

УДК 004.8

Применение больших языковых моделей для извлечения основных формул из текстов научных статей

Э.В. Саитов^[0009-0009-1980-8740], О.А. Невзорова^[0000-0001-8116-9446]

Казанский федеральный университет

Аннотация. В статье рассматриваются проблемы распознавания формул научных статей, представленных в формате pdf, а также преобразование выделенных формул из формата pdf в формат LaTeX. Предлагается решение данных проблем с использованием больших языковых моделей. Рассматривается извлечение главных формул математической статьи и их последующее представление в формате LaTeX с помощью большой языковой модели DeepSeek. Извлечение основных результатов статьи выполняется на основе аннотации. Затем осуществляется поиск фрагментов статьи, описывающих основные результаты и выделение внутри найденных фрагментов главных формул с помощью большой языковой модели. Получены оценки решения по метрикам точности, полноты и F-меры, средние значения F-меры по тестовой коллекции равны 0,93459 и 0,9294. В качестве способа хранения полученных данных предлагается система управления базами данных MySQL, приведено описание структуры базы данных.

Ключевые слова: большая языковая модель, математическая статья, основная формула, главная формула, формат LaTeX, извлечение формулы.

Using Large Language Models to Extract Key Formulas from Scientific Article Texts

E.V. Saitov, O.A. Nevzorova

Kazan Federal University

Abstract. This article studies the problem of extracting formulas of scientific articles presented in pdf format, as well as converting the selected formulas from pdf to LaTeX format. The extraction of the main formulas of a mathematical article and their subsequent presentation in LaTeX format using the large language model DeepSeek is considered. Extraction of the main results of the

papers bases on the paper annotations. Then, using a large language model we search fragments of papers describing the main results and highlighting the main formulas within the found fragments. The solution was evaluated using the metrics of precision, recall, and F-measure; the average values of the F-measure for the test collection was 0.93459 and 0,9294. The MySQL database management system is given as a method for storing the obtained data and the description of the database structure is given in text form: database tables, their attributes, as well as types of stored data.

Keywords: large language model, LLM, mathematical paper, basic formula, main formula, LaTeX format, formula extraction.

Введение

В последние годы произошел значительный прорыв в развитии больших языковых моделей с появлением целого ряда новых разработок от ведущих технологических компаний. Развитие больших языковых моделей открывает новые пути для решения ряда научных задач. Одна из актуальных задач управления математическим знанием в цифровом пространстве связана с разработкой семантического поисковика, позволяющего осуществлять поиск по формулам научных статей. В статье рассматривается подготовка базы данных, содержащей метаданные математических статей. Метаданные включают основные результаты, главные формулы, связанные с основными результатами, фрагменты текста с главными формулами с указанием логических единиц (леммы, теоремы и др.) и страниц документа, а также аннотацию статьи,

Основной трудностью при формировании базы данных семантического поисковика является распознавание главных формул в тексте научной статьи. Можно указать ряд статей, связанных с извлечением и аннотированием математических формул.

В работе [1] для выделения главных формул из текста статьи было предложено разделить текст на блоки для дальнейшей проверки количества слов в блоке с использованием регулярных выражений. В работе [2] рассматриваются методы аннотирования математических формул в PDF-документах и их семантическая разметка. В работе [3] предлагается алгоритм семантического поиска математических формул, основанный на анализе их структуры и контекста. В работе [4] проводится аналитический обзор существующих математических поисковых систем, предлагаются направления для их дальнейшего развития.

Проект ArqMath Lab направлен на разработку инструментов для семантического поиска математических формул в открытых научных базах данных [5]. Авторы исследуют возможности применения современных методов обработки естественного языка и машинного обучения для улучшения поиска и анализа математических формул в таких системах, как zbMATH. В работе [6] рассматриваются методы распознавания и извлечения математических выражений из документов на основе

комбинации структурных и семантических признаков для идентификации ключевых формул. В работе [7] предлагается подход к извлечению главных формул из научных статей, основанный на анализе пространственных и лексических особенностей текста. В работе [8] авторы исследуют применение методов глубокого обучения для автоматического выделения ключевых формул из научных статей. В работе [9] описывается гибридный метод извлечения доминирующих формул, сочетающий анализ контекста и структурных особенностей математических выражений.

