DOI 10.37882/2223-2966.2026.01.05

# МЕТОДИКА НАПОЛНЕНИЯ МОДЕЛИ ЦИФРОВОГО ОБЪЕКТА ИЗ РАЗНОРОДНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ НА ОСНОВЕ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ МЕТОДОВ ИЗВЛЕЧЕНИЯ ДАННЫХ

## METHODOLOGY FOR POPULATING A DIGITAL OBJECT MODEL FROM HETEROGENEOUS INFORMATION RESOURCES BASED ON QUANTITATIVE ASSESSMENT OF DATA EXTRACTION METHODS EFFICIENCY

***A. Artamonov***

*Summary*. The increasing volume of unstructured scientific data requires the development of effective methodologies for automated extraction and structuring of information. The relevance of the study is determined by the need to create holistic models of digital objects for subsequent comprehensive analysis under conditions of data source heterogeneity. The problem lies in the absence of a quantitatively substantiated methodology for selecting optimal approaches to data extraction from various types of documents while ensuring required reliability of results. The purpose of this work is to develop and verify a methodology for populating a digital object model from heterogeneous information resources with quantitative assessment of the effectiveness of applied data extraction methods using scientific publications as an example.

*Keywords*: data extraction, digital object, scientific publications, NLP, reliability assessment, NoSQL, document-oriented databases.

***Артамонов Алексей Анатольевич***
*Кандидат технических наук,*
*Национальный исследовательский ядерный университет «МИФИ», Москва*
*aaartamonov@mephi.ru*

*Аннотация*. Возрастающий объем неструктурированных научных данных требует разработки эффективных методик автоматизированного извлечения и структурирования информации. Актуальность исследования обусловлена необходимостью создания целостных моделей цифровых объектов для последующего многостороннего анализа в условиях гетерогенности источников данных. Проблема заключается в отсутствии количественно обоснованной методологии выбора оптимальных подходов к извлечению данных из различных типов документов при обеспечении требуемой достоверности результатов. Цель работы — разработать и верифицировать методику наполнения модели цифрового объекта из разнородных информационных ресурсов с количественной оценкой эффективности применяемых методов извлечения данных на примере научных публикаций.

*Ключевые слова*: извлечение данных, цифровой объект, научные публикации, NLP, оценка достоверности, NoSQL, документоориентированные базы данных.

## Введение

Стремительное накопление научных данных в цифровой среде создало фундаментальный вызов для современных информационных систем: как эффективно преобразовать разрозненные неструктурированные данные в целостные модели цифровых объектов, пригодные для аналитической обработки. По данным Scopus, количество научных публикаций увеличилось с 2,1 млн в 2010 году до 4,8 млн в 2023 году, что представляет собой рост в 2,3 раза за 13 лет [1]. Этот экспоненциальный рост сопровождается увеличением гетерогенности источников данных: научные публикации представлены в форматах PDF, HTML, XML, DOCX, причем каждый формат требует специфических методов обработки [2, 3].

Актуальность данного исследования обусловлена тем, что существующие подходы к извлечению данных из научных публикаций не предоставляют количественно обоснованной методологии выбора оптимальных методов для различных типов документов. Большинство исследований фокусируются на описании отдельных техник извлечения, не проводя систематического сравнительного анализа их эффективности [4, 5]. Между тем, выбор неоптимального метода может приводить к потере до 30–40 % ценной информации или требовать в 5–10 раз больше вычислительных ресурсов [6]. Проблема усугубляется необходимостью оценки достоверности извлекаемых данных, особенно в контексте критически важных технологических областей. Неверная или недостоверная научная информация может привести к ошибочным решениям на уровне государственной по-

литики в области науки и технологий [7]. Существующие библиометрические показатели (импакт-фактор, индекс Хирша, квартиль журнала) используются разрозненно, отсутствует интегральная метрика, позволяющая автоматизированно оценивать надежность источника с учетом множественных факторов [8, 9].

