

ОБРАБОТКА СКАНИРОВАННЫХ МАТЕМАТИЧЕСКИХ PDF-ДОКУМЕНТОВ В СЕРВИСЕ СЕМАНТИЧЕСКОГО ПОИСКА ФОРМУЛ

К. С. Николаев^[0000-0003-3204-238X]

*Национальный исследовательский центр «Курчатовский институт»,
г. Москва, Россия*

konnikolaeff@yandex.ru

Аннотация

Рассмотрена задача извлечения семантических связей между математическими формулами и концептами из неструктурированных PDF-документов, включая сканированные. Предложен сквозной конвейер обработки, объединяющий оптическое распознавание, восстановление структуры параграфов, выделение формул и переменных, а также их связывание с концептами на основе синтаксических шаблонов русского языка. Разработаны формальные модели для двух типов связей: «переменная – концепт» и «переменная – главная формула». Проведена экспериментальная оценка качества на коллекции из 150 математических статей; достигнуто повышение точности связывания на 18% по сравнению с предшествующей версией. Описаны архитектурные решения прототипа веб-сервиса, обеспечивающие асинхронную обработку и масштабируемость.

Ключевые слова: *семантический поиск формул, PDF-документы, научные библиотеки, онтологии, веб-сервис.*

ВВЕДЕНИЕ

Обработка научных документов в формате PDF остается актуальной задачей для цифровых библиотек, особенно для сканированных материалов, где отсутствует текстовый слой. Такие документы требуют распознавания символов (OCR) [1] и восстановления логической структуры. Проблема усугубляется для математических текстов, где формулы содержат сложную двумерную структуру, а их семантическая интерпретация зависит от контекста.

Существующие подходы к преобразованию PDF (например, PDFX [2], CiTalO [3]) ориентированы на текстовые файлы. Для сканированных документов применяются нейросетевые детекторы [4] и OCR-системы [1], однако они не обеспечивают связывание формул с концептами. Гибридные методы семантической обработки [5] используют комбинацию текстового и структурного анализа, но также требуют качественного текстового слоя. Методы извлечения информации из PDF на основе правил [6] показывают хорошие результаты для структурированных документов, но при этом не адаптированы к математическому содержанию. Работы по обработке математических выражений [7–11] фокусируются на распознавании структуры, они не решают задачу установления смысловых связей между переменными и концептами в тексте. Методы извлечения изображений из научных статей [12] и поисковые системы, такие как Semantic Scholar [13], ориентированы на графические объекты, а не на семантику формул. Оценка качества OCR-систем для научных текстов [14] показывает, что даже современные решения допускают ошибки в математических нотациях. Методы семантической аннотации [15, 16] используют онтологии, но требуют структурированного входного формата.

В работе [17] был представлен прототип сервиса семантического поиска формул, имевший ограничения: поддержка только цифровых PDF, использование эвристик на основе расстояния для связывания, отсутствие асинхронности и авторизации. Настоящее исследование направлено на устранение этих недостатков, а также на формализацию и экспериментальную оценку методов связывания переменных с математическими концептами и главными формулами.

Научная задача состоит в разработке и оценке методов автоматического извлечения семантических связей между математическими объектами (переменными, формулами) и концептами из предметной области в неструктурированных документах, включая сканированные, с использованием лингвистических и структурных признаков.

Введем следующие определения:

- **концепт** – это единица знания из формальной онтологии (OntoMathPRO [18] или MSC [19]), имеющая уникальный идентификатор и название;
- **математический концепт** – концепт, относящийся к математической области (например, «группа», «производная»);

- главная формула – формула, выделенная в отдельный абзац, часто нумерованная, содержащая основное утверждение;
- локальная переменная – символ или выражение, встречающееся внутри текстового абзаца и используемое для обозначения параметра или вспомогательной величины;
- семантическая связь – это отношение между переменной и концептом или между переменной и главной формулой, устанавливаемое на основе контекстных и структурных правил.

Далее представлены архитектура прототипа сервиса, модуль оцифровки, формальные модели связывания, модуль визуализации и дана экспериментальная оценка.

АРХИТЕКТУРА ПРОТОТИПА СЕРВИСА

Разработанный прототип веб-сервиса построен как приложение на Flask с использованием асинхронной очереди задач Celery [20] и брокера Redis [21] для хранения состояния задач. Архитектура (рис. 1) включает следующие модули:

- авторизация – регистрация и аутентификация пользователей с подтверждением по электронной почте. Все операции, кроме поиска, требуют авторизации. Форма авторизации приведена на рис. 2 (слева);
- управление задачами – каждый пользователь имеет собственный набор задач с возможностью отслеживания статуса («выполняется», «завершена», «ошибка»). Страница управления задачами показана на рис. 2 (справа);
- обработка PDF – определяет тип документа (цифровой или сканированный) и запускает соответствующий конвейер;
- оцифровка – для сканированных документов добавляет текстовый слой и восстанавливает структуру параграфов;
- извлечение формул и переменных – выделяет главные формулы и локальные переменные;
- связывание – устанавливает связи между переменными и концептами, а также между переменными и главными формулами;
- визуализация – отображает семантическую структуру документа с возможностью фильтрации элементов.