В работе [10] изучается комбинация layout-анализа и семантических признаков для поиска математических формул, с учетом как визуального расположения, так и смыслового содержания формул. В работе [11] предлагается правило-ориентированный подход для идентификации и извлечения математических формул из PDF-документов.

В настоящей статье для выделения основных формул предлагается использовать современные большие языковые модели (Large Language Model, LLM), в частности, модель DeepSeek. Задачей данного исследования является оценка эффективности использования современных больших языковых моделей для извлечения основных формул из текста научных статей.

Подготовка данных для исследования

В качестве данных для исследования были отобраны 50 научных статей с портала Math-Net.ru (<https://www.mathnet.ru/>) из разных разделов математики, таких как дифференциальные уравнения, математический анализ, теория вероятностей и математическая статистика

Для проводимого исследования была выбрана большая языковая модель DeepSeek из-за своей доступности на территории Российской Федерации, а также высокой релевантности выдаваемых ответов на вводимые пользователем запросы по сравнению с другими языковыми моделями.

Извлечение основных результатов статьи из аннотаций

В данной работе предлагается подход для выделения основных формул в тексте статьи путем связывания 3-х логических единиц: основных результатов статьи, фрагментов текста, связанными с основными результатами, и основных формул, принадлежащих к фрагментам текста.

Источником основных результатов, описанных в статье, является аннотация. Аннотация статьи прикрепляется в виде текста либо файла формата PDF к запросу большой языковой модели.

Далее в качестве примера рассматривается статья [12], в которой рассматриваются результаты из области обыкновенных дифференциальных уравнений.

Запрос к LLM DeepSeek на выделение главных результатов из аннотации статьи [12] представлен на рисунке 1.

Дана следующая аннотация:

Получена формальная классификация типичных локальных перестроек неявного обыкновенного дифференциального уравнения в его особых точках при изменении одного внешнего параметра. Она состоит из четырех нормальных форм, каждая из которых содержит функциональный инвариант. Доказано, что любую деформацию квадратичного по производной ростка уравнения в классе контактной эквивалентности, при которой он остается квадратичным по производной, можно получить с помощью деформации независимой и зависимой переменных. На обобщении этого результата для семейств уравнений основана наша классификация. В качестве ее приложения получена формальная классификация типичных локальных перестроек на плоскости линейного уравнения смешанного типа с частными производными второго порядка в точках, где области эллиптичности и гиперболичности испытывают морсовские перестройки.

Выделить из данной аннотации основные результаты, полученные в ходе исследования.

Рис. 1. Запрос 1: Основные результаты статьи

Главные результаты, выделенные из аннотации статьи [12] с помощью DeepSeek показаны на рисунке 2.



Основные результаты исследования, выделенные из аннотации:

1. **Формальная классификация типичных локальных перестроек** неявного обыкновенного дифференциального уравнения (ОДУ) в его особых точках при изменении одного внешнего параметра. Классификация включает:
 - Четыре нормальные формы, каждая из которых содержит функциональный инвариант.
2. **Доказательство возможности сведения деформаций** квадратичного по производной роста уравнения к деформациям независимой и зависимой переменных в классе контактной эквивалентности.
3. **Обобщение этого результата** на семейства уравнений, что легло в основу предложенной классификации.
4. **Применение классификации** к формальному описанию типичных локальных перестроек на плоскости для линейного уравнения смешанного типа (с частными производными второго порядка) в точках, где области эллиптичности и гиперболичности претерпевают морсовские перестройки.

Таким образом, ключевые достижения работы включают:

- Классификацию перестроек неявных ОДУ,
- Теоретическое обоснование сведения деформаций,
- Приложение результатов к уравнениям смешанного типа.