Таким образом, исследование направлено на разработку комплексной методики, которая не только описывает процесс извлечения и структурирования данных, но и предоставляет количественные критерии выбора оптимальных подходов для различных сценариев обработки, а также инструментарий для автоматизированной оценки достоверности полученной информации. Это позволит повысить качество формируемых баз знаний и эффективность последующего аналитического процесса.

## Материалы и методы исследования

В качестве эмпирической базы для настоящего исследования был использован корпус из 2847 научных публикаций, охватывающий период с 2020 по 2024 год и представленный в различных форматах. Распределение документов по форматам составило: PDF — 1423 документа (50,0 %), HTML — 768 документов (27,0 %), структурированный XML (включая JATS) — 456 документов (16,0 %), смешанные и прочие форматы — 200 документов (7,0 %). Публикации были отобраны из баз данных Scopus, Web of Science Core Collection и Российского индекса научного цитирования (РИНЦ) по тематическим направлениям информационных технологий, системного анализа и обработки данных. Для оценки качества извлечения данных использовалась контрольная выборка из 500 публикаций с ручной разметкой, выполненной экспертами предметной области. Разметка включала выделение всех целевых элементов: заголовков, авторов с аффилиациями, аннотаций, ключевых слов, основного текста, таблиц, изображений, библиографических списков. Метрики качества рассчитывались на основе стандартных показателей: точность (precision), полнота (recall) и F1-мера [10]. Методология исследования носила комплексный экспериментально-аналитический характер. Первым этапом была реализация и тестирование трех основных подходов к извлечению данных:

Rule-based подход — использование регулярных выражений и шаблонов для извлечения структурированных элементов. Реализован на базе библиотек Python (re, BeautifulSoup, lxml). Гибридный подход — комбинация правил с методами машинного обучения для повышения адаптивности. Использовались предобученные модели BERT и RuBERT для классификации разделов и именованного распознавания сущностей (NER).

Векторизованный подход — применение техник word embeddings и трансформеров для семантического анализа и извлечения. Реализован с использованием библиотек transformers и sentence-transformers. Вторым этапом проводился сравнительный анализ производительности и качества работы каждого подхода на различных типах документов. Измерялись следующие параметры: время обработки одного документа, точность извлечения метаданных, полнота извлечения текстового контента, способность к масштабированию (обработка пакетов документов) [11]. Третьим этапом была разработка интегрального показателя достоверности научной публикации на основе библиометрических критериев. Показатель рассчитывался по формуле:

$$D = w_1 Q^{\{-1\}} + w_2 H_{\{norm\}} + w_3 C_{\{max\}} + w_4 Aff_{\{max\}}$$

где Q — квартиль журнала (обратное значение: Q1=4, Q2=3, Q3=2, Q4=1); $H_{\{norm\}}$ — нормализованный индекс Хирша ведущего автора; $C_{\{max\}}$ — максимальный коэффициент авторитетности страны; $Aff_{\{max\}}$ — максимальный коэффициент престижности организации; $w_i$ — весовые коэффициенты, определяемые методом главных компонент, $\mathbf{\Sigma}_{\{i=1\}}^{\{4\}} w_i = 1$ [12].

Для определения весовых коэффициентов была собрана обучающая выборка из 300 публикаций с экспертными оценками достоверности (шкала 0–100). Применялся метод главных компонент (PCA) с последующей линейной регрессией для минимизации среднеквадратичной ошибки между расчетным показателем и экспертной оценкой.

Организация хранения данных осуществлялась в документоориентированной NoSQL базе данных MongoDB с использованием разработанной JSON-схемы. Схема включала поля для всех типов метаданных, текстового контента, извлеченных сущностей, геолокационных данных и временных меток. Для полнотекстового поиска дополнительно использовалась система Elasticsearch с настроенными анализаторами для русского и английского языков [13].

## Результаты и обсуждение

Анализ эффективности различных подходов к извлечению данных требует перехода от качественных описаний к строгим количественным метрикам. Центральным вопросом исследования является определение оптимального метода извлечения для различных типов документов с учетом баланса между точностью, полнотой и производительностью. Интуитивное предположение о превосходстве сложных методов машинного обучения над простыми правилами требует эмпирической верификации на репрезентативной выборке.

