УДК 004.021; 004.42

# СЕМАНТИЧЕСКОЕ АННОТИРОВАНИЕ МАТЕМАТИЧЕСКИХ ФОРМУЛ В PDF-ДОКУМЕНТАХ

**О. А. Невзорова[1] [0000-0001-8116-9446], К. С. Николаев[2] [0000-0003-3204-238X]**

[1, 2]*Казанский (Приволжский) федеральный университет, ул. Кремлевская, 35, г. Казань, 420008*

[1]onevzoro@gmail.com, [2]konnikolaeff@yandex.ru

***Аннотация***

Дан обзор существующих решений по семантическому анализу математических документов, а также описан метод автоматического семантического анализа документов, представленных в формате PDF. Разработанный метод позволяет выделять математические формулы внутри документа, анализировать их структуру, выполнять поиск локальных переменных формулы и их определений в документе, а также связывать переменные формулы и понятия из онтологии. Преимуществом разработанного метода перед другими существующими является независимость от разметки исходного PDF-документа, что расширяет область применения метода. Приведены оценки полноты, точности и F-меры для алгоритмов поиска переменных и связывания локальных переменных с формулами. Полученная семантическая разметка документа позволяет создавать коллекции документов, пригодных для сервиса семантического поиска формул, который является одним из сервисов цифровой библиотеки Lobachevskii-DML.

***Ключевые слова****: семантический анализ, PDF, обработка документов, научные журналы, Lobachevskii-DML*

## ВВЕДЕНИЕ

Семантический поиск ориентирован на поиск в цифровых коллекциях таких публикаций, которые представляют собой документы с семантической разметкой компонентов текста. Математические тексты отличаются высокой структурированностью, наличием компонентов фиксированной семантики, таких как теоремы, доказательства, формулы.

Задача поиска документов по математическим формулам является актуальной для проведения научных исследований, подготовки статей, изучения математических дисциплин. В работе [1] описан семантический поисковик по математическим формулам, который использует набор данных, построенный по коллекции научных статей журнала «Известия ВУЗов. Математика» за 1997–2009 гг. В настоящей статье представлены новые улучшенные алгоритмы для построения набора данных для семантического поиска по математическим формулам, что позволит качественно улучшить результаты поиска.

Математический поиск по формулам можно разделить на две категории – поиск формул по структуре и по содержанию. Поиск формул по структуре сводится к получению списка формул, которые частично или полностью совпадают со структурой формулы, заданной в поисковом запросе. Такой подход не учитывает семантики формул.

Более эффективным, но сложным в реализации является поиск математических статей по содержанию формулы. Для определения содержания формулы необходимо выделить переменные, входящие в ее состав, и определить математические понятия, обозначаемые переменными. Дополнительные сложности вносит разнообразие шаблонов оформления математических документов для разных информационных систем научных журналов. В настоящее время наиболее популярными форматами представления математических формул в научных статьях являются: графическое изображение (статьи в формате pdf); формулы в редакторе Microsoft Word; формат LaTeX; формат MathML. Коллекции статей в цифровых математических библиотеках представлены преимущественно в формате PDF. Распознавание текста и, в частности, математических выражений является основной задачей данного исследования. Математические формулы, извлекаемые при распознавании в научных статьях, и их описания являются исходными данными для построения улучшенного набора данных для математического поисковика.

Статья организована следующим образом. Вначале приведен обзор существующих решений, предназначенных для анализа структуры и семантики математических документов с формулами. Затем представлены новые алгоритмы семантического аннотирования математических формул, извлеченных из статей в

формате pdf. В заключении приведены оценки построенных алгоритмов и общие выводы.

## СУЩЕСТВУЮЩИЕ РЕШЕНИЯ ДЛЯ АНАЛИЗА СТРУКТУРЫ И СЕМАНТИКИ МАТЕМАТИЧЕСКИХ ДОКУМЕНТОВ

Существующие методы анализа структуры и семантики математических документов можно разделить на две категории. Методы первой группы ориентированы на извлечение метаданных, анализ логической структуры документа и извлечение внутритекстовых ссылок и цитирований. Методы второй группы дополнительно анализируют определения и формулы, присутствующие в тексте документа. Приведем некоторые работы из первой группы.

В работе [2] авторы предлагают протокол семантической разметки научных документов. Этот протокол основан на обработке XML, сегментации и семантической разметке текста. Наибольшее внимание авторы уделяют библиографии, внутритекстовым ссылкам и названиям разделов. Результатом работы является набор связанных данных для проекта Linked Open Data[1].

Авторы работы [3] также предлагают метод семантической аннотации научных документов. Для обработки документов используется программа PDFX, преобразующая PDF-документ в формат XML. Основными извлекаемыми данными в этой работе является расширенная информация о внутритекстовых цитатах.

В статье [4] авторы описывают процесс семантической разметки сборников CEUR-WS. Для каждого документа формируется информация о логической структуре документа, семантике текста (с помощью классификаторов SVM) и выполняется преобразование в формат RDF для последующей публикации в DBPedia[2].

В [5] приведен метод извлечения информации из документов в форматах PDF и XML. Из статей извлекается информация о логической структуре документа и вспомогательных материалах документа (базовые метаданные, информация о финансировании, названия таблиц и изображений, названии проекта и др.).

[1] https://lod-cloud.net/

[2] https://www.dbpedia.org/

Работы второй группы используют различные подходы (основанные на правилах, векторные представления текста, машинное обучение) для поиска определений математических формул в тексте. К примеру, в [6] авторами использована модель векторного представления слов для извлечения структуры математических документов и определений формул. Такой подход неплохо работает для извлечения структурных компонентов документов, но показывает нелучшие результаты при извлечении определений формул. Авторы указывают, что, по их данным, лишь для 70% всех формул в математических документах явно даются их определения [7].

В [8] обсуждается подход к извлечению формул из документов в формате LaTeX и представлению формул в специальном формате для математического поиска. Кроме того, данный метод способен анализировать текстовые документы с формулами в формате MathML. При этом связывание формул с их определениями не проводится.

В работе [9] проведены эксперименты по учету текстового контекста математических формул для повышения качества их преобразования в машиночитаемые форматы. Важным результатом этой работы является датасет, в котором размечены формулы, их компоненты и понятия, встречающиеся в тексте.

В [10] рассмотрен подход к аннотированию формул в формате XML на основе набора лингвистических шаблонов для установления легенды формульной переменной в тексте.

Большинство методов из работ, упомянутых выше, сильно зависит от входного документа, его формата и структуры. В настоящей статье предложен универсальный метод определения структуры документа в формате PDF и связывания переменных в тексте с главными формулами для определения семантического содержимого документа. Набор данных, построенный на основе разработанных алгоритмов, будет использован для улучшения качества семантического поиска математических формул.

## МЕТОД СЕМАНТИЧЕСКОГО АННОТИРОВАНИЯ ФОРМУЛ В PDF-ДОКУМЕНТЕ

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. Схема разработанного алгоритма семантического аннотирования приведена на Рис. 1.