Рис. 2. Ответ на запрос 1: Основные результаты статьи

Модель DeepSeek верно выделила результаты из аннотации загруженной статьи, а также дала подробное описание каждого результата.

Нахождение фрагментов текста и основных формул, относящихся к выделенным основным результатам

Следующим этапом исследования является нахождение фрагментов текста и основных формул, связанных с выделенными результатами на первом этапе.

Были сформулированы следующие требования к запросу к LLM. В качестве формата формул, выводимых в ответе, был выбран формат LaTeX. Также требовалось дать описание каждой переменной в выделенной формуле. При этом, если описание переменной не дано явно в тексте, LLM должна найти описание в другом источнике с указанием этого источника. Также требуется в ответе на запрос вывести ссылку в формате "страница статьи, на которой находится формула → тип логического элемента с номером (например, лемма, теорема и др.)".

Текст запроса к LLM для выделения фрагментов текста и основных формул представлен на рисунке 3.



Статья 49.pdf
PDF 728.06KB

Выделить из текста статьи фрагменты текста, содержащие элементы выделенных из аннотации основных результатов. Фрагменты текста должны быть взяты из текста статьи. Необходимо выделить фрагменты текста по каждому полученному из аннотации результату. Выделить в найденных фрагментах математические формулы, связанные с основными результатами по списку результатов. В одном фрагменте может быть несколько основных формул. Формулы и переменные должны быть записаны в формате LaTeX, например, $\$z_k\$$. Дать описание каждой переменной, указанной в формуле, в формате LaTeX. Описание должно быть дано снизу формул и содержать слово "где". Если описание переменной формулы не содержится в тексте статьи, дать ссылку на внешний источник, откуда было взято описание. Ссылка на внешний источник должна располагаться после описания переменной и содержать раздел и номер страницы источника, откуда было взято описание. Также под каждой формулой указать ссылку в формате "Страница статьи, на которой находится формула --> Тип утверждения с номером (Например, лемма, теорема и т.п.)". Отчёт страниц в статье начинать с первой страницы файла pdf.

Вот пример, как должны быть представлены формулы в формате LaTeX:

```
\begin{equation}
-\Delta u(x) - 2(a u_{x_1}(x) + b u_{x_1}(-x/p))_{x_2} = f(x)
\end{equation}
\begin{equation}
|x_1|^2 + |x_2|^2 + 2|x_1 x_2| \text{Re}(a \sqrt{p} b \lambda) > 0
\end{equation}
```

Рис. 3. Запрос 2: Фрагмент текста и основные формулы

Ответ модели DeepSeek на введенный запрос показан на рисунках 4-7.



Основные результаты и соответствующие фрагменты текста статьи

1. Формальная классификация типичных локальных перестроек неявного ОДУ

Фрагмент текста:

"Получена формальная классификация типичных локальных перестроек неявного обыкновенного дифференциального уравнения в его особых точках при изменении одного внешнего параметра. Она состоит из четырех нормальных форм, каждая из которых содержит функциональный инвариант."

Математические формулы:

```
\begin{equation}
\pm(y')^2 \pm x^2 + z^2 = \varepsilon + c(\varepsilon)z^3, \quad \text{где } z = y - \frac{1}{2}xy'
\end{equation}
```

где:

- y' — производная зависимой переменной y по независимой переменной x ,
- x, y — независимая и зависимая переменные,
- ε — внешний параметр,
- $c(\varepsilon)$ — функциональный инвариант.

Ссылка:

Страница 8 --> Теорема 1

Рис. 4. Ответ по запросу 2 для основного результата 1 из аннотации (см. рис.2): Фрагмент текста и основные формулы

2. Доказательство возможности сведения деформаций квадратичного роста

Фрагмент текста:

"Доказано, что любую деформацию квадратичного по производной роста уравнения в классе контактной эквивалентности, при которой он остается квадратичным по производной, можно получить с помощью деформации независимой и зависимой переменных."

