УДК 004.05+004.4

# МЕТОДИКА СРАВНЕНИЯ ПРОГРАММНЫХ РЕШЕНИЙ РАСПОЗНАВАНИЯ ТЕКСТОВ НАУЧНЫХ ПУБЛИКАЦИЙ ПО КАЧЕСТВУ ИЗВЛЕЧЕНИЯ МЕТАДАННЫХ

**И. И. Кузнецов[1] [0009-0001-6287-8295], О. П. Новиков[2] [0009-0009-3494-3799], Д. Ю. Ильин[3] [0000-0002-0241-2733]**

*[1, 2]Российский государственный университет им. А.Н. Косыгина (Технологии. Дизайн. Искусство), г. Москва, 119071, Россия*

*[3]МИРЭА – Российский технологический университет, г. Москва, 119454, Россия*

[1]iliya-kuznetsov@mail.ru, [2]novikovop55@rambler.ru, [3]i@dmitryilin.com

***Аннотация***

Метаданные научных публикаций используются для построения каталогов, определения цитируемости публикаций и решения других задач. Автоматизация извлечения метаданных из PDF-файлов позволяет ускорить выполнение обозначенных задач, а от качества извлеченных данных зависит возможность их дальнейшего использования. Проанализированы существующие программные решения, в итоге отобраны три: GROBID, CERMINE, ScientificPdfParser. Предложена методика сравнения этих программных решений распознавания текстов научных публикаций по качеству извлечения метаданных. На основе методики проведен эксперимент по извлечению четырех типов метаданных (название, аннотация, дата публикации, имена авторов). Для сравнения программных решений использован набор из 112457 публикаций с разбиением на 23 предметные области, сформированный на основе данных Semantic Scholar. Приведен пример выбора эффективного программного решения извлечения метаданных в условиях заданных приоритетов для предметных областей и типов метаданных с использованием взвешенной суммы. Определено, что для приведенного примера CERMINE показывает эффективность на 10,5% выше, чем GROBID, и на 9,6% выше, чем ScientificPdfParser.

***Ключевые слова****: распознавание текста, научные публикации, метаданные, качество извлечения данных, методика.*

## ВВЕДЕНИЕ

Научные статьи являются одним из основных способов публикации и предоставления результатов разнообразных научных исследований. Помимо непосредственно информации об исследованиях, содержащейся в текстах статей, интерес представляют и относящиеся к этим статьям метаданные. К таким метаданным относятся названия статей, сведения об авторах, дате и месте публикации, используемом цитировании и др. Сбор и анализ подобных метаданных может служить различным целям – рассмотрению связей статей на уровне цитирования [1], анализу научной области на предмет того, насколько активно ведется публикация связанных с ней работ в какой-то период времени и какие направления исследований в ней преобладают [2], формирования выборки статей по авторам и другим признакам, и многое другое.

Сбор метаданных, а также какого-либо контента статей вручную является трудоемким. Электронные библиотеки, содержащие большие объемы научных статей, помимо текстов статей могут предоставлять и связанные с ними метаданные, однако на практике далеко не все источники статей содержат их метаданные или же предоставляют лишь часть из возможных данных. Кроме того, такие метаданные могут быть низкого качества: содержать ошибки, неточности, пустые поля и т. д. [3]. Таким образом, актуальным становится извлечение метаданных и контента из статей в автоматизированном режиме, и для этого существует ряд программных решений.

Исследования, связанные с автоматизированным извлечением метаданных и конкретных фрагментов статей, ведутся в различных направлениях. В ряде работ, например [4–6], предложены оригинальные модели и программные решения для извлечения метаданных. При этом одни решения специализируются на извлечении библиографических ссылок, такие как модель полного цикла извлечения и постобработки [7, 8], другие – на извлечении непосредственно метаданных статьи [9]. Некоторые работы посвящены извлечению формул и результатов измерений из научных статей: в [10] описана такая система, разработанная на основании GROBID, в [11] – система, использующаяся для извлечения из текстов статей информации о различных материалах (названий, формул, свойств).

Аналогичные разработки имеются для извлечения таких данных из научных публикаций в области сверхпроводников [12] или, например, в геологии [13].

Программные решения, изначально созданные для работы с научными статьями, находят применение и при извлечении данных из других документов сходной структуры: так, в [14] описано использование инструмента на основе GROBID для извлечения результатов измерений из медицинских отчетов. Помимо разработки новых моделей для извлечения метаданных или же извлечения таких фрагментов, как результаты измерений, некоторые исследователи предлагают методы использования результатов измерений для оценки качества самих статей или их оформления. В [15] извлечение названия, ключевых слов и полного набора библиографических ссылок использовано для проверки того, насколько содержание статьи и тематики цитируемых работ соответствуют заявленной тематике исходной статьи. В работе [16] извлечение ключевых слов и фрагментов статьи из документа использовано для оценки уровня репрезентативности аннотации относительно содержимого статьи, а в [17] распознавание структуры и извлечение элементов текста применены для создания аннотаций и кратких описаний статей, в которых представлены основные тезисы исследования.