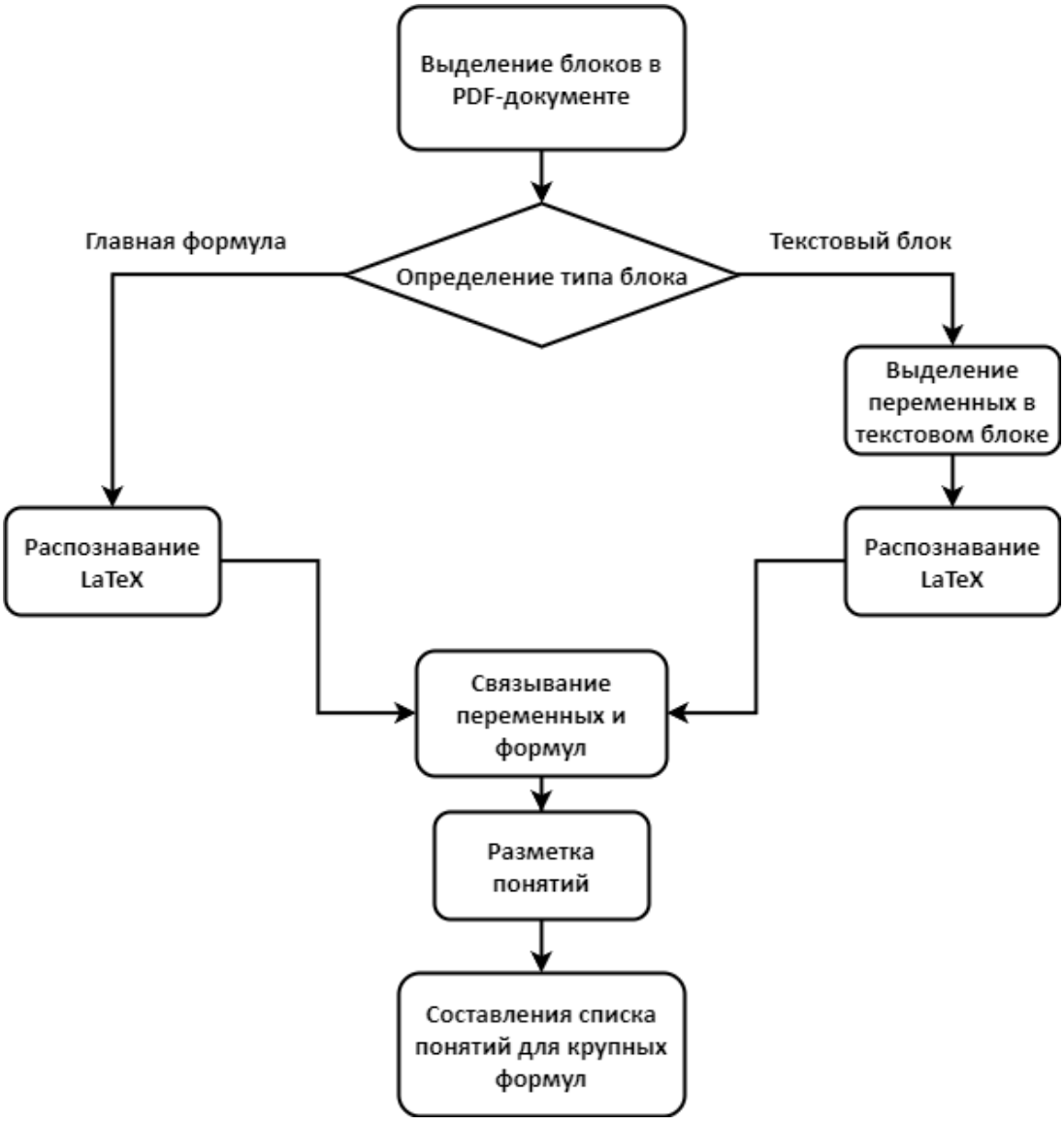



Рис. 1. Схема алгоритма семантического аннотирования PDF-документов

Главная задача семантического аннотирования формул заключается в разработке программного решения, позволяющего выделить набор переменных в формулах математического документа и связать переменные с математическими понятиями, используя математическую онтологию. Полученная семантическая разметка документа позволит создать коллекцию документов, пригодных для сервиса семантического поиска формул, являющегося частью набора сервисов цифровой платформы Lobachevskii-DML.

В качестве тестовой коллекции мы используем набор случайно выбранных документов, размещенных на портале MathNet.ru. Документы в коллекции представлены в виде текстового PDF, что заметно упрощает анализ документа. Примеры, использованные в настоящей работе, содержатся в статьях Р. Р. Кучарова, Ю. Х. Эшкабилова, Ю. Г. Никонорова, А. И. Парфенова[3].Для решения задачи семантического аннотирования формул в PDF-документах необходимо решить следующие задачи:

1. Разделение документа на блоки.
2. Выделение главных формул и текстовых блоков.
3. Поиск переменных в текстовых блоках.
4. Распознавание главных формул и локальных переменных.
5. Связывание формул и локальных переменных.
6. Разметка математических понятий в текстовых блоках на основе онтологии OntoMathPro[4].
7. Связывание выделенных понятий с переменными формулы.

Опишем подробнее каждую из указанных подзадач.

**Разделение документа на блоки.** Под блоками будем понимать обособленные участки символов, такие как формулы и текстовые абзацы. Данный этап выполняется с помощью функционала анализа разметки документа, встроенного в библиотеку pdfminer, и группировки отдельных элементов в блоки. На Рис. 5 приведен пример результата разделения страницы документа на блоки. При этом

[3] http://mi.mathnet.ru/mt270, http://mi.mathnet.ru/mt271, http://mi.mathnet.ru/mt272
[4] https://github.com/CLLKazan/OntoMathPro

для последующих этапов обработки документа сохраняется информация об отдельных строках документа и символах в строках. Кроме того, выполняется фильтрация служебных блоков, таких как номер формулы (например, «(1)»), и номеров страниц.

Введем понятия главной формулы, текстового блока и локальной переменной текстового блока.

Под главной формулой будем понимать формулу, отделенную от текстовой информации разрывами строк (например, такую, как на Рис. 2).

Текстовым блоком называется блок, содержащий как текстовую, так и формульную информацию. Примеры текстовых блоков приведены на Рис. 3.

Под локальной переменной текстового блока понимаются формулы, расположенные внутри текстовых блоков. Примеры локальных переменных приведены на Рис. 4

**Выделение главных формул и текстовых блоков.** Для дальнейшего анализа структуры документа необходимо различать блоки с текстом и главные формулы (такие, как формула $H = H_0 - (T_1 + T_2)$ на Рис. 5). Для этого применяется проверка количества слов в блоке с помощью регулярного выражения. В случае, если количество слов в блоке меньше порогового, блок считается главной формулой. На этом этапе алгоритма для главных формул сохранялась соответствующая часть изображения страницы из исходного документа. Например, для формулы $H = H_0 - (T_1 + T_2)$ на Рис. 5Рис. 5 сохраняется изображение, приведенное на Рис. 6.

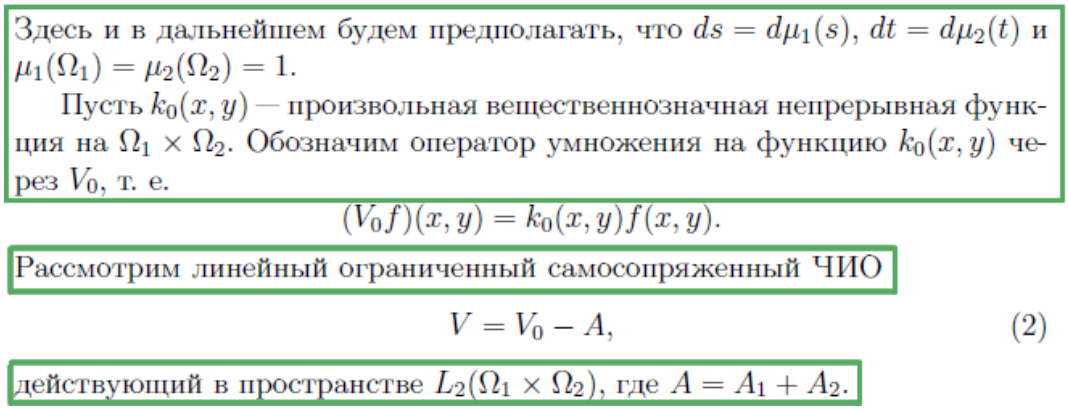
Пусть $u(x)$ и $v(y)$ — неотрицательные непрерывные функции на $\Omega_1$ и $\Omega_2$, $0 \in \mathrm{Ran}(u) \cap \mathrm{Ran}(v)$, $k_0(x,y) = u(x)v(y)$ и $\varphi_j(\cdot)$ — вещественнозначные непрерывные функции на $\Omega_j$, для которых выполняются следующие равенства:

$$\int_{\Omega_j} \varphi_j(\xi) d\mu_j(\xi) = 0, \quad \int_{\Omega_j} \varphi_j^2(\xi) d\mu_j(\xi) = 1, \quad j = 1, 2.$$

В этом параграфе мы рассмотрим вопросы о существовании отрицательных собственных значений, лежащих ниже нижнего края ЧИО $H$ вида $(1)$, и их количестве при вышеуказанных предположениях.