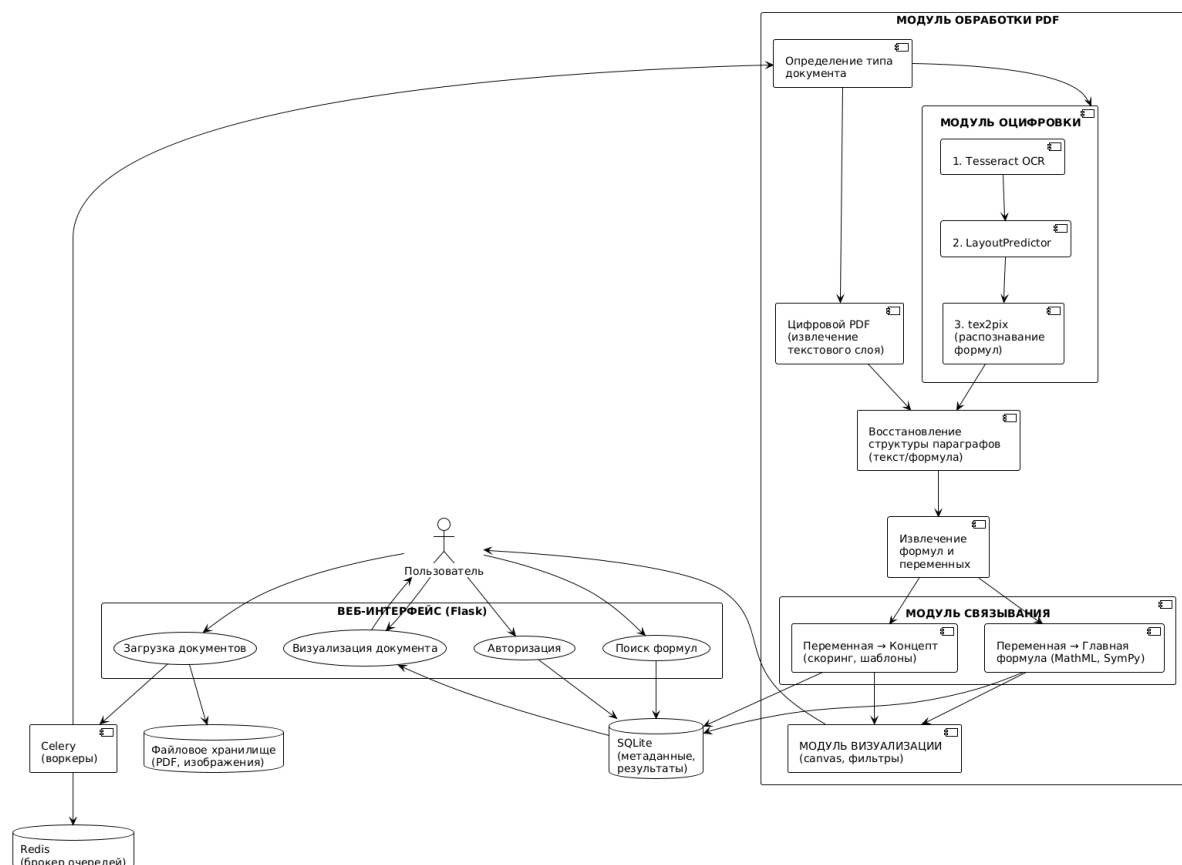


Рис. 1. Архитектура сервиса семантического поиска формул.

