{type:RightTriangle}$, $ABC_{type:Angle}$, $BAC_{type:Angle}$, $BCA_{type:RightAngle}$. isAngleOf (ABC, ABC) ∧ isAngleOf (BAC, ABC) ∧ isAngleOf (BCA, ABC) ∧ A ≠ B ∧ B ≠ C ∧ A ≠ C → degreeMesure (ABC) + degreeMesure (BAC) = 90

При генерации заданий второго типа выделяется поддерево иерархии понятий онтологии OntoMath^{Edu} и формируется задание на восстановление структуры данного поддерева.

В третьей главе описана реализация программных продуктов, в которых применены методы, предложенные во второй главе.

Разработана программная реализация сервиса для управления автоматической разметкой математических документов, визуализации и ручной корректировки результатов разметки, с возможностью генерации веб-документа со ссылками на результаты работы метода визуализации подграфов семантических сетей. На Рис. 2 показан процесс визуализации и редактирования результатов разметки. Также было подсчитано значение F-меры для данного алгоритма. Для каждого документа рассчитывались корректные распознавания (True positive); определенные методом, но не присутствующие в тексте понятия (False positive); пропущенные понятия (False negative). Среднее значение на 9 документах, содержащих информацию о разных задачах в рамках планиметрии, составило 87%. В таблице 2 представлена подробная информация об эксперименте по подсчету F-меры алгоритма.

Разработана программная реализация сервиса автоматического пополнения онтологии. Был проведен эксперимент на фрагменте учебного пособия по планиметрии в 7–9 классе Р.К. Гордина, объемом 348 заданий. Для пар концептов, сформированных из концептов условия и постановки задачи, был произведен поиск иерархических цепочек понятий и цепочек, состоящих из любых отношений. В результате было обнаружено 164 связи между парами понятий, среди которых 150 имели оптимальные связи. Под оптимальными связями понимаются связи, состоящие из любых отношений между понятиями (не только иерархическими).

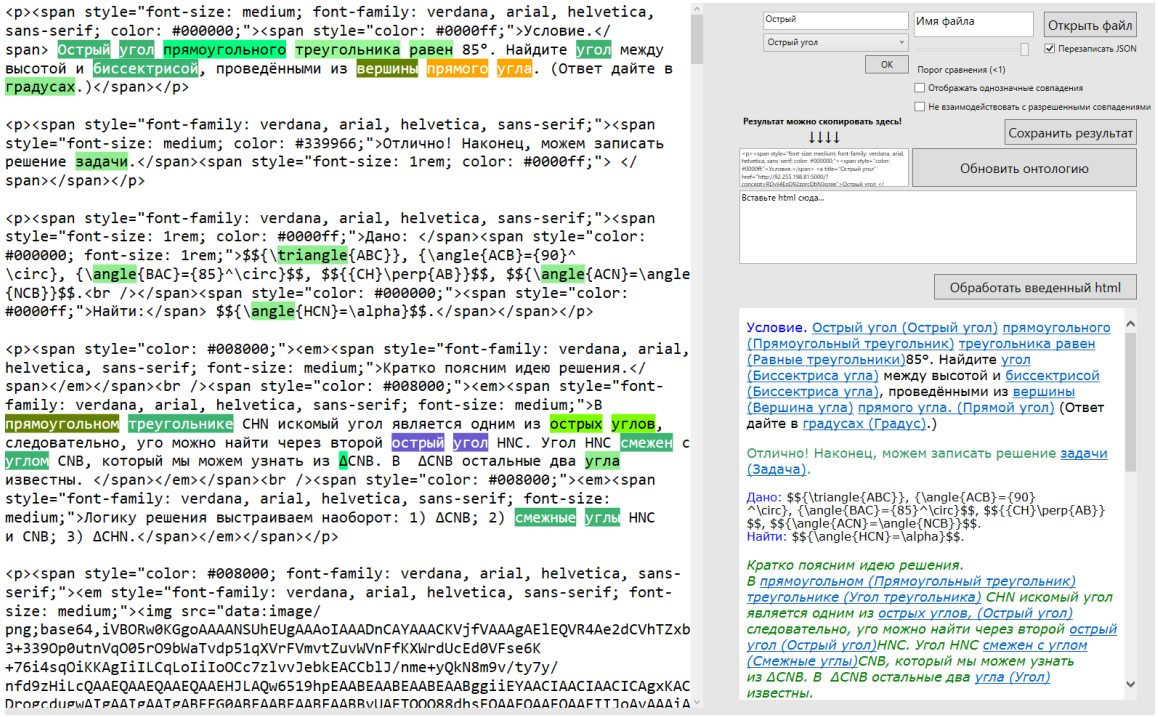

Рис. 2. Интерфейс программы для визуализации и редактирования разметки

Из проведенного эксперимента следует, что для 14 из 164 пар контекстуально близких понятий существуют только цепочки, состоящие из иерархических отношений, что говорит о высокой степени горизонтальной связности онтологии OntoMath^{Edu}. Множество таких цепочек направлено экспертам для определения необходимости обновления онтологии OntoMath^{Edu} прямыми отношениями между понятиями в парах. В таблице 3 приведен пример оптимальной связи, обнаруженной в онтологии OntoMath^{Edu}. В таблице 4 приведен пример неоптимальной связи, обнаруженной в онтологии OntoMath^{Edu}.