Рис. 2. Пример главной формулы в документе (выделена цветным

прямоугольником)


Здесь и в дальнейшем будем предполагать, что $ds = d\mu_1(s)$, $dt = d\mu_2(t)$ и $\mu_1(\Omega_1) = \mu_2(\Omega_2) = 1$.

Пусть $k_0(x, y)$ — произвольная вещественнозначная непрерывная функция на $\Omega_1 \times \Omega_2$. Обозначим оператор умножения на функцию $k_0(x, y)$ через $V_0$, т. е.

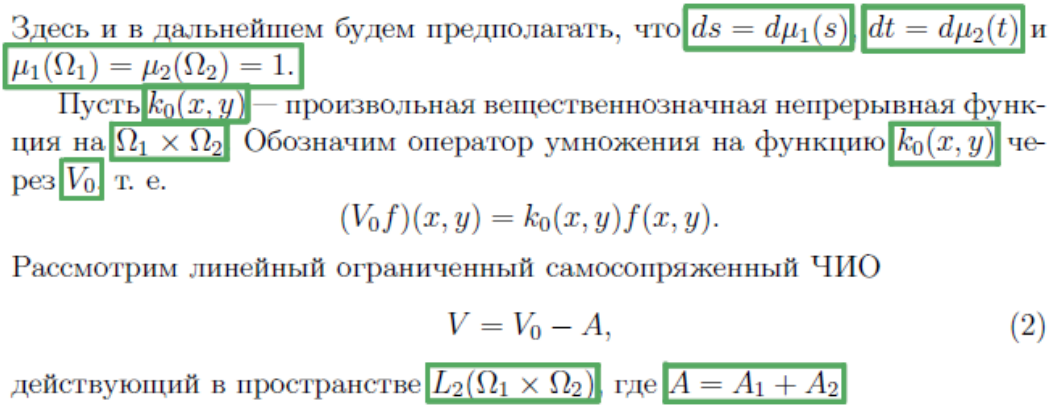
$$(V_0 f)(x, y) = k_0(x, y) f(x, y).$$

Рассмотрим линейный ограниченный самосопряженный ЧИО

$$V = V_0 - A, \qquad (2)$$

действующий в пространстве $L_2(\Omega_1 \times \Omega_2)$, где $A = A_1 + A_2$.

Рис. 3. Примеры текстового блока (выделен цветным прямоугольником)


Здесь и в дальнейшем будем предполагать, что $ds = d\mu_1(s)$, $dt = d\mu_2(t)$ и $\mu_1(\Omega_1) = \mu_2(\Omega_2) = 1$.

Пусть $k_0(x, y)$ — произвольная вещественнозначная непрерывная функция на $\Omega_1 \times \Omega_2$. Обозначим оператор умножения на функцию $k_0(x, y)$ через $V_0$, т. е.

$$(V_0 f)(x, y) = k_0(x, y) f(x, y).$$

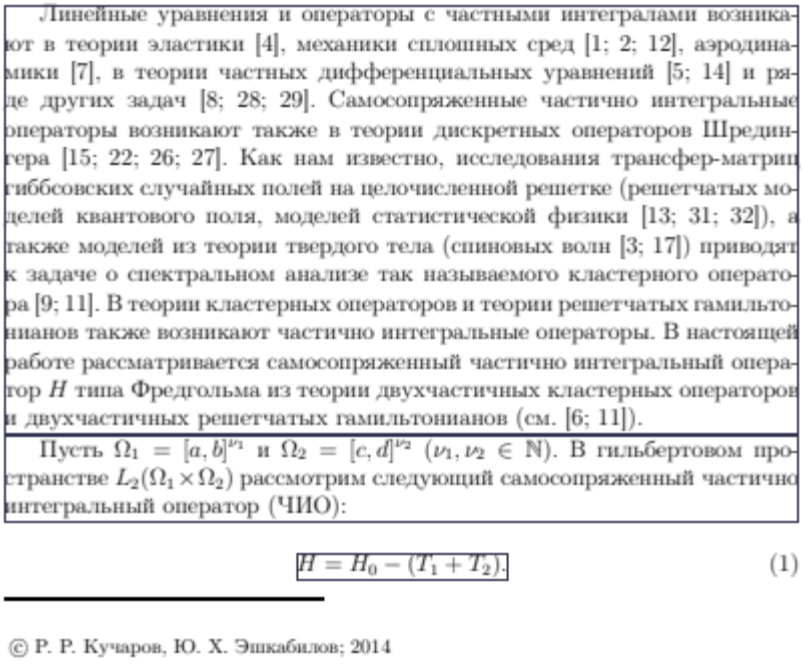
Рассмотрим линейный ограниченный самосопряженный ЧИО

$$V = V_0 - A, \qquad (2)$$

действующий в пространстве $L_2(\Omega_1 \times \Omega_2)$, где $A = A_1 + A_2$

Рис. 4. Примеры локальных переменных (выделены цветным прямоугольником)

На Рис. 7 приведен результат определения типа блока – синим цветом выделены текстовые блоки, красным – блоки с главной формулой.



Линейные уравнения и операторы с частными интегралами возникают в теории эластики [4], механики сплошных сред [1; 2; 12], аэродинамики [7], в теории частных дифференциальных уравнений [5; 14] и ряде других задач [8; 28; 29]. Самосопряженные частично интегральные операторы возникают также в теории дискретных операторов Шредингера [15; 22; 26; 27]. Как нам известно, исследования трансфер-матриц гиббсовских случайных полей на целочисленной решетке (решетчатых моделей квантового поля, моделей статистической физики [13; 31; 32]), а также моделей из теории твердого тела (спиновых волн [3; 17]) приводят к задаче о спектральном анализе так называемого кластерного оператора [9; 11]. В теории кластерных операторов и теории решетчатых гамильтонианов также возникают частично интегральные операторы. В настоящей работе рассматривается самосопряженный частично интегральный оператор $H$ типа Фредгольма из теории двухчастичных кластерных операторов и двухчастичных решетчатых гамильтонианов (см. [6; 11]).