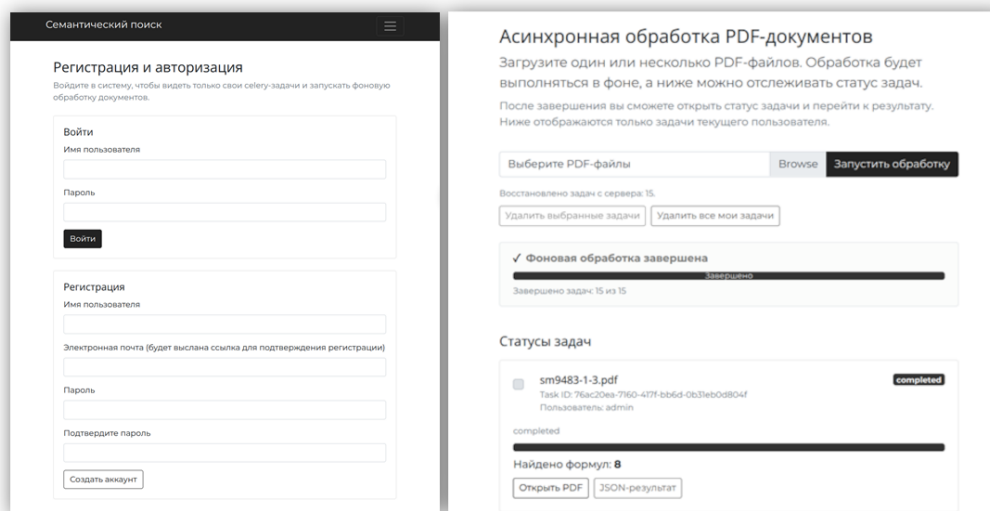


Рис. 2. Форма авторизации и регистрации с подтверждением через электронную почту (слева) и форма загрузки PDF-документов и отслеживания прогресса запущенных задач (справа).

Связь между модулями обработки и визуализации осуществляется через общую базу данных, где хранятся результаты аннотации.

Целевой аудиторией являются исследователи в области математики и компьютерных наук, сотрудники библиотек, издательства, заинтересованные в автоматической индексации и поиске по математическому содержанию.

Перечислим технические требования к вычислительной инфраструктуре: сервер с объемом оперативной памяти не менее 16 ГБ, хранилище на 500 ГБ; поддержка параллельной обработки до 10 документов одновременно; среднее время обработки одного документа – не более 5 мин для 20-страничной статьи.

Инфраструктура прототипа включает: веб-сервер на языке Python, базу данных SQLite для хранения метаданных и результатов, очередь задач Redis. Контейнеризация не используется для упрощения эксплуатации на этапе разработки.

Модуль поиска формул содержит поле поиска на основе коллекции математических концептов (рис. 3).

Семантический поиск

Семантический поиск формул

Этот сервис поможет найти научные документы. Основное отличие этого сервиса от других - это возможность поиска понятий в документах, независимо от их обозначения в тексте.

Коллекция документов состоит из многочисленных работ, размещенных на портале **Math-Net.Ru** и в коллекции журнала **Электронные библиотеки**

☐ Включить подклассы в результаты

Рис. 3. Форма поиска формул по математическим концептам.

МОДУЛЬ ОЦИФРОВКИ СКАНИРОВАННЫХ ДОКУМЕНТОВ

Модуль предназначен для преобразования сканированных PDF в структурированный текстовый формат. Процесс включает три этапа.

1. Распознавание текста – с помощью ocrmypdf (обертка над Tesseract OCR [1]) создается текстовый слой с координатами строк. Используются русский и английский языки.

2. Восстановление структуры параграфов – применена модель LayoutPredictor из библиотеки surya-ocr, обученная на размеченных научных статьях. Модель выделяет прямоугольники, соответствующие текстовым блокам, формулам, колонтитулам и т. д. Точность определения границ параграфов составляет 0.92 по метрике IoU на тестовой выборке. Визуализация работы LayoutPredictor представлена на рис. 4.

3. Группировка строк – координаты строк из результата OCR сопоставляются с прямоугольниками параграфов; строки, попавшие в один прямоугольник, объединяются в параграф. Каждому параграфу присваивается тип: «текст», «формула», «другое».

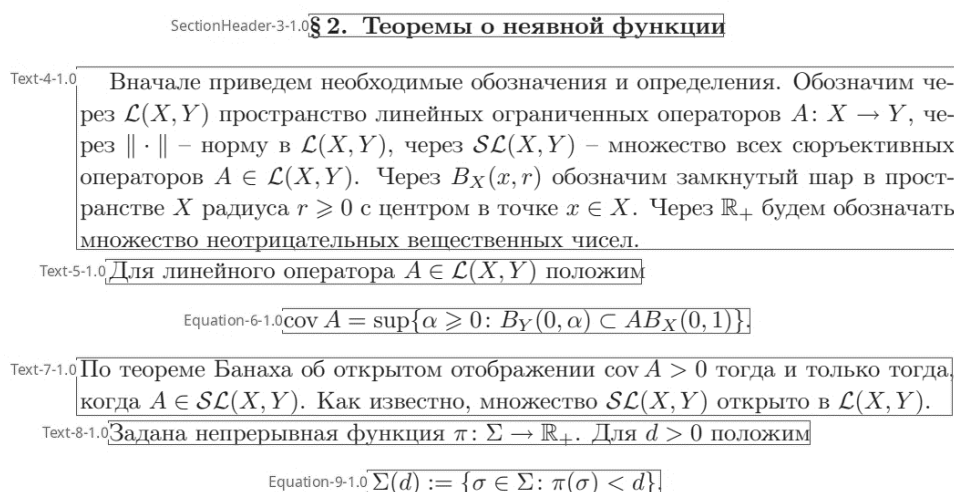


Рис. 4. Пример работы LayoutPredictor для определения границ параграфов и формул.