Существующие решения, как правило, работают с одним языком, например английским, ввиду чего еще одним направлением является разработка решений для поддержки мультиязычных наборов статей. Например, в [18] была предложена модель извлечения библиографических ссылок, поддерживающая многие языки (включая корейский), а в работах [19] и [20] авторы сосредоточились на создании размеченных наборов данных для поддержки статей, написанных на кириллице.

Поскольку существуют готовые программные решения для извлечения метаданных, ссылок и контента статей, проводятся и исследования (например, [21, 22]) по сравнению их эффективности и качества извлечения. В [21] авторы провели сравнение трех инструментов для извлечения библиографических ссылок на основе статей из индонезийской базы научных журналов и предположили использовать эти результаты при создании приложения для корректного распознавания и извлечения библиографических ссылок. В работе [22] авторами пред-

ложена инфраструктура для измерения качества и производительности извлечения метаданных при сравнении качества извлечения несколькими различными программными решениями.

При извлечении метаданных могут быть допущены различные ошибки: неверное распознавание элементов текста и ошибки разметки, ошибки при извлечении конкретных символов или их неверная интерпретация, некорректное распознание и извлечение строки текста и др. Большое количество ошибок будет означать низкое качество извлечения метаданных, а полученные данные будет затруднительно или невозможно использовать для последующих анализа и обработки. Поскольку извлечение метаданных может осуществляться на начальных этапах в цепочке анализа данных, от качества извлечения будет зависеть корректность результатов всей последующей части цепочки.

Целью настоящей работы является разработка методики сравнения программных решений распознавания текстов научных публикаций по качеству извлечения метаданных. Отличиями от ранее представленных работ являются используемая процедура определения пороговых значений метрик схожести строк и учет влияния принадлежности публикаций к различным предметным областям на качество извлечения метаданных. При этом под качеством понимаются точность и полнота извлечения метаданных. Точность отражает верное распознавание всех символов, а также отсутствие ошибок и искажений при извлечении строк и фрагментов, содержащих метаданные. Полнота же является показателем того, что все метаданные указанного типа будут извлечены из документа (например, будут извлечены все авторы публикации).

## 1. АНАЛИЗ СУЩЕСТВУЮЩИХ ПРОГРАММНЫХ РЕШЕНИЙ

Существует ряд программных решений, позволяющих извлекать метаданные из научных статей в формате PDF. Они различаются по набору извлекаемых метаданных, уровню поддержки, затрачиваемым ресурсам и т. д. При этом алгоритм их работы состоит, как правило, из набора этапов, схожих по принципу организации. На первом этапе происходят распознание символов и реконструкция базовой структуры документа: каждой странице ставится в соответствие набор текстовых блоков. На втором этапе текстовые блоки классифицируются, происходит определение принадлежности блоков тому или иному типу элементов статьи.

На третьем этапе из распознанных элементов извлекаются соответствующие им метаданные (а также текстовый контент, изображения, таблицы и т. д. в зависимости от возможностей программы). В совокупности, на втором и третьем этапах для классификации блоков и распознавания и извлечения метаданных и контента используются различные методы: подходы на основе правил [23], эвристических предположений [24], различных алгоритмов машинного обучения [24–26] или же комбинации этих методов [27].

Рассмотрены существующие решения для извлечения метаданных и контента из научных статей, а именно: CERMINE [28], GROBID [29], ParsCit [30], Neural-ParsCit [31], PDFX [32], ScientificPdfParser, Science Parse и SPv2. При распознавании и извлечении метаданных в этих решениях используются различные методы. В CERMINE – это метод опорных векторов и алгоритм кластеризации k-средних для извлечения ссылок. В GROBID алгоритм основан на методе условных случайных полей. Этот же метод используется в ParsCit в сочетании с эвристическими методами на основе регулярных выражений. Решение Neural-ParsCit, являющееся усовершенствованной версией ParsCit, использует модель глубокого обучения LSTM. ScientificPdfParser имеет три режима извлечения: на основании заранее определенных правил и эвристики, с использованием машинного обучения на основе наивного байесовского алгоритма, а также гибридный режим с генерацией обучающей модели. Science Parse и SPv2 тоже используют методы машинного обучения, включая метод условных случайных полей. В табл. 1 приведены некоторые характеристики перечисленных решений, которые рассматривались при их отборе для эксперимента.