Пусть $\Omega_1 = [a,b]^{\nu_1}$ и $\Omega_2 = [c,d]^{\nu_2}$ $(\nu_1, \nu_2 \in \mathbb{N})$. В гильбертовом пространстве $L_2(\Omega_1 \times \Omega_2)$ рассмотрим следующий самосопряженный частично интегральный оператор (ЧИО):

$$H = H_0 - (T_1 + T_2). \qquad (1)$$



Рис. 5. Пример разделения страницы документа на блоки

$$H = H_0 - (T_1 + T_2)$$

Рис. 6. Пример изображения, привязанного к блоку с главной формулой

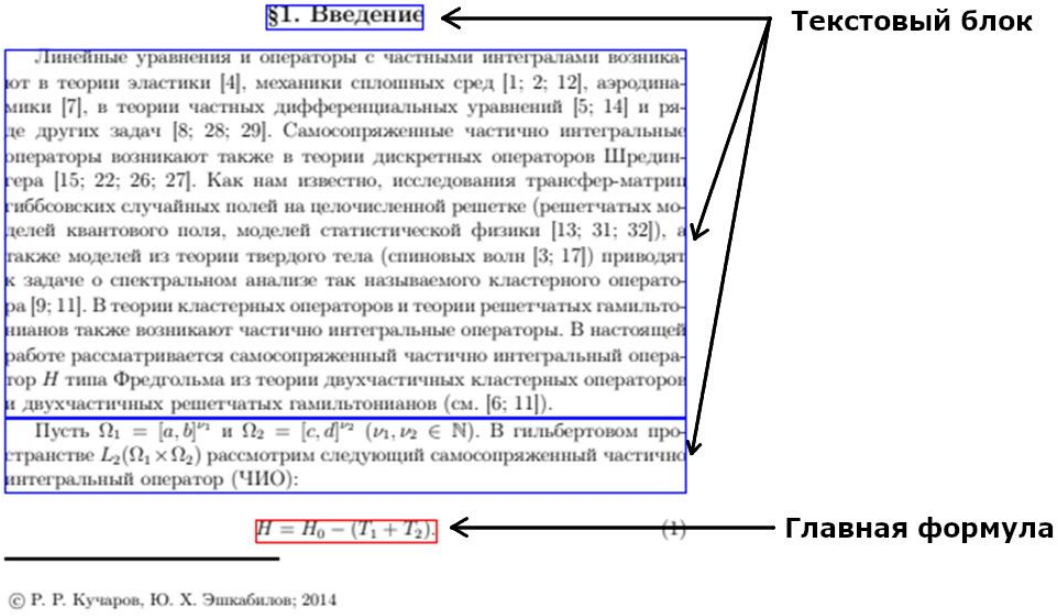



Рис. 7. Определение типа блока

**Поиск переменных в текстовых блоках.** Для выделения переменных в текстовых блоках выполняются сегментация текста на предложения и затем токенизация (выделение слов) в предложениях. В список локальных переменных добавляются все участки текста, не входящие в список слов текстового блока. Схема алгоритма поиска приведена на Рис. 8.

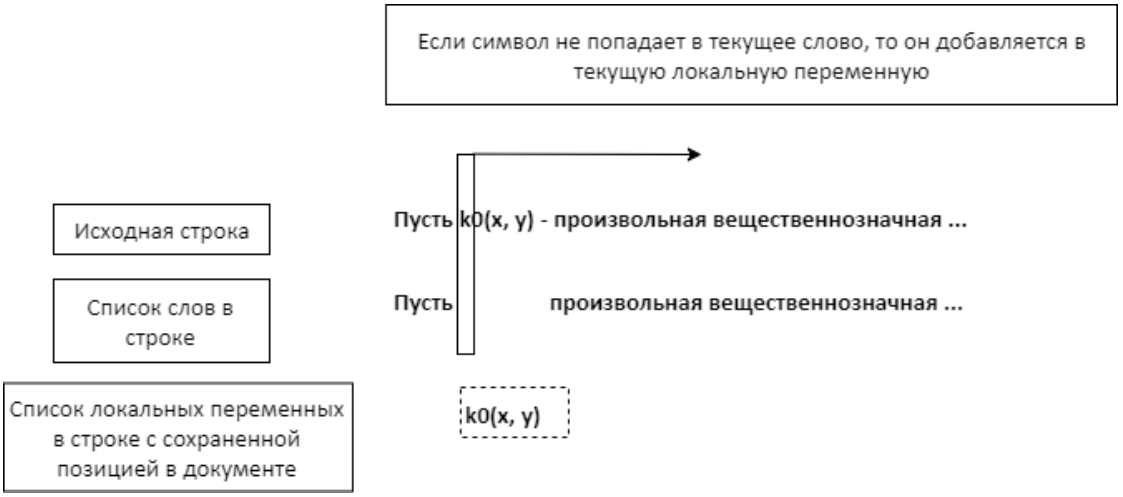
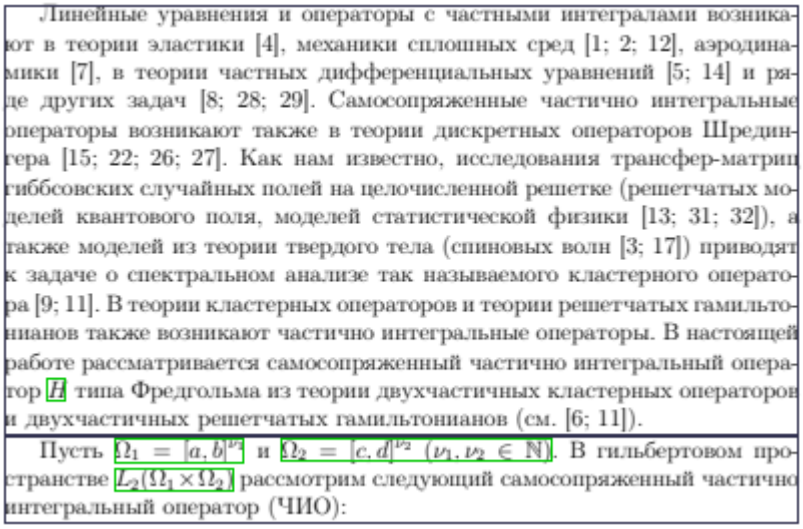



Рис. 8. Поиск локальных переменных в текстовых блоках

В результате этой операции формируется набор локальных переменных, и для каждой собирается набор символов, входящих в имя переменной. Однако некоторые специфические математические символы (например, символ интеграла, символ суммы) не распознаются библиотекой pdfminer. По этой причине для каждой локальной переменной дополнительно извлекается соответствующая часть изображения страницы из исходного документа. На Рис. 9 прямоугольниками зеленого цвета отмечены части изображения страницы, соответствующие локальным переменным.



Линейные уравнения и операторы с частными интегралами возникают в теории эластики [4], механики сплошных сред [1; 2; 12], аэродинамики [7], в теории частных дифференциальных уравнений [5; 14] и ряде других задач [8; 28; 29]. Самосопряженные частично интегральные операторы возникают также в теории дискретных операторов Шредингера [15; 22; 26; 27]. Как нам известно, исследования трансфер-матриц гиббсовских случайных полей на целочисленной решетке (решетчатых моделей квантового поля, моделей статистической физики [13; 31; 32]), а также моделей из теории твердого тела (спиновых волн [3; 17]) приводят к задаче о спектральном анализе так называемого кластерного оператора [9; 11]. В теории кластерных операторов и теории решетчатых гамильтонианов также возникают частично интегральные операторы. В настоящей работе рассматривается самосопряженный частично интегральный оператор $H$ типа Фредгольма из теории двухчастичных кластерных операторов и двухчастичных решетчатых гамильтонианов (см. [6; 11]).

Пусть $\Omega_1 = [a, b]^{\nu_1}$ и $\Omega_2 = [c, d]^{\nu_2}$ $(\nu_1, \nu_2 \in \mathbb{N})$. В гильбертовом пространстве $L_2(\Omega_1 \times \Omega_2)$ рассмотрим следующий самосопряженный частично интегральный оператор (ЧИО):

Рис. 9. Локальные переменные и их позиции на изображении страницы (отмечены прямоугольниками зеленого цвета)

**Распознавание главных формул и локальных переменных.** Для более корректного связывания переменных в тексте с главными формулами выполняется преобразование главных формул и локальных переменных в формат LaTeX. Для этого применяется библиотека pix2tex, использующая предобученную нейросетевую модель для распознавания формул LaTeX на изображениях. В Таблице 1 приведены результаты распознавания с указанием исходного текста формулы и переменных в документе и распознанного LaTeX представления. Распознавание не всегда происходит корректно, но для успешного связывания переменной и формулы нет необходимости в идеальном совпадении переменной и части формулы.