Для распознавания математических формул, выделенных на этапе 2, применяется библиотека tex2pix, которая принимает на вход изображение формулы и возвращает LaTeX-представление формулы с высокой точностью. Библиотека tex2pix основана на нейросетевой модели, обученной на размеченных наборах

изображений формул и соответствующих им LaTeX-последовательностей. Использование tex2pix позволяет значительно повысить качество распознавания сложных математических выражений по сравнению со стандартными OCR-системами, которые не специализированы на математической нотации. В рамках разработанного конвейера tex2pix вызывается для каждого блока, классифицированного как «формула»; полученное LaTeX-представление сохраняется в структурированном виде для дальнейшей семантической обработки, включая связывание переменных с концептами и главными формулами.

Ограничение метода состоит в том, что Tesseract не распознает сложные многоэтажные формулы (например, цепные дроби) с достаточной точностью. По результатам тестирования на 50 сканированных страницах с формулами среднего уровня сложности точность распознавания символов составила 0.78 для формул против 0.95 для обычного текста. Использование tex2pix позволяет повысить точность распознавания формул до 0.89 на том же наборе данных, что существенно улучшает качество последующего семантического связывания. Это, в свою очередь, накладывает ограничения на дальнейшее связывание: для документов с большим числом сложных формул качество работы снижается.

МОДУЛЬ ФОРМИРОВАНИЯ СЕМАНТИЧЕСКОЙ СТРУКТУРЫ

Формальная постановка задачи связывания переменных и концептов

Пусть имеются множество локальных переменных $V = \{v_1, \dots, v_n\}$, обнаруженных в тексте, и множество концептов $C = \{c_1, \dots, c_m\}$ из онтологии. Для каждой переменной v_i заданы ее левый и правый контексты – последовательности слов L_i и R_i (до и после переменной). Требуется установить бинарное отношение $S \subseteq V \times C$ такое, что $(v_i, c_j) \in S$, если концепт c_j семантически связан с переменной v_i .

Для решения использована функция $f(v_i, c_j)$, агрегирующая следующие признаки:

- вхождение названия концепта (в исходной или лемматизированной формах) в контекст: +1 балл;
- наличие маркеров определения в левом контексте («обозначим», «назовем», «введем обозначение», «через» и др.): +3 балла;

- наличие маркеров определения в правом контексте («это», «есть», «обозначает», «называется», «представляет собой»): +3 балла;
- одновременное присутствие концепта в правом контексте и маркера в левом («где», «здесь», «пусть»): +2 балла;
- присутствие концепта в контекстном окне (6 слов до и 6 после переменной): +2 балла.

Порог принятия решения $\tau = 3$: если $f(v_i, c_j) \geq 3$, то связь считается установленной.

Обоснование порога выбрано эмпирически на валидационной выборке из 30 документов как значение, максимизирующее F-меру.

Формальная постановка задачи связывания переменных и главных формул

Пусть имеется множество главных формул $F = \{f_1, \dots, f_k\}$, каждая из формул представлена в виде LaTeX-выражения. Для каждой локальной переменной v_i также известно ее LaTeX-представление. Требуется определить отображение $g: V \rightarrow F \cup \{\emptyset\}$, где $g(v_i) = f_j$, если переменная входит в состав формулы f_j как параметр или составная часть.

Используем три критерия для установления связи между главной формулой и переменной.

1. Строковое вхождение – LaTeX-код переменной является подстрокой кода формулы.
2. Сравнение символов в MathML – оба выражения конвертируются в MathML согласно спецификации [8]; извлекаются неделимые символы (идентификаторы, числа, операторы); если множество символов переменной есть подмножество символов формулы – то кандидат принимается.
3. Семантический разбор через SymPy – выражения парсятся с помощью `parse_latex` из библиотеки `sympy`; сравниваются не только символы, но и структура операторов и функции.

Связывание главной формулы и переменной производится, если выполняется хотя бы один из критериев. Приоритет отдается критерию 3 как наиболее полному.

МОДУЛЬ ВИЗУАЛИЗАЦИИ СЕМАНТИЧЕСКОЙ СТРУКТУРЫ

Для представления результатов разработан модуль визуализации на основе веб-шаблонов Flask, использующих элементы canvas. На странице документа (рис. 5) отображаются:

- локальные переменные;
- главные формулы;
- связи переменных с главными формулами.