Для объективной оценки было проведено систематическое сравнение трех подходов (rule-based, гибридный,

Таблица 1.

Сравнительная эффективность методов извлечения базовых метаданных из документов различных форматов

| Тип документа | Rule-based подход | | Гибридный подход | | Векторизованный подход | |
|---|---|---|---|---|---|---|
| | Точность, % | Полнота, % | Точность, % | Полнота, % | Точность, % | Полнота, % |
| Структурированный XML (JATS) | 94,8 | 89,3 | 93,1 | 92,7 | 89,4 | 91,8 |
| HTML с разметкой | 87,3 | 72,5 | 90,8 | 88,9 | 92,1 | 94,3 |
| PDF с распознаванием текста | 76,2 | 61,8 | 89,5 | 85,3 | 91,7 | 89,6 |
| PDF отсканированные документы | 58,4 | 45,2 | 82,6 | 78,9 | 87,3 | 85,1 |
| Средневзвешенное значение | 82,1 | 69,8 | 89,7 | 86,7 | 90,5 | 90,4 |

векторизованный) по метрикам точности и полноты извлечения базовых метаданных из документов различных форматов. Гипотеза исследования заключалась в том, что эффективность метода существенно зависит от степени структурированности исходного документа, и не существует универсального оптимального подхода для всех типов данных. Количественные результаты, представленные в таблице 1, призваны верифицировать эту гипотезу.

Анализ данных таблицы 1 выявляет сложную зависимость эффективности извлечения от степени структурированности исходных документов и выбранного метода. Для высокоструктурированных XML-документов rule-based подход демонстрирует максимальную точность (94,8 %), что объясняется наличием четких тегов разметки и стандартизованной структуры JATS XML. Однако полнота извлечения составляет только 89,3 %, поскольку жесткие правила не способны обработать документы с отклонениями от стандарта.

Гибридный подход показывает более сбалансированные результаты: незначительное снижение точности до 93,1 % компенсируется повышением полноты до 92,7 %. Математически это выражается в росте F1-меры с 91,96 до 92,90, что свидетельствует о лучшей общей эффективности. Векторизованный подход демонстрирует снижение точности до 89,4 % из-за вероятностной природы нейросетевых предсказаний, но обеспечивает стабильную полноту 91,8 %. Наиболее драматические различия наблюдаются при обработке слабоструктурированных PDF-документов. Для отсканированных PDF rule-based методы показывают точность всего 58,4 % и полноту 45,2 %, что делает их практически неприменимыми. Гибридный подход повышает метрики до 82,6 % и 78,9 % соответственно, увеличивая F1-меру с 50,98 до 80,69 — рост на 58,3 %. Векторизованный подход достигает максимальных показателей: 87,3 % точности и 85,1 % полноты, что дает F1-меру 86,19. Средневзвешенные значения (с учетом распределения типов документов в выборке) показывают преимущество современных подходов: переход от rule-based к гибридному методу повышает точность на 9,3 % и полноту на 24,2 %, а векторизованный подход обеспечивает дальнейший прирост до 90,5 % и 90,4 % соответственно. Критически важно, что выигрыш сложных методов максимален именно для наиболее проблемных типов документов, что оправдывает их применение в гетерогенных корпусах.

Однако высокая эффективность извлечения данных теряет практическую ценность, если время обработки делает систему неприменимой для работы с большими объемами информации. Поэтому необходимо оценить производительность каждого подхода и выявить зависимость времени обработки от сложности метода и объема обрабатываемых данных (табл. 2).

Таблица 2.

Производительность методов извлечения данных и масштабируемость при обработке корпусов различного объема

| Метод извлечения | Время обработки 1 документа, сек | Время обработки пакета (100 док.), сек | Время обработки корпуса (1000 док.), сек | Коэффициент масштабируемости |
|---|---|---|---|---|
| Rule-based подход | 1,3 | 118,7 | 1247,3 | 0,96 |
| Гибридный подход (CPU) | 8,7 | 782,1 | 8356,8 | 0,96 |
| Гибридный подход (GPU) | 8,7 | 247,3 | 1893,4 | 0,22 |
| Векторизованный (CPU) | 12,4 | 1098,5 | 11782,9 | 0,95 |
| Векторизованный (GPU) | 12,4 | 156,2 | 982,7 | 0,08 |

*Примечание*: коэффициент масштабируемости рассчитан как $(T_{\{1000\}} / (T_1\ 1000))–1$, где отрицательные значения указывают на сверхлинейное ускорение при пакетной обработке.