**Связывание формул и локальных переменных** производилось с помощью выделения обозначений формул с учетом нижних и верхних индексов (например, $\mathrm{f}(x)$, $f_0(x, y)$) 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. Для переменных, не обозначающих формулу с аргументами (например, на Рис. 9 выделен частично интегральный оператор H), проводится прямой поиск в главной формуле. На Рис. 10Рис. 10 приведена схема алгоритма связывания главных формул с локальными переменными. На Рис. 11 дан пример

связывания главной формулы с локальными переменными.

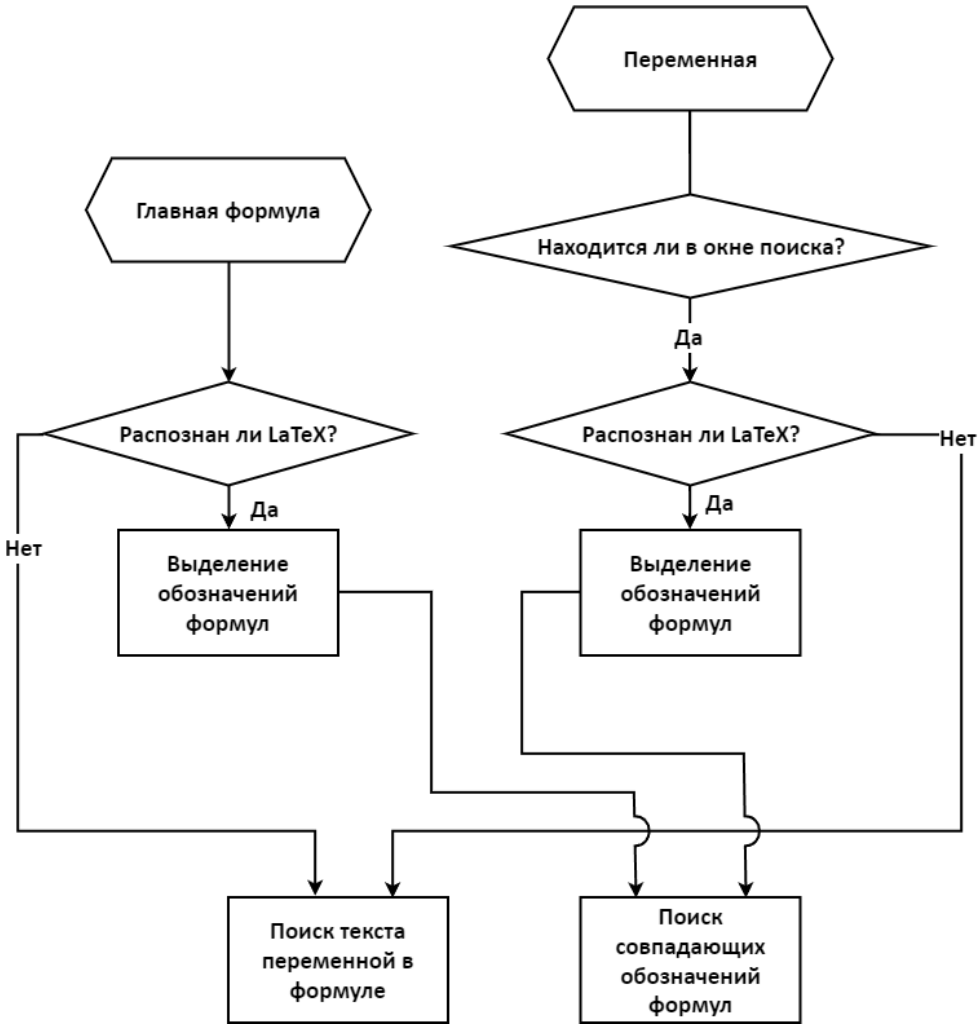



Рис. 10. Схема алгоритма связывания локальных переменных и главных формул

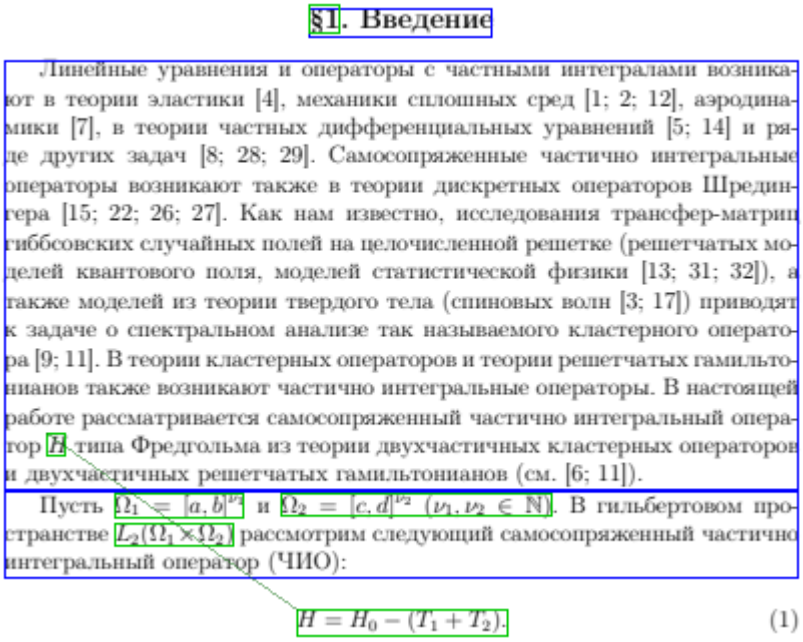



Рис. 11. Связывание главной формулы и переменной

**Таблица 1**. Примеры распознанных формул и переменных

| **Исходное текстовое представление формулы** | **Распознанное представление в LaTeX** |
|---|---|
| H = H0 – (T1 + T2) | H=H_{0}-(T_{1}+T_{2}) |
| H0f (x, y) = k0(x, y)f (x, y) | H_{0}f(x,y)=k_{0}(x,y)f(x,y) |
| Ω1 × Ω2 | \\Omega_{1}\\times\\Omega_{2} |
| k2(x, t, y)f (x, t)dt | k_{2}(x,t,y)f(x,t)d u |

В Таблице 2 приведены оценки полноты и точности алгоритма связывания локальных переменных и главных формул. В качестве экспериментальной выборки были использованы 9 документов на портале Math-Net.ru. Показатель полноты алгоритма связывания указывает на необходимость дальнейшего улучшения процедуры связывания локальных переменных и главных формул.

**Таблица 2.** Оценки полноты и точности для алгоритма связывания локальной переменной и главной формулы

| **Корректные связывания (TP)** | **Некорректные связывания (FP)** | **Пропущенные связывания (FN)** | **Точность** | **Полнота** | **F-мера** |
|---|---|---|---|---|---|
| 38 | 6 | 14 | 0,86 | 0,73 | 0,79 |
| 30 | 4 | 49 | 0,88 | 0,38 | 0,53 |
| 51 | 24 | 24 | 0,68 | 0,68 | 0,68 |
| 48 | 6 | 42 | 0,89 | 0,53 | 0,67 |
| 110 | 60 | 35 | 0,65 | 0,76 | 0,70 |
| 68 | 6 | 30 | 0,92 | 0,69 | 0,79 |
| 98 | 12 | 6 | 0,89 | 0,94 | 0,92 |
| 100 | 32 | 6 | 0,76 | 0,94 | 0,84 |
| 90 | 24 | 18 | 0,79 | 0,83 | 0,81 |
| **Средние значения** | | | **0,81** | **0,72** | **0,75** |

**Разметка математических понятий в текстовых блоках на основе онтологии OntoMathPro**

Для определения понятия, связанного с локальной формулой, был доработан и применен метод автоматической разметки текстовых документов понятиями из математической онтологии. Этот метод применяется при подготовке образовательных курсов по математике в Казанском федеральном университете. Основная идея данного алгоритма заключается в поиске всех цепочек слов в предложении и сравнении их с аналогично построенными цепочками в понятиях онтологии. Алгоритм принимает на вход документы в формате html, поэтому для его применения в данной задаче был создан метод генерации промежуточного html-представления PDF-документа. В тегах такого представления было получено текстовое представление исходного документа с указанием порядкового номера блока. Подробная схема алгоритма разметки математических документов приведена на Рис. 12.