Просмотр PDF: 610b1aеc-25e5-4d79-b1c6-9a0fe76c811c

- ☒ Показать переменные
- ☒ Показать формулы
- ☒ Показать концепты
- ☒ Показать связи

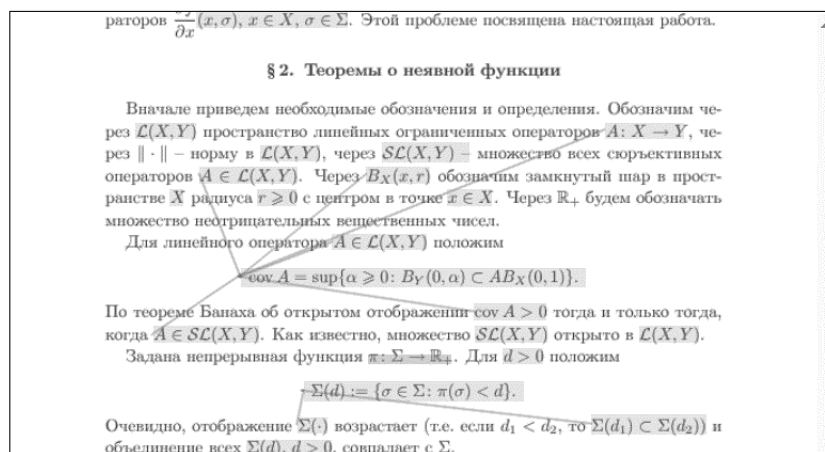


Рис. 5. Фрагмент страницы визуализации семантической структуры документа.

Пользователь может включать/отключать отображение каждой группы элементов через панель управления. Такой подход обеспечивает единообразие для всех браузеров, в отличие от встроенных PDF-аннотаций, поддержка которых различается в разных программах просмотра.

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА

Датасет и метрики

Для оценки качества была использована коллекция из 150 математических статей из открытых архивов, из них 50 – сканированные (без текстового слоя).

Все статьи размечены вручную по трем классам связей: переменная – концепт, переменная – главная формула.

Перечислим выбранные метрики:

- точность (Precision) – доля верно установленных связей среди всех установленных системой;
- полнота (Recall) – доля верно установленных связей среди всех истинных связей в разметке;
- F-мера – гармоническое среднее точности и полноты.

Результаты связывания переменных и концептов

В таблице 1 представлено сравнение качества связывания переменных с математическими концептами для двух подходов: эвристики на основе расстояния, использовавшейся в предыдущей версии прототипа, и предложенного скорингового метода, основанного на синтаксических шаблонах русского языка. Оценка выполнена на всей коллекции из 150 документов.

Табл. 1. Значения метрик для различных методов.

Метод	Precision	Recall	F-мера
Расстояние (по [17])	0.62	0.58	0.60
Предложенный скоринговый метод	0.81	0.79	0.80

Прирост F-меры составил 20 процентных пунктов. Для сканированных документов точность ниже (0.72) из-за ошибок OCR, однако метод сохраняет работоспособность благодаря использованию лемматизации и множественных контекстных признаков.

Результаты связывания переменных и главных формул

В таблице 2 приведено сравнение двух подходов к связыванию локальных переменных с главными формулами: с использованием только сравнения символов в MathML и с добавлением семантического разбора через SymPy. Использование SymPy позволяет учитывать не только символьный состав формулы, но и её структуру, что особенно важно для различения семантически разных выражений с одинаковыми символами (например, $f(x)$ и $f \cdot x$).

Табл. 2. Сравнение значений метрик точности, полноты и F-меры для различных подходов к связыванию переменных и главных формул.

Подход	Precision	Recall	F-мера
Только сравнение MathML	0.68	0.65	0.66
С преобразованием в SymPy	0.85	0.82	0.83

Использование SymPy существенно улучшает распознавание параметров формул, обеспечивая прирост F-меры на 17 процентных пунктов.

Границы применимости

На основе анализа ошибок определены следующие границы применимости разработанных методов.

Система эффективна (F-мера ≥ 0.70) для документов, где

- OCR-качество текста ≥ 0.90 (по символьной точности);
- формулы имеют стандартную структуру (не содержат редких нотаций, например нотации Кнута или специальных операторов);
- язык документа – русский.