Таблица 1. Характеристики программных решений извлечения метаданных научных статей

| Признак | CERMINE | GROBID | ParsCit/Neural-ParsCit | PDFX | ScientificPdfParser | Science Parse | SPv2 |
|---|---|---|---|---|---|---|---|
| Язык программы | Java | Java | Perl | Java | Java | Java + Scala | Java + Scala |
| Открытый исходный | Да | Да | Да | Нет | Да | Да | Да |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| код | | | | | | | |
| Извлечение метаданных | Да | Да | Нет | Да | Да | Да | Да |
| Извлечение ссылок | Да | Да | Да | Да | Да | Да | Да |
| Извлечение метаданных для ссылок | Да | Да | Да | Нет | Да | Да | Нет |
| Количество типов метаданных | 11 | 7 | - | 4 | 5 | 4 | 3 |
| Тип выходного файла | XML | JSON | XML | XML | JSON | XML | XML |

После рассмотрения этих решений для эксперимента были отобраны несколько из них. Условиями для отбора служили: наличие открытого исходного кода и возможность извлекать метаданные статьи (не только библиографические ссылки). Согласно этим условиям были исключены PDFX как не обладающий открытым исходным кодом и обе версии ParsCit, поскольку они извлекают только библиографические данные. Кроме того, из рассмотрения были исключены Science Parse и SPv2, так как из-за долгого отсутствия поддержки авторами кодовой базы не удалось достичь их корректной и стабильной работы без дополнительной актуализации и доработки кода. Таким образом, для эксперимента были взяты три решения: CERMINE, GROBID и ScientificPdfParser.

## 2. МЕТОДИКА СРАВНЕНИЯ ПРОГРАММНЫХ РЕШЕНИЙ ПО КАЧЕСТВУ ИЗВЛЕЧЕНИЯ МЕТАДАННЫХ

Для эксперимента должны использоваться научные статьи в формате PDF, собранные на основании какой-либо электронной библиотеки таких статей. Эта же библиотека должна предоставлять метаданные для выбранных статей, что послужит контрольным набором для сравнения с результатами извлечения. Кроме того, должны быть отобраны статьи из различных научных направлений, что позволит получить как общие оценки качества извлечения, так и оценки для различ-

ных направлений, поскольку принципы оформления и структуры статей могут отличаться от направления к направлению. Для проведения эксперимента по сравнению программных решений по качеству извлечения метаданных из научных статей была разработана методика, содержащая следующие шаги.

1. Выбор подходящей электронной библиотеки, содержащей метаданные для статей различных научных направлений и PDF-файлы с текстами научных публикаций либо ссылки на них.
2. Выгрузка из библиотеки набора метаданных, относящихся к статьям различных направлений.
3. Очистка метаданных от некорректных и неполных образцов.
4. Выгрузка набора статей в формате PDF, соответствующих набору метаданных.
5. Очистка набора статей от невалидных и пустых документов.
6. Извлечение метаданных одинаковых типов с помощью отобранных программных решений.
7. Постобработка результатов извлечения и приведение их к единому формату.
8. Сравнение результатов извлечения с выгруженным ранее набором метаданных, вычисление значений метрик качества извлечения.
9. Анализ полученных значений метрик, сравнение программных решений между собой на основании результатов эксперимента.

При сравнении качества извлечения должны быть использованы заранее извлеченные метаданные, позволяющие определить, какой результат работы программных решений может считаться правильным. На основании полученных значений метрик для всех типов метаданных должны быть рассчитаны точность, полнота и $F_1$-мера, которые и послужат итоговой оценкой качества извлечения и основанием для выводов о сравнительных возможностях исследуемых программных решений.

## 3. ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ПРОГРАММНЫХ РЕШЕНИЙ

На первом этапе эксперимента была осуществлена загрузка документов статей и соответствующих им метаданных. В качестве источника метаданных был

выбран Semantic Scholar. Этот интернет-ресурс предоставляет широкие возможности для выгрузки метаданных научных статей по различным направлениям и гибкие возможности настройки выгружаемых метаданных. Среди прочего ресурс предоставляет ссылку на исходный документ статьи в формате PDF. При настройке выгрузки есть возможность выбрать 23 различных предметных области для статей. При первоначальной выгрузке метаданных с помощью предлагаемого этим ресурсом API было выгружено 230 000 наборов метаданных, по 10 000 для каждой из предложенных областей, затем из этих наборов были исключены те, в которых присутствовали пустые значения для рассматриваемых типов метаданных. Далее для оставшихся наборов с помощью содержащихся в них ссылок была осуществлена загрузка полных текстов статей в формате PDF. После выгрузки из выборки были исключены поврежденные или пустые PDF-файлы. Итоговая выборка наборов метаданных с соответствующим ими PDF-документами составила 112 457 записей, т. е. около 5000 записей на каждую из рассматриваемых предметных областей.