Таблица 2. Полнота, точность и F-мера для алгоритма автоматической разметки

Тема документа	**Корректные распознавания (TP)**	**Разметка несуществующих понятий (FP)**	**Пропущенные понятия (FN)**	**Полнота**	**Точность**	**F-score**
Угол треугольника	19	2	2	0,90	0,90	0,90
Трапеция	15	1	7	0,94	0,68	0,79
Площади параллелограмма	17	6	4	0,74	0,81	0,77
Задача о нахождении отношения площадей	24	2	6	0,92	0,80	0,86
Углы, связанные с окружностью	55	15	5	0,79	0,92	0,85
Поворот	33	2	5	0,94	0,87	0,91
Параллельный перенос	49	4	1	0,92	0,98	0,95
Симметрия	50	8	1	0,86	0,98	0,92
Гомотетия	45	6	3	0,88	0,94	0,91
Средние значения				**0,88**	**0,88**	**0,87**

Разработана программная реализация сервиса семантической разметки PDF документов. Сервис семантической разметки PDF-документов реализован в виде скрипта на языке программирования Python, принимающего на вход путь к PDF-файлу или к папке, содержащей набор PDF-документов для пакетной обработки. При первом запуске скрипта происходит установка всех необходимых библиотек. В результате работы скрипта в папке с исходным файлом формируется JSON-файл, содержащий информацию об обнаруженных в документе главных формулах, локальных переменных, и распознанных понятиях в текстовых блоках. Кроме того, в JSON-файле хранится семантическое наполнение каждой главной формулы, представленное в виде множества понятий, содержащихся в этой формуле (результат работы метода семантической разметки документов и связывания локальных переменных и главных формул).

Таблица 3. Пример оптимальной связи между концептами в условии задания и формулировке задания

Текст задания	**Концепт в условии задачи**	**Концепт в постановке задачи**	**Цепочка концептов из любых отношений**	**Иерархическая цепочка**
ABCD - **параллелограмм**, точки M и N лежат на сторонах AD и CD соответственно. AN пересекается с BM в точке P. AN пересекается с CM в точке Q, BN пересекается с CM в точке R. Найдите PM/RN, если ∠PRM = ∠RMN и около PRNM можно **описать окружность**.	«Параллелограмм»	«Описанная окружность»	**«Описанная окружность»** → *суперкласс* → «Описанная окружность многоугольника» → *суперкласс* → «Описанная окружность четырехугольника» → *онтологически зависит* → «Четырехугольник» → *суперкласс* → **«Параллелограмм»**	**«Описанная окружность»** → *подкласс* → «Окружность» → *подкласс* → «Кривая постоянной ширины» → *подкласс* → «Секущая к кривой» → *подкласс* → «Секущая прямая» → *подкласс* → «Прямая» → *подкласс* → «Кривая» → *подкласс* → «Линия» → *подкласс* → «Геометрическая фигура на плоскости» → *суперкласс* → «Часть плоскости» → *суперкласс* → «Ограниченная часть плоскости» → *суперкласс* → «Многоугольник» → *суперкласс* → «Четырехугольник» → *суперкласс* → **«Параллелограмм»**

Таблица 4. Пример неоптимальной связи между концептами в условии задания и формулировке задания

Текст задания	**Концепт в условии задачи**	**Концепт в постановке задачи**	**Цепочка концептов из любых отношений**	**Иерархическая цепочка**
На плоскости даны четыре **точки**. Найдите множество центров **прямоугольников**, образуемых четырьмя прямыми, проходящими соответственно через данные точки.	«Точка»	«Прямоугольник»	**«Прямоугольник»** → *подкласс* → «Параллелограмм» → *подкласс* → «Четырехугольник» → *подкласс* → «Многоугольник» → *подкласс* → «Ограниченная часть плоскости» → *подкласс* → «Ограниченная часть плоскости» → *подкласс* → «Часть плоскости» → *подкласс* → «Геометрическая фигура на плоскости» → *суперкласс* → **«Точка»**	**«Прямоугольник»** → *подкласс* → «Параллелограмм» → *подкласс* → «Четырехугольник» → *подкласс* → «Многоугольник» → *подкласс* → «Ограниченная часть плоскости» → *подкласс* → «Ограниченная часть плоскости» → *подкласс* → «Часть плоскости» → *подкласс* → «Геометрическая фигура на плоскости» → *суперкласс* → **«Точка»**

Для каждой формулы, у которой есть связанные локальные переменные с привязанными понятиями, формируется набор связанных понятий и, как следствие, её семантическое наполнение.