Для документов с большим количеством сложных формул (например, многоуровневые индексы, матрицы, длинные цепные дроби) качество падает до 0.55–0.60. Это связано с ограничениями Tesseract при распознавании двумерных структур и ограничениями парсера LaTeX в SymPy. Использование tex2pix для распознавания формул позволяет повысить точность распознавания до 0.89 для формул среднего уровня сложности, однако для крайне сложных выражений (например, многострочные системы уравнений или матрицы большого размера) точность tex2pix также снижается до 0.72–0.78, что приводит к ошибкам в извлечении LaTeX-представления и последующем семантическом связывании. Кроме того, tex2pix, как и любая нейросетевая модель, чувствителен к качеству входного изображения: низкое разрешение, шумы, артефакты сканирования или наклон страницы могут существенно ухудшить результат распознавания. Перспективным решением видится использование специализированных моделей распознавания формул на основе графов [11] или нейросетевых подходов с увеличенным объемом обучающих данных для сложных нотаций.

Система также не применима для документов, содержащих формулы исключительно в виде изображений без текстового контекста, поскольку для связывания необходим лингвистический контекст. Для таких документов требуется дополнительный этап извлечения контекстных предложений из окружающего текста, что не всегда возможно в случае отсутствия связного текстового описания формулы.

ЗАКЛЮЧЕНИЕ

Решена задача создания сквозного конвейера для семантической обработки математических PDF-документов, включая сканированные. Предложены формальные модели и методы связывания переменных с математическими концептами и главными формулами. Разработанный конвейер объединяет несколько ключевых компонентов: оптическое распознавание текста на основе Tesseract, восстановление логической структуры документа с помощью LayoutPredictor, распознавание математических формул с использованием tex2pix, а также лингвистический анализ для установления семантических связей на основе синтаксических шаблонов русского языка.

Экспериментально показано значительное повышение качества по сравнению с предыдущей версией прототипа: прирост F-меры для связывания переменной и концепта составил 20 п.п. (с 0.60 до 0.80), для связывания переменных с главными формулами – 17 п.п. (с 0.66 до 0.83). Внедрение tex2pix позволило повысить точность распознавания формул в сканированных документах с 0.78 до 0.89 для выражений среднего уровня сложности, что непосредственно улучшило качество последующего семантического связывания. Определены границы применимости: система эффективна для документов со стандартной структурой формул и качеством OCR не ниже 0.90; для сложных двумерных формул (многострочные системы, матрицы большого размера) качество снижается, что требует дальнейших исследований.

Архитектурные решения (асинхронная очередь задач на базе Celery и Redis, модульная структура, система авторизации) обеспечивают масштабируемость и готовность прототипа к промышленному развертыванию. Разработан-

ный модуль визуализации с использованием canvas предоставляет интерактивный инструмент для анализа семантической структуры документов, что может быть полезно для исследователей и сотрудников библиотек.

Благодарности

Работа выполнена в рамках государственного задания НИЦ «Курчатовский институт».

СПИСОК ЛИТЕРАТУРЫ

1. *Smith R.* An overview of the tesseract OCR engine // Proc. Int. Conf. Doc. Anal. Recognition, ICDAR. 2007. Vol. 2. P. 629–633.
<https://doi.org/10.1109/ICDAR.2007.4376991>
2. *Constantin A., Pettifer S., Voronkov A.* PDFX: Fully-automated PDF-to-XML conversion of scientific literature // DocEng 2013 – Proc. 2013 ACM Symp. Doc. Eng. 2013. P. 177–180. <https://doi.org/10.1145/2494266.2494271>
3. *Ciancarini P., Iorio A. Di, Nuzzolese A. G., Peroni S., Vitali F.* Semantic annotation of scholarly documents and citations // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2013. Vol. 8249 LNAI. P. 336–347. https://doi.org/10.1007/978-3-319-03524-6_29
4. *Amethiya Y., Bajwa G.* Automatic Table Detection and Tabular Data Extraction from Scanned Documents // Smart Innovation, Systems and Technologies. 2025. P. 75–87. https://doi.org/10.1007/978-981-96-0143-1_7
5. *Bertin M., Atanassova I.* Hybrid Approach for the Semantic Processing of Scientific Papers // Semant. Publ. Chall. Track 11 th Eur. Semant. Web Conf. (ESWC 2014). 2014. https://doi.org/10.1007/978-981-96-0143-1_7
6. *Ahmad R., Afzal M. T., Qadir M. A.* Information extraction from PDF sources based on rule-based system using integrated formats // Commun. Comput. Inf. Sci. 2016. Vol. 641. P. 293–308. https://doi.org/10.1007/978-3-319-46565-4_23
7. *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 // Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2018. P. 233–242. <https://doi.org/10.1145/3197026.3197058>