Данные таблицы 2 демонстрируют критическую важность выбора вычислительной архитектуры для сложных методов извлечения. Rule-based подход показывает наименьшее время обработки одного документа (1,3 секунды) благодаря простоте операций. Коэффициент масштабируемости 0,96 указывает на практически линейную зависимость времени от объема данных, что является ожидаемым для последовательной обработки. Гибридный подход требует значительно больше времени на один документ (8,7 секунд) из-за применения моделей машинного обучения. При CPU-вычислениях обработка корпуса из 1000 документов занимает 8356,8 секунд (около 2,3 часов), что может быть неприемлемо для оперативных задач. Однако использование GPU-ускорения снижает время до 1893,4 секунд (31,6 минуты) — сокращение в 4,4 раза. Коэффициент масштабируемости 0,22 для GPU-версии показывает существенный выигрыш от пакетной обработки.

Наиболее впечатляющие результаты демонстрирует векторизованный подход при GPU-ускорении. Несмотря на максимальное время обработки одного документа (12,4 секунды), пакетная обработка 1000 документов занимает всего 982,7 секунды (16,4 минуты) благодаря эффективному распараллеливанию матричных операций. Коэффициент масштабируемости 0,08 означает, что обработка корпуса требует менее 8 % времени от наивной оценки $T_1$ 1000, что свидетельствует о почти десятикратном ускорении.

Математический анализ отношения качества к производительности показывает, что оптимальный выбор метода зависит от конкретных требований задачи. Для задач реального времени с высокоструктурированными данными rule-based подход остается конкурентоспособным. Для корпусных исследований с гетерогенными документами векторизованный подход с GPU-ускорением обеспечивает наилучший баланс: максимальное качество (F1 = 90,45) при приемлемой производительности (менее 1 секунды на документ в пакетном режиме).

Высокая эффективность извлечения данных создает основу для построения репрезентативных моделей цифровых объектов, однако практическая ценность собранной информации критически зависит от ее достоверности. Особую важность оценка надежности приобретает в контексте научно-технической информации, где ошибочные данные могут привести к неверным исследовательским или управленческим решениям. Разработанный интегральный показатель достоверности требует эмпирической валидации для подтверждения его практической применимости (табл. 3).

Анализ данных таблицы 3 показывает критическую важность правильного выбора весовых коэффициентов для интегрального показателя достоверности. Наивный подход с равными весами ($w_i$ = 0,25) демонстрирует корреляцию 0,67 с экспертными оценками и RMSE 18,4, что указывает на недостаточную прогностическую силу такой модели. Среднеквадратичная ошибка 18,4 на шкале 0–100 означает, что предсказания могут отклоняться от экспертной оценки почти на 20 %, что неприемлемо для практического применения.

Подход с весами, предложенными экспертами предметной области ($w_{1=0}$,40, $w_{2=0}$,35), основан на их интуитивном понимании относительной важности критериев. Этот набор повышает корреляцию до 0,79 и снижает RMSE до 12,7, что представляет собой улучшение на 31 % по сравнению с равными весами. Эксперты придают наибольшее значение квартилю журнала и индексу Хирша автора, рассматривая географические и институциональные факторы как второстепенные. Оптимизация весов методом главных компонент с последующей регрессией дает наилучшие результаты: корреляция 0,84 и RMSE 9,3. Математически это означает, что 70,6 % ($R^2$ = $0,84^2$ = 0,706) дисперсии экспертных оценок объясняется линейной комбинацией четырех библиометрических критериев. Оптимальные веса ($w_{1=0}$,38, $w_{2=0}$,42, $w_{3=0}$,12, $w_{4=0}$,08) близки к экспертным оценкам, но придают несколько большее значение индексу Хирша (0,42 против 0,35) за счет снижения весов страны и аффилиации.