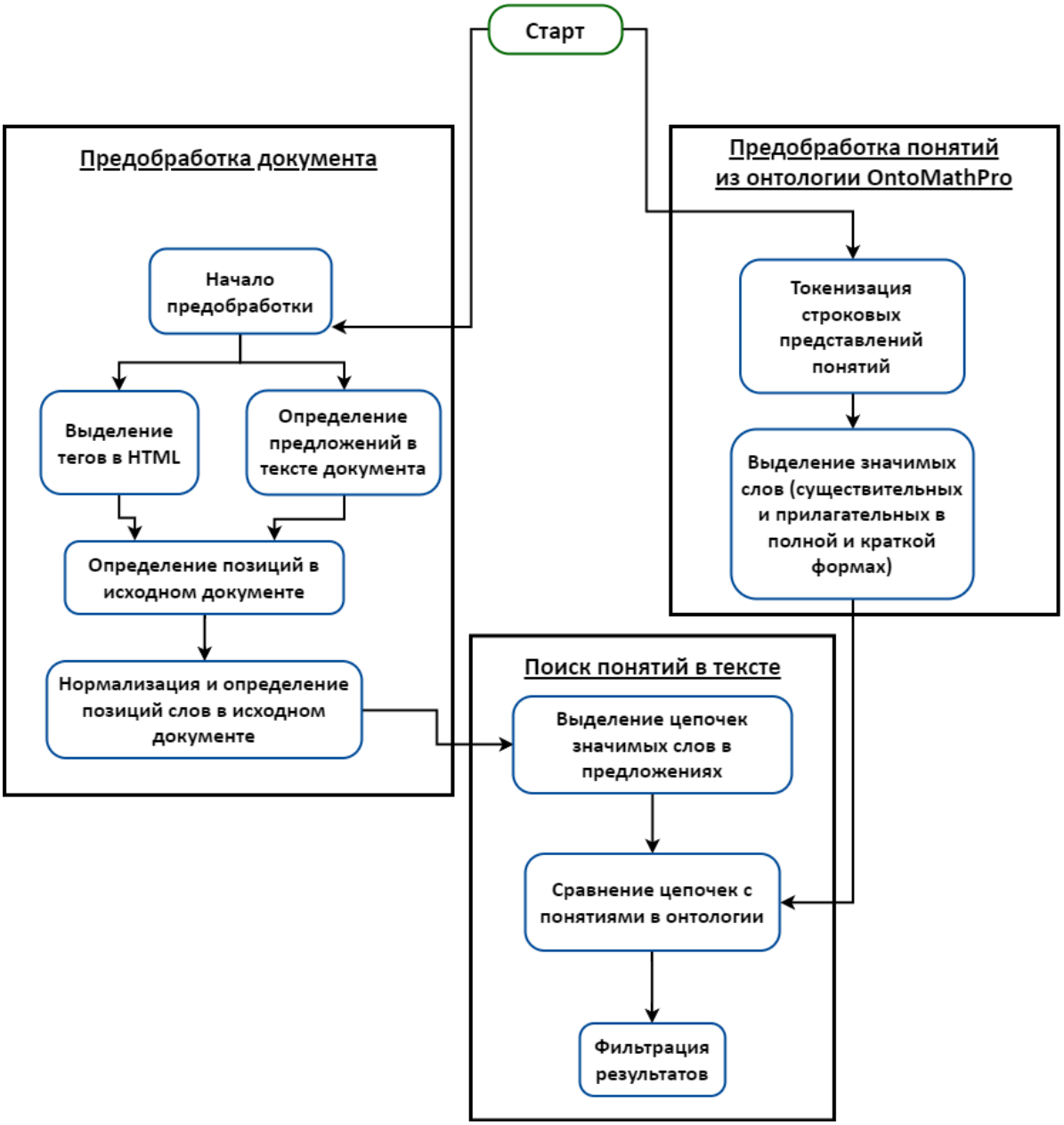



Рис. 12. Схема алгоритма разметки математических документов

Для применения метода разметки математических документов в выделении понятий в текстовых блоках PDF-документов была применена онтология профессиональной математики OntoMathPro [11].

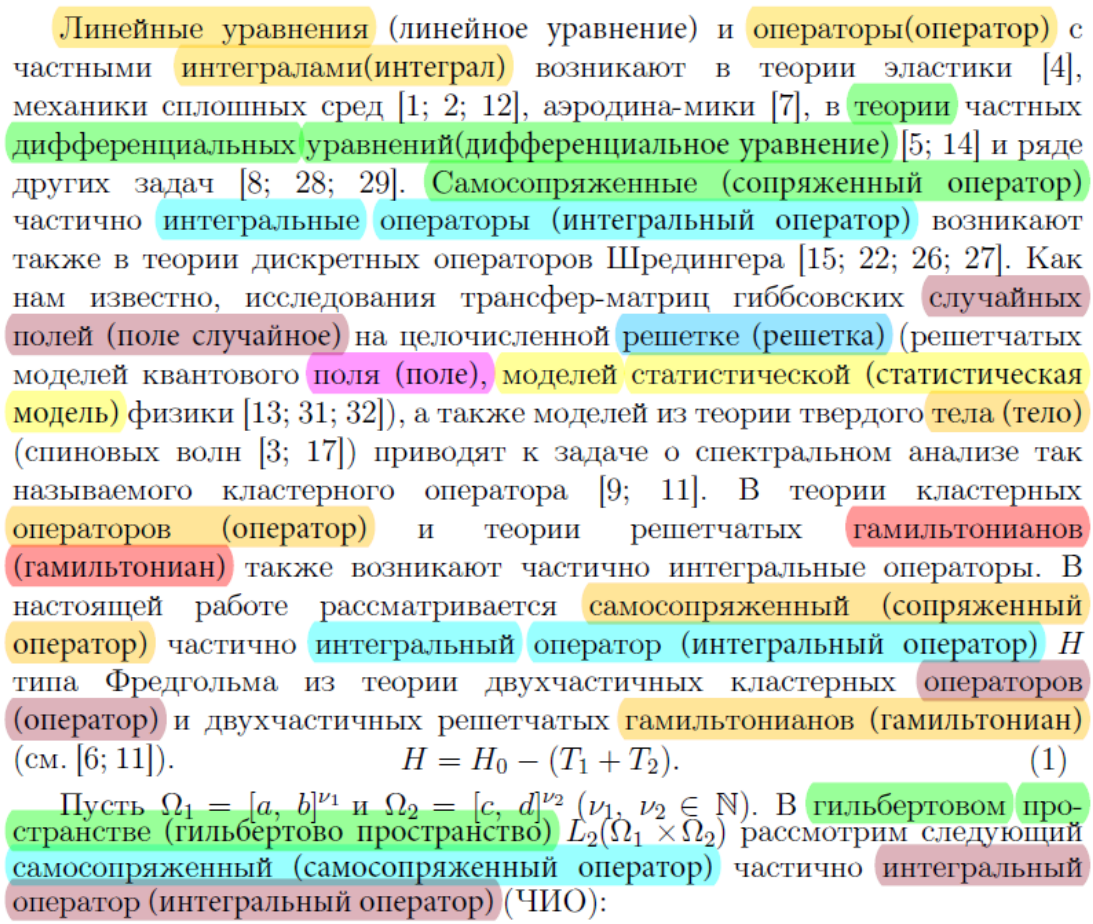
Онтология профессиональной математики OntoMathPro – это прикладная онтология для автоматической обработки профессиональных математических статей на русском и английском языках, разработанная в Казанском федеральном университете. Эта онтология охватывает широкий спектр областей математики,

таких как теория чисел, теория множеств, алгебра, анализ, геометрия, теория вычислений, дифференциальные уравнения, численный анализ, теория вероятностей и статистика. Каждый концепт онтологии имеет аннотацию, имя на русском и английском языках, включая синонимы. В онтологии можно выделить две таксономии по отношению *ISA* – иерархия областей математики и иерархия объектов математического знания (Рис. 13). Понятия онтологии, включенные в иерархию объектов математического знания, используются в разработанном алгоритме связывания переменных формулы.