8. Mathematical Markup Language (MathML) Version 3.0 2nd Edition. W3C Recommendation, 10 April 2014. URL: <https://www.w3.org/TR/MathML3/>
9. Greiner-Petter A., Youssef A., Ruas T., Miller B. R., Schubotz M., Aizawa A., Gipp B. Math-word embedding in math search and semantic extraction // *Scien-tometrics*. 2020. Vol. 125, No. 3. P. 3017–3046.
<https://doi.org/10.1007/s11192-020-03502-9>
10. Nevzorova O., Kirillovich A., Nevzorov V., Nikolaev K. The semantic con-text models of mathematical formulas in scientific papers // *CEUR Workshop Proc.* 2018. Vol. 2277. P. 33–40.
11. Shah A. K., Dey A., Zanibbi R. A Math Formula Extraction and Evaluation Framework for PDF Documents // *Lecture Notes in Computer Science*. 2021. P. 19–34. https://doi.org/10.1007/978-3-030-86331-9_2
12. Siegel N., Lourie N., Power R., Ammar W. Extracting Scientific Figures with Distantly Supervised Neural Networks // *Proc. ACM/IEEE Jt. Conf. Digit. Libr.* 2018. P. 223–232. <https://doi.org/10.1145/3197026.3197040>
13. Semantic Scholar. URL: <https://www.semanticscholar.org/>
14. Neudecker C., Baierer K., Gerber M., Clausner C., Antonacopoulos A., Pletschacher S. A survey of OCR evaluation tools and metrics // *The 6th International Workshop on Historical Document Imaging and Processing*. New York, NY, USA: ACM, 2021. P. 13–18. <https://doi.org/10.1145/3476887.3476888>
15. Jaradeh M. Y., Oelen A., Farfar K. E., Prinz M., D'Souza J., Kismihók G., Stocker M., Auer S. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge // *K-CAP 2019 – Proceedings of the 10th Interna-tional Conference on Knowledge Capture*. New York, NY, USA: ACM, 2019. P. 243–246. <https://doi.org/10.1145/3360901.3364435>
16. Peroni S., Shotton D. FaBiO and CiTO: Ontologies for describing biblio-graphic resources and citations // *J. Web Semant.* 2012. Vol. 17. P. 33–43.
<https://doi.org/10.1016/j.websem.2012.08.001>
17. Николаев К.С. Сервис семантического поиска формул по коллекции математических PDF-документов // *Информационные технологии и вычисли-тельные системы*. 2025. № 3. С. 34–43. <https://doi.org/10.14357/20718632250304>

18. Kirillovich A. V., Nevzorova O. A., Lipachev E. K. OntoMathPRO 2.0 Ontology: Updates of Formal Model // Lobachevskii J. Math. 2022. Vol. 43, No. 12. P. 3504–3514. <https://doi.org/10.1134/S1995080222150136>
 19. MSC2020 –Mathematics Subject Classification System.
URL: <https://msc2020.org/>.
 20. Celery –Distributed Task Queue. URL: <https://docs.celeryq.dev/>.
 21. Redis – The real-time context engine for AI apps. URL: <https://redis.io/>.
-

PROCESSING OF SCANNED MATHEMATICAL PDF DOCUMENTS IN THE SEMANTIC FORMULA SEARCH SERVICE

K. S. Nikolaev^[0000-0003-3204-238X]

National Research Centre "Kurchatov Institute", Moscow, Russia

konnikolaeff@yandex.ru

Abstract

The article addresses the problem of extracting semantic relations between mathematical formulas and concepts from unstructured PDF documents, including scanned ones. A complete processing pipeline is proposed that combines optical character recognition, paragraph structure recovery, formula and variable extraction, and their linking to concepts using syntactic patterns of the Russian language. Formal models are developed for two types of relations: "variable-concept" and "variable-main formula". Experimental evaluation on a collection of 150 mathematical articles shows a significant improvement in linking accuracy, with an increase of 18 percentage points in F-measure compared to the previous version. Architectural improvements of the web service prototype are described, providing asynchronous processing and scalability.

Keywords: *semantic formula search, PDF documents, scientific libraries, ontologies, web service.*