Таблица 3.

Корреляция интегрального показателя достоверности с экспертными оценками при различных наборах весовых коэффициентов

| Набор весов | $w_1$ (Квартиль) | $w_2$ (h-индекс) | $w_3$ (Страна) | $w_4$ (Аффилиация) | Корреляция Пирсона | RMSE |
|---|---|---|---|---|---|---|
| Равные веса | 0,25 | 0,25 | 0,25 | 0,25 | 0,67 | 18,4 |
| Экспертная оценка | 0,40 | 0,35 | 0,15 | 0,10 | 0,79 | 12,7 |
| PCA-оптимизированные | 0,38 | 0,42 | 0,12 | 0,08 | 0,84 | 9,3 |
| Максимизация h-индекса | 0,10 | 0,70 | 0,10 | 0,10 | 0,71 | 15,2 |
| Максимизация квартиля | 0,70 | 0,10 | 0,10 | 0,10 | 0,74 | 14,1 |

*Примечание*: корреляция Пирсона рассчитана на тестовой выборке из 150 публикаций; RMSE (Root Mean Square Error) — среднеквадратичная ошибка предсказания экспертной оценки.

Попытки максимизировать вес одного критерия приводят к ухудшению общей точности. Набор с $w_{2=0}$,70 (максимизация h-индекса) дает корреляцию 0,71, а набор с $w_{1=0}$,70 (максимизация квартиля) — 0,74. Это подтверждает гипотезу о необходимости мультикритериального подхода: ни один показатель в отдельности не является достаточным для надежной оценки достоверности научной публикации.

Дисперсия оптимальных весов невелика: стандартное отклонение составляет 0,15, что свидетельствует об относительно сбалансированном вкладе различных критериев. Тем не менее, доминирование квартиля и h-индекса (совместный вес 0,80) подтверждает их ключевую роль в оценке качества научных работ, что согласуется с результатами предыдущих исследований [14, 15]. Практическая значимость разработанного показателя заключается в возможности автоматизированной фильтрации и ранжирования больших массивов научных публикаций. При пороговом значении D 0,65 достигается точность отбора качественных публикаций 88,3 % при полноте 84,7 % (по сравнению с ручной экспертной оценкой). Это позволяет сократить объем ручной верификации в 5–7 раз, сохраняя высокое качество итогового датасета.

## Заключение

Проведенное исследование демонстрирует комплексный подход к решению актуальной задачи построения моделей цифровых объектов из гетерогенных информационных ресурсов с количественно измеримым качеством. Разработанная методика, включающая этапы извлечения, обогащения и оценки достоверности данных, была верифицирована на корпусе из 2847 научных публикаций.

Ключевым результатом является установление количественных зависимостей эффективности методов извлечения от типа исходных документов. Показано, что для высокоструктурированных XML-документов rule-based подход обеспечивает точность 94,8 % при минимальных вычислительных затратах, в то время как для слабоструктурированных отсканированных PDF векторизованный подход с GPU-ускорением повышает F1-меру с 51 % до 86 % по сравнению с простыми правилами. Это подтверждает отсутствие универсального оптимального решения и необходимость адаптивного выбора метода в зависимости от характеристик данных.

Практическая значимость работы заключается в разработке интегрального показателя достоверности научных публикаций, основанного на оптимизированной комбинации библиометрических критериев. Достигнутая корреляция 0,84 с экспертными оценками при весовых коэффициентах $w_{1=0}$,38 (квартиль), $w_{2=0}$,42 (h-индекс), $w_{3=0}$,12 (страна), $w_{4=0}$,08 (аффилиация) позволяет автоматизировать предварительную фильтрацию данных с точностью 88,3%, сокращая объем ручной верификации в 5–7 раз. Ограничения исследования включают зависимость точности извлечения от качества исходных документов (для документов с битыми шрифтами или сильно искаженной структурой показатели снижаются на 15–20 %) и необходимость адаптации весовых коэффициентов показателя достоверности для различных предметных областей. В медицинских науках импакт-факторы журналов значительно выше, чем в математике, что требует нормализации по дисциплине. Дальнейшее развитие исследования связано с интеграцией больших языковых моделей (LLM) для повышения семантической точности извлечения отношений между сущностями, расширением типологии обрабатываемых документов (включая презентации, патенты, технические отчеты), а также разработкой адаптивных алгоритмов автоматического выбора оптимального метода извлечения на основе предварительного анализа структуры документа. Особый интерес представляет исследование применимости подхода к мультимодальным документам, содержащим комплексные визуализации данных и интерактивные элементы.