Математические формулы:

$$\begin{equation}$$

$$f(x, y, p) = 0, \quad \text{где } f(a_0) = \partial_p f(a_0) = 0 \text{ и } \partial_p^2 f(a_0) \neq 0$$

$$\end{equation}$$

где:

- f — функция, задающая неявное ОДУ,
- a_0 — точка приложения роста,
- p — производная dy/dx .

Ссылка:

Страница 9 --> Определение 3

Рис. 5. Ответ по запросу 2 для основного результата 2 из аннотации (см. рис.2): Фрагмент текста и основные формулы

3. Применение классификации к уравнениям смешанного типа

Фрагмент текста:

"В качестве ее приложения получена формальная классификация типичных локальных перестроек на плоскости линейного уравнения смешанного типа с частными производными второго порядка в точках, где области эллиптичности и гиперболичности испытывают морсовские перестройки."

Математические формулы:

$$\begin{equation}$$

$$\alpha(x, y, \epsilon)u_{xx} + 2\beta(x, y, \epsilon)u_{xy} + \gamma(x, y, \epsilon)u_{yy} = \sigma(x, y,$$

$$\epsilon, u, u_x, u_y)$$

$$\end{equation}$$

где:

- α, β, γ — коэффициенты уравнения,
- ϵ — параметр,
- u_{xx}, u_{xy}, u_{yy} — частные производные функции u .

Ссылка:

Страница 16 --> Теорема 5

Рис. 6. Ответ по запросу 2 для основного результата 3 из аннотации (см. рис.2): Фрагмент текста и основные формулы

4. Обобщение результата для семейств уравнений

Фрагмент текста:

"На обобщении этого результата для семейств уравнений основана наша классификация."

Математические формулы:

$$\begin{equation}$$

$$F(x, y, p, \varepsilon) = 0, \quad \text{где } F(A_0) = \partial_p F(A_0) = 0 \text{ и } \partial_p^2 F(A_0) \neq 0$$

$$\end{equation}$$

где:

- F — функция, задающая семейство неявных ОДУ,
- A_0 — точка приложения роста,
- ε — параметр семейства.

Ссылка:

Страница 12 --> Теорема 4

Рис. 7. Ответ по запросу 2 для основного результата 4 из аннотации (см. рис.2): Фрагмент текста и основные формулы

Анализ ответов (рис. 4-7) показывает, что модель DeepSeek корректно выделяет фрагменты текста и основные формулы по каждому результату из аннотации. Также у LLM корректно представляет формулы в формате LaTeX и для каждой формулой указывает ссылку на страницу и логическую единицу, в которой дана формула.

Оценка эффективности извлечения основных результатов и формул

На данных тестовой коллекции из 50 статей была проведена оценка эффективности извлечения основных результатов и формул с помощью модели DeepSeek. Сравнивалось количество извлеченных основных результатов и формул статей на основе экспертной оценки и средствами модели DeepSeek по метрикам точности, полноты и F-меры. Результаты оценки приведены на рисунках 8 и 9.