Был проведен эксперимент по оценке точности, полноты и F–меры алгоритма связывания выделенных понятий с главной формулой на коллекции из 9 документов, размещенных на портале Math-Net.Ru (https://www.mathnet.ru/). Средние значения показателей составили 0.89, 0.88, 0.88 соответственно.

Разработан сервис визуализации подграфов семантических сетей с возможностью встраивания в существующие веб-документы для расширения семантического наполнения учебных материалов. Также был реализован функционал встраивания результатов работы сервиса визуализации подграфов семантических сетей в учебный материал по общеобразовательной геометрии. Всплывающее окно с предпросмотром отображается при наведении курсора на соответствующее понятие в тексте.

Реализован сервис генерации тестовых заданий на основе понятий и отношений между понятиями в онтологии. Был проведен эксперимент по генерации коллекции тестовых заданий на основе 30 записей корпуса формул. Для каждой записи было сформировано 50 заданий. Сгенерированные 150 тестовых заданий были встроены в образовательный курс по планиметрии, разработанный в Казанском федеральном университете. В таблице 5 приведены примеры заданий, сгенерированных методом, основанным на корпусе формул математических формул в формате OpenMath.

Таблица 5. Примеры сгенерированных заданий

Математическое выражение	Текст задания	Неизвестная переменная
Площадь прямоугольного треугольника равна половине произведения его катетов	Дано: Прямоугольный треугольник *ABC,* сторона треугольника *b*, сторона треугольника *c*, площадь треугольника $S_{ABC.}$ $b = 10$, $S_{ABC} = 20$ Найти: a.	*a* (сторона треугольника)
Площадь трапеции равна произведению полусуммы ее оснований на высоту	Дано: Трапеция *ABCD* Площадь трапеции S_{ABCD}, $BC = 36$, $AD = 10$, $S_{ABCD} = 28$ Найти: BH.	*BH* (высота трапеции)

Заключение

В ходе выполнения диссертационной работы получены следующие результаты:

1. Разработана архитектура экосистемы OntoMathEdu, которая направлена на применение семантических технологий в решении научных и научно-образовательных задач, и включает взаимосвязанные сервисы для обработки математических документов (сервис выделения математических понятий, сервис обогащения онтологии, сервис семантического аннотирования формул в PDF документах, сервис генерации тестовых заданий, сервис визуализации подграфов семантических сетей). Разработаны методы и алгоритмы для извлечения, хранения и обработки математического контента и их реализация в форме сервисов на основе семантических технологий и специализированных математических ресурсов.

2. Разработан метод автоматического аннотирования математических понятий в образовательных математических текстах на основе онтологии математического знания, в котором учитываются структурные свойства образовательных документов и применяются встроенные в онтологию дидактические отношения, а также его программная реализация.

3. Разработан метод полуавтоматической оценки структурной полноты горизонтальных связей онтологии на основе анализа структурных свойств графа классов онтологии, включающий этап формирования рекомендаций для эксперта и последующую экспертную оценку по принятию рекомендаций о внедрении новых отношений между концептами онтологии, а также его программная реализация.

4. Разработан метод автоматического извлечения и семантического аннотирования математических формул в PDF документах, в котором учитываются структурные свойства научных математических документов, производится связывание компонентов формул с математическими понятиями в тексте с целью формирования семантического представления формулы, а также его программная реализация

5. Разработаны алгоритмы для визуализации и генерации образовательного контента, включая алгоритм визуализации подграфов семантических сетей для обеспечения совместной работы с методом аннотирования математических понятий в образовательных математических текстах, и алгоритм генерации тестовых заданий на основе математической онтологии и коллекции размеченных математических формул.

6. На основе предложенных методов и алгоритмов разработан прототип сервисов экосистемы OntoMathEdu для автоматического извлечения, обработки, хранения и использования математических понятий.

Разработанный комплекс программных средств может существенно упростить работу специалистов при подготовке образовательных курсов по математике, а также улучшить степень усвоения материала лекций. Несомненным плюсом сервисов является возможность их применения для обработки контента различных математических дисциплин при наличии соответствующей предметной онтологии. Все описанные методы нашли применение при подготовке курса дистанционного образования «Планиметрия для школьников», разрабатываемого в Институте математики и механики имени Н.И. Лобачевского Казанского (Приволжского) федерального университета.