Рис. 13. Фрагмент онтологии OntoMathPro: разделы математики и элементы математического знания

**Связывание выделенных понятий с переменными формулы**

В результате работы метода разметки математических документов формируется список понятий и позиций слов в документе, к которым привязано соответствующее понятие из онтологии OntoMathPro. Кандидаты на связывание с переменной выбираются в пределах предложения, в котором находится переменная. На данном этапе к переменной привязываются непосредственно соседние понятия в тексте. На Рис. 14 приведен пример разметки текста понятиями из онтологии OntoMathPro. В скобках после каждого цветного выделения приведено название распознанного понятия.



Линейные уравнения (линейное уравнение) и операторы(оператор) с частными интегралами(интеграл) возникают в теории эластики [4], механики сплошных сред [1; 2; 12], аэродина-мики [7], в теории частных дифференциальных уравнений(дифференциальное уравнение) [5; 14] и ряде других задач [8; 28; 29]. Самосопряженные (сопряженный оператор) частично интегральные операторы (интегральный оператор) возникают также в теории дискретных операторов Шредингера [15; 22; 26; 27]. Как нам известно, исследования трансфер-матриц гиббсовских случайных полей (поле случайное) на целочисленной решетке (решетка) (решетчатых моделей квантового поля (поле), моделей статистической (статистическая модель) физики [13; 31; 32]), а также моделей из теории твердого тела (тело) (спиновых волн [3; 17]) приводят к задаче о спектральном анализе так называемого кластерного оператора [9; 11]. В теории кластерных операторов (оператор) и теории решетчатых гамильтонианов (гамильтониан) также возникают частично интегральные операторы. В настоящей работе рассматривается самосопряженный (сопряженный оператор) частично интегральный оператор (интегральный оператор) $H$ типа Фредгольма из теории двухчастичных кластерных операторов (оператор) и двухчастичных решетчатых гамильтонианов (гамильтониан) (см. [6; 11]).

$$H = H_0 - (T_1 + T_2). \qquad (1)$$

Пусть $\Omega_1 = [a,\ b]^{\nu_1}$ и $\Omega_2 = [c,\ d]^{\nu_2}$ $(\nu_1,\ \nu_2 \in \mathbb{N})$. В гильбертовом пространстве (гильбертово пространство) $L_2(\Omega_1 \times \Omega_2)$ рассмотрим следующий самосопряженный (самосопряженный оператор) частично интегральный оператор (интегральный оператор) (ЧИО):

Рис. 14. Размеченные понятия в тексте документа

Для каждой формулы, у которой есть связанные локальные переменные с привязанными понятиями, формируются набор связанных понятий и, как следствие, её семантическое наполнение. В Таблице 3 приведены примеры формул с распознанными понятиями по онтологии. Так, переменная, обозначающая понятие «Мера Лебега», была привязана к формуле, несмотря на то, что для этой переменной использовались различные символьные обозначения в разных документах.

В Таблице 4 приведены оценки точности, полноты и F-мера алгоритма связывания выделенных понятий с главной формулой. В эксперименте использовалась коллекция из 9 документов под авторством Р. Р. Кучарова, Ю. Х. Эшкабилова, Ю. Г. Никонорова, А. И. Парфенова, Н. М. Алиева, М. Э. Муминова, Л. Н. Ромакиной, Г. П. Арзикулова, В. Н. Берестовского, И. А. Зубаревой, И. С. Борисова,

С. Е. Хрущева, А. А. Боровкова, А. А. Могульского, Н. В. Васильевой, Н. В. Красощёк[5].

**Таблица 3.** Примеры формул с распознанными понятиями

| **Главная формула** | **Переменная** | **Понятие онтологии** |
|---|---|---|
| $\int_{\Omega_j} \varphi_j(\xi) d\mu_j(\xi) = 0, \quad \int_{\Omega_j} \varphi_j^2(\xi) d\mu_j(\xi) = 1$ | $\mu_j(\cdot)$ | Мера Лебега |
| $t \mapsto \frac{n!}{\sqrt{\sum_{i=1}^n (x_i - x_0)^2}} \cdot \mathrm{mes}\,(U(t)),$ | $\mathrm{mes}\,(U(t))$ | Мера Лебега |
| $H = H_0 - (T_1 + T_2).$ | $H$ | Интегральный оператор |
| $H_2(\beta)\psi(y) = u(\beta)v(y)\psi(y) - \int_{\Omega_2} (\mu_0 + \mu\varphi_2(y)\varphi_2(t))\,\psi(t)dt.$ | $\{H_2(\beta)\}_{\beta\in\Omega_1}$ | Семейство операторов |
| $P_H(x) = \sum_{k=0}^{l} \sum_{m=0}^{p_k-1} \sum_{s=0}^{m} \frac{f^{(m)}(z_k)}{(m-s)!(p_k-m-1)!} \frac{d^{m-s}}{dz^{m-s}} \left[\frac{(z-z_k)^{p_k}}{q(z)}\right]\Big\|_{z=z_k} \frac{q(x)}{(x-z_k)^{s+1}}$ (7) | $P_H$ | Многочлен Эрмита |

[5] http://mi.mathnet.ru/mt270, http://mi.mathnet.ru/mt271, http://mi.mathnet.ru/mt272, http://mi.mathnet.ru/mt273, http://mi.mathnet.ru/mt274, http://mi.mathnet.ru/mt275, http://mi.mathnet.ru/mt276, http://mi.mathnet.ru/mt277, http://mi.mathnet.ru/mt278

**Таблица 4.** Оценки полноты и точности для алгоритма определения семантического наполнения формулы

| **Корректные распознавания (TP)** | **Разметка несуществующих понятий (FP)** | **Пропущенные понятия (FN)** | **Точность** | **Полнота** | **F-мера** |
|---|---|---|---|---|---|
| 190 | 16 | 18 | 0,92 | 0,91 | 0,92 |
| 120 | 8 | 56 | 0,94 | 0,68 | 0,79 |
| 136 | 48 | 36 | 0,74 | 0,79 | 0,76 |
| 192 | 18 | 48 | 0,91 | 0,80 | 0,85 |
| 495 | 120 | 50 | 0,80 | 0,91 | 0,85 |
| 340 | 12 | 45 | 0,97 | 0,88 | 0,92 |
| 441 | 32 | 10 | 0,93 | 0,98 | 0,95 |
| 400 | 64 | 9 | 0,86 | 0,98 | 0,92 |
| 450 | 54 | 24 | 0,89 | 0,95 | 0,92 |
| **Средние значения** | | | **0,89** | **0,88** | **0,88** |

## ЗАКЛЮЧЕНИЕ

В статье представлен метод семантического аннотирования математических документов в формате PDF. Описан способ определения структуры документа с помощью разделения на блоки с текстом и главными формулами. Разработан метод связывания локальных переменных в тексте с главными формулами. С помощью метода разметки математических понятий в тексте формируется семантическое представление главных формул. Разработанный метод применяется для подготовки набора данных для семантического поисковика по формулам, который функционирует на цифровой платформе Lobachevskii-DML.