Номер статьи	Количество основных результатов (экспертная оценка)	Количество основных результатов (оценка модели)	Точность	Полнота	F-мера	F-мера (сред)
1	4	4	1	1	1	0,934586
2	7	5	1	0,7142857	0,833333	
3	3	4	0,75	1	0,857143	
4	3	3	1	1	1	
5	3	2	1	0,6666667	0,8	
6	2	3	0,6666667	1	0,8	
7	5	5	1	1	1	
8	5	6	0,83333333	1	0,909091	
9	5	4	1	0,8	0,888889	
10	3	5	0,6	1	0,75	
11	3	3	1	1	1	
12	9	9	1	1	1	
13	5	5	1	1	1	
14	3	5	0,6	1	0,75	
15	3	3	1	1	1	
16	4	3	1	0,75	0,857143	
17	5	5	1	1	1	
18	5	5	1	1	1	
19	3	4	0,75	1	0,857143	
20	5	5	1	1	1	
21	4	4	1	1	1	
22	5	5	1	1	1	
23	3	3	1	1	1	
24	2	3	0,6666667	1	0,8	
25	2	3	0,6666667	1	0,8	
26	6	5	1	0,8333333	0,909091	
27	3	4	0,75	1	0,857143	
28	3	4	0,75	1	0,857143	
29	3	3	1	1	1	
30	2	2	1	1	1	
31	3	3	1	1	1	
32	4	4	1	1	1	
33	4	3	1	0,75	0,857143	
34	3	3	1	1	1	
35	3	3	1	1	1	
36	4	3	1	0,75	0,857143	
37	4	5	0,8	1	0,888889	
38	4	6	0,6666667	1	0,8	
39	3	3	1	1	1	
40	2	3	0,6666667	1	0,8	
41	4	4	1	1	1	
42	5	5	1	1	1	
43	5	5	1	1	1	
44	5	5	1	1	1	
45	4	4	1	1	1	
46	4	4	1	1	1	
47	7	7	1	1	1	
48	3	3	1	1	1	
49	4	4	1	1	1	
50	6	6	1	1	1	

Рис. 8 – Оценка эффективности извлечения основных результатов

Номер статьи	Количество основных формул (экспертная оценка)	Количество основных формул (оценка модели)	Точность	Полнота	F-мера	F-мера (сред)
1	9	7	1	0,777778	0,875	0,929401
2	11	10	1	0,909091	0,952381	
3	7	6	1	0,857143	0,923077	
4	8	5	1	0,625	0,769231	
5	8	10	0,8	1	0,888889	
6	7	7	1	1	1	
7	8	8	1	1	1	
8	9	10	0,9	1	0,947368	
9	8	6	1	0,75	0,857143	
10	9	7	1	0,777778	0,875	
11	5	4	1	0,8	0,888889	
12	8	10	0,8	1	0,888889	
13	7	9	0,777778	1	0,875	
14	9	9	1	1	1	
15	10	9	1	0,9	0,947368	
16	7	6	1	0,857143	0,923077	
17	7	7	1	1	1	
18	5	5	1	1	1	
19	9	8	1	0,888889	0,941176	
20	5	7	0,714286	1	0,833333	
21	4	5	0,8	1	0,888889	
22	7	10	0,7	1	0,823529	
23	7	5	1	0,714286	0,833333	
24	6	7	0,857143	1	0,923077	
25	6	6	1	1	1	
26	7	7	1	1	1	
27	8	8	1	1	1	
28	4	6	0,666667	1	0,8	
29	5	5	1	1	1	
30	7	4	1	0,571429	0,727273	
31	6	7	0,857143	1	0,923077	
32	8	8	1	1	1	
33	4	4	1	1	1	
34	6	6	1	1	1	
35	5	6	0,833333	1	0,909091	
36	5	5	1	1	1	
37	8	7	1	0,875	0,933333	
38	7	9	0,777778	1	0,875	
39	4	4	1	1	1	
40	6	6	1	1	1	
41	6	8	0,75	1	0,857143	
42	8	12	0,666667	1	0,8	
43	6	6	1	1	1	
44	8	6	1	0,75	0,857143	
45	7	7	1	1	1	
46	7	7	1	1	1	
47	6	6	1	1	1	
48	8	8	1	1	1	
49	4	4	1	1	1	
50	7	8	0,875	1	0,933333	

Рис. 9 – Оценка эффективности извлечения основных формул

Анализ полученных результатов свидетельствует, что модель DeerpSeek выдает релевантные результаты по главным результатам статей и связанным с ними формулам. Метрики точности, полноты и F-меры являются достаточно высокими, средние оценки F-меры по коллекции равны 0,93459 и 0,9294.

6. Концепция базы данных

Выделенные по коллекции статей основные результаты и формулы сохраняются в базе данных. Предложена следующая структура базы данных, содержащая таблицы статей, основных результатов, фрагментов текста основных результатов, формул и переменных формул.