REFERENCES

1. *Smith R.* An overview of the tesseract OCR engine // Proc. Int. Conf. Doc. Anal. Recognition, ICDAR. 2007. Vol. 2. P. 629–633.
<https://doi.org/10.1109/ICDAR.2007.4376991>
2. *Constantin A., Pettifer S., Voronkov A.* PDFX: Fully-automated PDF-to-XML conversion of scientific literature // DocEng 2013 – Proc. 2013 ACM Symp. Doc. Eng. 2013. P. 177–180. <https://doi.org/10.1145/2494266.2494271>
3. *Ciancarini P., Iorio A. Di, Nuzzolese A. G., Peroni S., Vitali F.* Semantic annotation of scholarly documents and citations // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2013. Vol. 8249 LNAI. P. 336–347. https://doi.org/10.1007/978-3-319-03524-6_29
4. *Amethiya Y., Bajwa G.* Automatic Table Detection and Tabular Data Extraction from Scanned Documents // Smart Innovation, Systems and Technologies. 2025. P. 75–87. https://doi.org/10.1007/978-981-96-0143-1_7
5. *Bertin M., Atanassova I.* Hybrid Approach for the Semantic Processing of Scientific Papers // Semant. Publ. Chall. Track 11 th Eur. Semant. Web Conf. (ESWC 2014). 2014. https://doi.org/10.1007/978-981-96-0143-1_7
6. *Ahmad R., Afzal M. T., Qadir M. A.* Information extraction from PDF sources based on rule-based system using integrated formats // Commun. Comput. Inf. Sci. 2016. Vol. 641. P. 293–308. https://doi.org/10.1007/978-3-319-46565-4_23
7. *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 // Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2018. P. 233–242. <https://doi.org/10.1145/3197026.3197058>
8. Mathematical Markup Language (MathML) Version 3.0 2nd Edition. W3C Recommendation, 10 April 2014. URL: <https://www.w3.org/TR/MathML3/>
9. *Greiner-Petter A., Youssef A., Ruas T., Miller B. 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.
<https://doi.org/10.1007/s11192-020-03502-9>

10. *Nevzorova O., Kirillovich A., Nevzorov V., Nikolaev K.* The semantic context models of mathematical formulas in scientific papers // CEUR Workshop Proc. 2018. Vol. 2277. P. 33–40.
11. *Shah A. K., Dey A., Zanibbi R.* A Math Formula Extraction and Evaluation Framework for PDF Documents // Lecture Notes in Computer Science. 2021. P. 19–34. https://doi.org/10.1007/978-3-030-86331-9_2
12. *Siegel N., Lourie N., Power R., Ammar W.* Extracting Scientific Figures with Distantly Supervised Neural Networks // Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2018. P. 223–232. <https://doi.org/10.1145/3197026.3197040>
13. Semantic Scholar. URL: <https://www.semanticscholar.org/>
14. *Neudecker C., Baierer K., Gerber M., Clausner C., Antonacopoulos A., Pletschacher S.* A survey of OCR evaluation tools and metrics // The 6th International Workshop on Historical Document Imaging and Processing. New York, NY, USA: ACM, 2021. P. 13–18. <https://doi.org/10.1145/3476887.3476888>
15. *Jaradeh M. Y., Oelen A., Farfar K. E., Prinz M., D'Souza J., Kismihók G., Stocker M., Auer S.* Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge // K-CAP 2019 – Proceedings of the 10th International Conference on Knowledge Capture. New York, NY, USA: ACM, 2019. P. 243–246. <https://doi.org/10.1145/3360901.3364435>
16. *Peroni S., Shotton D.* FaBiO and CiTO: Ontologies for describing bibliographic resources and citations // J. Web Semant. 2012. Vol. 17. P. 33–43. <https://doi.org/10.1016/j.websem.2012.08.001>
17. *Nikolaev K.S.* Servis semanticheskogo poiska formul po kollekcii matematicheskikh PDF-dokumentov // Informacionnye tehnologii i vychislitel'nye sistemy. 2025. № 3. S. 34–43. <https://doi.org/10.14357/20718632250304>
18. *Kirillovich A. V., Nevzorova O. A., Lipachev E. K.* OntoMathPRO 2.0 Ontology: Updates of Formal Model // Lobachevskii J. Math. 2022. Vol. 43, No. 12. P. 3504–3514. <https://doi.org/10.1134/S1995080222150136>
19. MSC2020 –Mathematics Subject Classification System. URL: <https://msc2020.org/>.
20. Celery –Distributed Task Queue. URL: <https://docs.celeryq.dev/>.
21. Redis – The real-time context engine for AI apps. URL: <https://redis.io/>.

СВЕДЕНИЯ ОБ АВТОРЕ



НИКОЛАЕВ Константин Сергеевич – младший научный сотрудник Лаборатории суперкомпьютерного моделирования Отделения суперкомпьютерных систем и параллельных вычислений ФГБУ «Национальный исследовательский центр «Курчатовский институт». Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

NIKOLAEV Konstantin Sergeevich – junior researcher at the Supercomputer Modeling Laboratory of the Department of Supercomputer Systems and Parallel Computing of the Federal State Budgetary Institution "National Research Center "Kurchatov Institute". Major fields of scientific research are natural language processing, artificial intelligence.

email: konnikolaeff@yandex.ru

ORCID: 0000-0003-3204-238X

Материал поступил в редакцию 30 июня 2026 года