Будущие разработки связаны с расширением возможностей метода, в частности, для связывания многих переменных математической функции, а также разработки метода фильтрации понятий, обнаруженных в предложении, для увеличения точности метода аннотирования. Другим направлением для повышения универсальности метода можно считать планирование внедрения OCR-модуля

для распознавания текстов в отсканированных PDF, что особенно актуально для математических статей, изданных в доцифровую эпоху.

**БЛАГОДАРНОСТИ**

Исследование выполнено при поддержке Российского научного фонда, проект № 21-11-00105.

**СПИСОК ЛИТЕРАТУРЫ**

*1. Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E.* Bringing math to LOD: A semantic publishing platform prototype for scientific collections in mathematics // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013. Vol. 8218 LNCS. No. 1. P. 379–394.

2. *Bertin M., Atanassova I.* Hybrid Approach for the Semantic Processing of Scientific Papers // Semantic Publishing Challenge Track in 11th European Semantic Web Conference (ESWC 2014). 2014. P. 1–5.

*3. Ciancarini P., Di Iorio A., Nuzzolese A.G., Silvio P., Fabio V.* Semantic annotation of scholarly documents and citations // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013. Vol. 8249 LNAI. P. 336–347.

4. *Ronzano F., Del Bosque G.C., Saggion H.* CEUR-WS proceedings: Towards the automatic generation of highly descriptive scholarly publishing linked datasets // Communications in Computer and Information Science. 2014. Vol. 475. P. 83–88.

5. *Ahmad R., Afzal M.T., Qadir M.A.* Information extraction from PDF sources based on rule-based system using integrated formats // Communications in Computer and Information Science. 2016. Vol. 641. P. 293–308.

6. *Greiner-Petter A., Youssef A., Ruas T., Miller, Bruce R., Schubotz M., Aizawa A., Gipp B.* Math-word embedding in math search and semantic extraction // Scientometrics. 2020. Vol. 125. No. 3. P. 3017–3046.

7. *Wolska M., Grigore M.* Symbol declarations in mathematical writing // Proceedings of the 3rd Workshop on Digital Mathematics Libraries. 2010. P. 119–127.

8. *Líška M., Sojka P., Ružička M., Mravec P.* Web interface and collection for

mathematical retrieval WebMIaS and MREC // DML 2011 – Towards a Digital Mathematics Library, Proceedings. 2011. P. 77–84.

9. *Schubotz M., Greiner-Petter A., Scharpf P., Meuschke N., Cohl H.S., Gipp B.* Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context // Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. New York, NY, USA: ACM. 2018. P. 233–242.

10. *Nevzorova O., Kirillovich A., Nevzorov V., Nikolaev K.* The semantic context models of mathematical formulas in scientific papers // CEUR Workshop Proceedings. 2018. Vol. 2277. P. 33–40.

11. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* OntoMath-PRO: Ontology of Mathematical Knowledge // Dokl. Math. 2022. https://doi.org/10.1134/S1064562422700016

## SEMANTIC ANNOTATION OF MATHEMATICAL FORMULAS IN PDF-DOCUMENTS

**O. A. Nevzorova[1] [0000-0001-8116-9446], K. S. Nikolaev[2] [0000-0003-3204-238X]**

[1-2] *Kazan (Volga Region) Federal University, 35 Kremlyovskaya str., Kazan, 420008*
[1]onevzoro@gmail.com, [2]konnikolaeff@yandex.ru

***Abstract***

This article provides an overview of existing solutions for semantic analysis of mathematical documents, and also presents a method for automatic semantic analysis of documents in PDF format. This method searches for local variables in the text of the article, extracts their definitions and connects concepts with formulas. The advantage of the method over the existing ones is independence from the markup of the original PDF document, which expands the scope of the method. We provide estimates of recall, precision and F-measure for algorithms for finding variables and linking local variables with formulas. The resulting semantic markup of the document will be used to create a collection of documents suitable for the semantic formula search service, which is part of the set of services of the Lobachevskii-DML digital publishing system.

***Keywords**: semantic analysis, PDF, document processing, scientific journals, Lobachevskii-DML.*

**REFERENCES**

1. *Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E.* Bringing math to LOD: A semantic publishing platform prototype for scientific collections in mathematics // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013. Vol. 8218 LNCS. No. 1. P. 379–394.

2. *Bertin M., Atanassova I.* Hybrid Approach for the Semantic Processing of Scientific Papers // Semantic Publishing Challenge Track in 11th European Semantic Web Conference (ESWC 2014). 2014. P. 1–5.

3. *Ciancarini P., Di Iorio A., Nuzzolese A.G., Silvio P., Fabio V.* Semantic annotation of scholarly documents and citations // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013. Vol. 8249 LNAI. P. 336–-347.

4. *Ronzano F., Del Bosque G.C., Saggion H.* CEUR-WS proceedings: Towards the automatic generation of highly descriptive scholarly publishing linked datasets // Communications in Computer and Information Science. 2014. Vol. 475. P. 83–88.

5. *Ahmad R., Afzal M.T., Qadir M.A.* Information extraction from PDF sources based on rule-based system using integrated formats // Communications in Computer and Information Science. 2016. Vol. 641. P. 293–308.

6. *Greiner-Petter A., Youssef A., Ruas T., Miller, Bruce R., Schubotz M., Aizawa A., Gipp B.* Math-word embedding in math search and semantic extraction // Scientometrics. 2020. Vol. 125. No. 3. P. 3017–3046.

7. *Wolska M., Grigore M.* Symbol declarations in mathematical writing // Proceedings of the 3rd Workshop on Digital Mathematics Libraries. 2010. P. 119–127.

8. *Líška M., Sojka P., Ružička M., Mravec P.* Web interface and collection for mathematical retrieval WebMIaS and MREC // DML 2011 – Towards a Digital Mathematics Library, Proceedings. 2011. P. 77–84.

9. *Schubotz M., Greiner-Petter A., Scharpf P., Meuschke N., Cohl H.S., Gipp B.*

Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context // Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. New York, NY, USA: ACM. 2018. P. 233–242.

10. *Nevzorova O., Kirillovich A., Nevzorov V., Nikolaev K.* The semantic context models of mathematical formulas in scientific papers // CEUR Workshop Proceedings. 2018. Vol. 2277. P. 33–40.

11. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* OntoMath-PRO: Ontology of Mathematical Knowledge // Dokl. Math. 2022. https://doi.org/10.1134/S1064562422700016

## СВЕДЕНИЯ ОБ АВТОРАХ

***НЕВЗОРОВА Ольга Авенировна*** – доцент кафедры информационных систем Института вычислительной математики и информационных технологий Казанского федерального университета, к. т. н. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

***Olga Avenirovna NEVZOROVA*** – Kazan Federal University, Institute of Computational Mathematics and Information Technologies, Associated Professor of the Department of Information System, PhD. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: onevzoro@gmail.com
ORCID: 0000-0001-8116-9446

***НИКОЛАЕВ Константин Сергеевич*** – ассистент кафедры системного анализа и информационных технологий Института Вычислительной математики и информационных технологий Казанского федерального университета. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

***Konstantin Sergeevich NIKOLAEV*** – Assistant of the Department of System Analysis and Information Technologies of the Institute of Computational Mathematics and Information Technologies of Kazan Federal University. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: konnikolaeff@yandex.ru
ORCID: 0000-0003-3204-238X

*Материал поступил в редакцию 28 октября 2022 года*