Ниже даны описания столбцов соответствующих таблиц:

- Статья (атрибуты: номер (int), год издания(date), наименование журнала (varchar (100)), УДК (int), аннотация (varchar (700)))
- Результат (атрибуты: номер (int), номер статьи (int) текст (varchar (4000)))
- Фрагмент текста (атрибуты: номер (int), номер результата (int), текст (varchar (4000)))
- Формула (атрибуты: номер (int), номер фрагмента (int), текст (varchar (1000)))
- Переменная (атрибуты: номер (int), номер формулы (int), текст (varchar (20)))

Связи таблиц в базе данных задаются следующим образом:

- Статья содержит один или множество результатов. Результат может быть отнесен только к одной статье.
- Результат содержит один или множество фрагментов текста. Фрагмент текста может быть отнесен только к одному результату.
- Фрагмент текста содержит одну или множество формул. Формула может быть отнесена только к одному фрагменту текста.
- Формула содержит одну или множество переменных. Переменная может быть отнесена только к одной формуле.

Заключение

В статье предложен подход к выделению основных формул с помощью современных больших языковых моделей. С помощью модели DeerpSeek было проведено исследование по выделению основных результатов научных статей по аннотациям, на основании чего были выделены фрагменты из основного текста статьи и основные формулы, связанные с главными результатами.

Основной задачей создания семантического поисковика был вопрос хранения выделенных формул в машиночитаемом формате. В качестве

такого формата был выбран LaTeX. Для оценки релевантности ответов LLM использовались метрики точности, полноты и F-меры. Результаты исследования показывают, что современные языковые модели, подобные DeepSeek, успешно справляются с поставленными задачами, так среднее значение F-меры по тестовой коллекции равно 0,93459 для оценки эффективности извлечения основных результатов и равно 0,9294 для оценки эффективности извлечения основных формул.

Будущие исследования могут быть проведены в области написания программного кода, который позволял бы выделять главные формулы из текста загружаемых научных статей формата pdf с последующей их выдачей в формате LaTeX. Локализованный программный код мог бы снять с большой языковой модели задачи выделения основных результатов и формул. Это обеспечит более надёжный подход для решения поставленных задач.

Список литературы

1. Невзорова О.А. Николаев К.С. СЕМАНТИЧЕСКОЕ АННОТИРОВАНИЕ МАТЕМАТИЧЕСКИХ ФОРМУЛ В PDF-ДОКУМЕНТАХ // Электронные библиотеки. – 2022. – Том 25, № 6: Тематический выпуск по материалам XXIV Всероссийской научной конференции «Научный сервис в сети Интернет», 19–23 сентября 2022 года. Часть 1. DOI: <https://doi.org/10.26907/1562-5419-2022-25-6-616-639>
2. NIKOLAEV K., NEVZOROVA O. ANNOTATION OF MATHEMATICAL FORMULAS IN PDF DOCUMENTS // ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И НАНОТЕХНОЛОГИИ (ИТНТ-2023) сборник трудов по материалам IX Международной конференции и молодежной школы: в 6 т. Том 5. Самара, 2023
3. Ohashi S. et al. Efficient algorithm for math formula semantic search // IEICE TRANSACTIONS on Information and Systems. – 2016. – Т. 99. – №. 4. – С. 979-988.
4. Dhar S., Roy S., Das S. K. A critical survey of mathematical search engines // International Conference on Computational Intelligence, Communications, and Business Analytics. – Singapore: Springer Singapore, 2018. – С. 193-207.
5. Scharpf P. et al. Arqmath lab: An incubator for semantic formula search in zbmath open? // arXiv preprint arXiv:2012.02413. – 2020.
6. Zanibbi R., Blostein D. Recognition and retrieval of mathematical expressions // International Journal on Document Analysis and Recognition (IJДАР). – 2012. – Т. 15. – №. 4. – С. 331-357. DOI: 10.1007/s10032-011-0174-4
7. Kristianto G. Y. et al. Extracting main formulae from scientific documents using spatial and lexical features // Proceedings of the 15th ACM/IEEE-CS