Публикации автора по теме диссертации

Статьи в изданиях, входящих в международную базу цитирования Scopus:

1. The semantic context models of mathematical formulas in scientific papers / Nevzorova O., Kirillovich A., Nevzorov V., **Nikolaev K.** // CEUR Workshop Proceedings. – 2018. –Т.2277. – С.33–40. – 0.5 / 0.12 п.л.
2. Towards a semantically annotated corpus of educational mathematical texts in Russian / Nevzorova O., Kirillovich A., **Nikolaev K.,** Galiaskarova K. // CEUR Workshop Proceedings. – 2019. – Т.2523. – С.299–305. – 0.44 / 0.1 п.л.
3. Adapting the LodView RDF browser for navigation over the linguistic linked open data cloud in Russian and the languages of Russia / **Nikolaev K.,** Kirillovich A. // CEUR Workshop Proceedings. – 2020. – Т.2790. – С.350–361. – 0.75 / 0.4 п.л.
4. Towards a parallel informal/formal corpus of educational mathematical texts in Russian / Kirillovich A., Nevzorova O., **Nikolaev K.**, Galiaskarova K.// Advances in Intelligent Systems and Computing. – 2020. – Т.1127. – С.325–334. – 0.62 / 0.25 п.л.
5. A corpus-based approach to elementary geometry knowledge test generation / **Nikolaev K.,** Kirillovich A., Nevzorova O. // INTED Proceedings: 14th International Technology, Education and Development Conference (INTED2020, 2-4 March, 2020, Valencia, Spain). – Valencia: IATED, 2020. – С.6342–6348. – 0.44 / 0.23 п.л.
6. Developing the OntoMathEdu ecosystem for educational applications / **Nikolaev K.S.,** Nevzorova O.A., Falileeva M.V. // CEUR Workshop Proceedings. – 2021. – Т.2910. - С.81–87. – 0.44 / 0.24 п.л.

В прочих научных изданиях:

7. К методу семантического аннотирования математических формул в научных статьях: учет контекстных ограничений / **Николаев К.С.,** Невзорова О.А. // Ученые записки ИСГЗ. – 2018. – Т.16. №1. – С.367–373. – 0.69 / 0.35 п.л.
8. Структура и методы пополнения хранилища формул для онтологии школьной математики / **Николаев К.С.,** Невзорова О.А. // Ученые записки ИСГЗ. – 2019. – Т.17. №1. - С.367–372. – 0.38 / 0.2 п.л.
9. Towards ontology-based software services for teaching school mathematics / Nevzorova O., Shakirova L., Falileeva M., **Nikolaev K.** // INTED Proceedings: 15th International Technology, Education and

Development Conference (INTED2021, 8-9 March, 2021). – Valencia: IATED, 2021. – С.5288–5297. – 0.6 п.л. // 0.2 п.л.

10. Научные издательские сервисы на платформе Lobachevskii-DML / Невзорова О.А., **Николаев К.С.** // Электронные библиотеки. - 2022. - Т.25. №1. - С.42–63. – 1,38 / 0.69 п.л.

11. Семантическое аннотирование математических формул в PDF-документах / Невзорова О.А., **Николаев К.С.** // Электронные библиотеки. - 2022. - Т.25. №6. - С.616–639. – 1,5 / 0.75 п.л.

12. Метод автоматической семантической разметки математических образовательных текстов / **Николаев К.С.,** Невзорова О.А. // Информационные технологии в образовании и науке (ИТОН - 2022) и II International Workshop "Digital Technologies for Teaching and Learning (DTTL)": материалы III Международного форума по математическому образованию, IFME’2022 (Казань, 28 марта – 2 апреля 2022 г.). – Казань: Издательство Казанского университета, 2022. – С.181–190. – 0,62 / 0,43 п.л.

13. Сервис генерации учебных карточек математических понятий для дистанционного курса по геометрии / **Николаев К.С.** // Электронные библиотеки. - 2023. - Т.26. №3. - С.365–377. – 0.81 п.л.

14. Семантические сервисы цифровой экосистемы OntoMath для математического образования / Невзорова О.А., Липачёв Е.К., **Николаев К.С.** // Электронные библиотеки. - 2023. - Т.26. №4. - С.538–564. – 1.69 / 0.56 п.л.

Свидетельство о государственной регистрации программы для ЭВМ:

15. Свидетельство о государственной регистрации программы для ЭВМ №2022680211. Программное обеспечение для управления разметкой математических понятий в образовательных документах / **К.С. Николаев**; заявитель Николаев Константин Сергеевич. – №2022680211; заявление 18.10.2022; опубликовано 28.10.2022, Реестр программ для ЭВМ. – 1 с.;

16. Свидетельство о государственной регистрации программы для ЭВМ № 2024611586. Метод семантической разметки и визуализации математических документов в формате PDF / **К.С. Николаев**; заявитель Николаев Константин Сергеевич. – №2024611586; заявление 27.10.2023; опубликовано 23.01.2024, Реестр программ для ЭВМ. – 1 с.;