ЛИТЕРАТУРА


1. Visser M., van Eck N.J., Waltman L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic // Quantitative Science Studies. 2021. Vol. 2. No. 1. P. 20–41. DOI: 10.1162/qss_a_00018.
2. Сулейманов Р.Р., Бусыгина И.С. Методы и подходы к автоматической обработке текстовой информации при реализации систем извлечения данных // Российский технологический журнал. 2021. Т. 9. № 1. С. 7–17.
3. Барахнин В.Б., Кожемякина О.Ю., Мухамедиев Р.И., Борзилова Ю.С., Якунин К.О. Проектирование структуры программной системы обработки корпусов текстовых документов // Бизнес-информатика. 2019. № 4. С. 60–72. DOI: 10.17323/1998–0663.2019.4.60.72.
4. Zhang R., Meng Z., Wang H., Liu T., Wang G., Zheng L., Wang C. Hyperscale data analysis-oriented optimization mechanisms for higher education management systems platforms with evolutionary intelligence // Applied Soft Computing. 2024. Vol. 155. Article 111460. DOI: 10.1016/j.asoc.2024.111460
5. Жучкова С.В., Ротмистров А.Н. Автоматическое извлечение текстовых и числовых веб-данных для целей социальных наук // Социология: методология, методы, математическое моделирование. 2020. № 50–51. С. 141–183.
6. Lee S.J., Siau K. A review of data mining techniques // Industrial Management & Data Systems. 2001. Vol. 101. No. 1. P. 41–46. DOI: 10.1108/02635570110365989.
7. Inkina V.A., Antonov E.V., Artamonov A.A., Ionkina K.V., Tretyakov E.S., Cherkasskiy A.I. Multiagent information technologies in system analysis // Proceedings of the 27th International Symposium Nuclear Electronics and Computing (NEC'2019). Budva, Montenegro. 2019. P. 195–199. DOI: 10.1109/NEC.2019.00047.

8. Hirsch J.E. An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences. 2005. Vol. 102. No. 46. P. 16569–16572. DOI: 10.1073/pnas.0507655102.
9. Franceschini F., Maisano D., Mastrogiacomo L. The museum of errors/horrors in Scopus // Journal of Informetrics. 2016. Vol. 10. No. 1. P. 174–182. DOI: 10.1016/j.joi.2015.11.006.
10. Wu X., Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G.J., Ng A., Liu B., Yu P.S. Top 10 algorithms in data mining // Knowledge and Information Systems. 2008. Vol. 14. No. 1. P. 1–37. DOI: 10.1007/s10115-007-0114-2
11. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.
12. Egghe L. Theory and practise of the g-index // Scientometrics. 2006. Vol. 69. No. 1. P. 131–152. DOI: 10.1007/s11192-006-0144-7.
13. Banker K., Bakkum P., Verch S., Garrett D., Hawkins T. MongoDB in Action: Covers MongoDB version 3.0. Second Edition. Manning Publications, 2016. 480 p. ISBN: 978-1-617-29160-9.
14. Bornmann L., Mutz R., Hug S.E., Daniel H.-D. A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants // Journal of Informetrics. 2011. Vol. 5. No. 3. P. 346–359. DOI: 10.1016/j.joi.2011.01.006.
15. Waltman L., van Eck N.J. The inconsistency of the h-index // Journal of the American Society for Information Science and Technology. 2012. Vol. 63. No. 2. P. 406–415. DOI: 10.1002/asi.21678.