- Joint Conference on Digital Libraries. – 2015. – С. 91-100. DOI: 10.1145/2756406.2756918
8. Li X. et al. Key formula extraction from scientific documents based on deep learning // IEEE Access. – 2020. – Т. 8. – С. 152380-152390. DOI: 10.1109/ACCESS.2020.3017217
 9. Yuan W. et al. A hybrid approach for extracting dominant formulae in academic papers // Proceedings of the 12th International Conference on Natural Language Processing. – 2015. – С. 210-218.
 10. Davila K., Zanibbi R. Layout and semantics: Combining representations for mathematical formula search // Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2017. – С. 1165-1168. DOI: 10.1145/3077136.3080741
 11. Lin, Xiao-Rong et al. "Mathematical Formula Identification in PDF Documents." 2011 International Conference on Document Analysis and Recognition (2011): 1419-1423.
 12. Богаевский И.А. Неявные обыкновенные дифференциальные уравнения: перестройки и усиление эквивалентности // Известия РАН. Серия математическая. – 2014. –Том 78, № 6. – С. 5–20.

References

1. Nevzorova O.A. Nikolaev K.S. SEMANTICHESKOE ANNOTIROVANIE MATEMATICHESKIH FORMUL V PDF-DOKUMENTAH // Elektronnye biblioteki. – 2022. – Том 25, № 6: Tematicheskij vypusk po materialam XXIV Vserossijskoj nauchnoj konferencii «Nauchnyj servis v seti Internet», 19–23 sentyabrya 2022 goda. Chast' 1. DOI: <https://doi.org/10.26907/1562-5419-2022-25-6-616-639>
2. NIKOLAEV K., NEVZOROVA O. ANNOTATION OF MATHEMATICAL FORMULAS IN PDF DOCUMENTS // ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И НАНОТЕХНОЛОГИИ (ИТНТ-2023) сборник трудов по материалам IX Международной конференции и молодежной школы: в 6 т.. Том 5. Самара, 2023
3. Ohashi S. et al. Efficient algorithm for math formula semantic search // IEICE TRANSACTIONS on Information and Systems. – 2016. – Т. 99. – №. 4. – С. 979-988.
4. Dhar S., Roy S., Das S. K. A critical survey of mathematical search engines // International Conference on Computational Intelligence, Communications, and Business Analytics. – Singapore: Springer Singapore, 2018. – С. 193-207.
5. Scharpf P. et al. Arqmath lab: An incubator for semantic formula search in zbmath open? // arXiv preprint arXiv:2012.02413. – 2020.
6. Zanibbi R., Blostein D. Recognition and retrieval of mathematical expressions // International Journal on Document Analysis and Recognition

(IJДАР). – 2012. – Т. 15. – №. 4. – С. 331-357. DOI: 10.1007/s10032-011-0174-4

7. Kristianto G. Y. et al. Extracting main formulae from scientific documents using spatial and lexical features // Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. – 2015. – С. 91-100. DOI: 10.1145/2756406.2756918
8. Li X. et al. Key formula extraction from scientific documents based on deep learning // IEEE Access. – 2020. – Т. 8. -- С. 152380-152390. DOI: 10.1109/ACCESS.2020.3017217
9. Yuan W. et al. A hybrid approach for extracting dominant formulae in academic papers // Proceedings of the 12th International Conference on Natural Language Processing. – 2015. – С. 210-218.
10. Davila K., Zanibbi R. Layout and semantics: Combining representations for mathematical formula search // Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2017. – С. 1165-1168. DOI: 10.1145/3077136.3080741
11. Lin, Xiao-Rong et al. “Mathematical Formula Identification in PDF Documents.” 2011 International Conference on Document Analysis and Recognition (2011): 1419-1423.
12. Bogaevskii I.A. Neqavnye obyknovennye differencial'nye uravneniya: perestrojki i usilenie ekvivalentnosti // Izvestiya RAN. Seriya matematicheskaya. – 2014. – Tom 78, № 6. – С. 5–20.