И наконец, с помощью CERMINE, GROBID и ScientificPdfParser из текстов статей были извлечены метаданные. Для GROBID использовался интерфейс SciPDF_Parser, совместимый с Python и предоставляющий настройки для извлекаемых полей. Для всех статей извлекались следующие метаданные: название статьи, фамилии авторов, год публикации и аннотация. Такой набор, с одной стороны, опирается на возможности рассматриваемых решений, с другой – позволяет делать выводы об извлечении отличающихся данных: текст из нескольких предложений в аннотации, набор из имен авторов, численные значения даты. Для GROBID и ScientificPdfParser данные были извлечены в формате JSON, для CERMINE – в формате cermxml, а затем также переведены в формат JSON. Это было обусловлено тем, что наборы метаданных, полученных от Semantic Scholar, также представлены в формате JSON.

Для оценки правильности извлечения метаданных использовались две метрики сходства. Для длинных строк, таких как название статьи и аннотация, использовалась метрика сходства на основе расстояния Левенштейна [10]. Для имен авторов и даты публикации использовалась метрика сходства на основе расстояния Джаро–Винклера, как более подходящая для коротких строк с учетом сравнения совпадающего префикса, в том числе для сравнения имен [11].

Метрика сходства на основе расстояния Левенштейна для строк $s_1$ и $s_2$ определяется по формуле

$$\mathrm{sim}_{\mathrm{Lev}}(s_1, s_2) = 1 - \frac{\mathrm{Lev}(s_1, s_2)}{\max(|s_1|, |s_2|)}$$

где $\mathrm{Lev}(s_1, s_2)$ – расстояние Левенштейна для строк $s_1$ и $s_2$, а $|s_1|$ и $|s_2|$ – длины строк.

Метрика сходства на основе расстояния Джаро–Винклера для двух строк $s_1$ и $s_2$ определяется по формуле

$$\mathrm{sim}_{\mathrm{JW}}(s_1, s_2) = 1 - (d_{\mathrm{j}}(s_1, s_2) + (l \cdot p \cdot (1 - d_{\mathrm{j}}(s_1, s_2)))),$$

где $d_{\mathrm{j}}(s_1, s_2)$ – расстояние Джаро для строк $s_1$ и $s_2$, $l$ – длина совпадающего префикса (максимум до 4 символов), $p$ – коэффициент масштабирования. Стандартным значением для $p$ является 0.1, оно и было взято для вычислений в рамках этого исследования.

Для определения качества извлечения метаданных необходимо было определить пороговое значение для оценок, получаемых с помощью метрик сходства, после которого извлечение считалось бы корректным. Это обусловлено тем, что существует некоторое отклонение от точного совпадения, обусловленное незначительными неточностями (отсутствием пробелов, нестандартными текстовыми символами, знаками препинания), в рамках которого данные все еще можно считать корректными для дальнейшего использования. Правильное выявление этого порога является важной задачей, поскольку именно на нем строится итоговая оценка качества извлечения метаданных и, как следствие, работы всего программного решения. Алгоритм выявления порогового значения для обоих метрик сходств был одинаков и состоял из следующих шагов.

1. Для поиска порогового значения $t$ берется интервал [0.5,1], где 1 означает полное совпадение. Интервал разбивается с шагом 0.05, таким образом рассматривается $t \in \{0.05x \mid x \in \{10,11,\ldots,20\}\}$.
2. Для каждого возможного порогового значения $t$ из общего числа статей выбираются те, у которых вычисленное значение метрики сходства попадает в интервал $[t; t + 0.01]$. Отбор производится таким образом, чтобы в вы-

борке были равным образом представлены статьи из всех предметных областей, извлеченные всеми рассматриваемыми программными решениями.

3. Извлеченные метаданные, соответствующие каждому набору статей, рассматриваются на предмет количества статей с некорректно извлеченными метаданными $v_e$ и его соотношения с общим количеством статей $v_a$.
4. Итоговым пороговым значением $t$ выбирается то, при котором показатель $v = \frac{v_a - v_e}{v_a}$ начинал превышать 0.95.

Для выявления порога $t_{Lev}$ по метрике сходства на основе расстояния Левенштейна использовалось название статьи, для порога $t_{JW}$ по метрике сходства на основе расстояния Джаро-Винклера – дата публикации.

На рис. 1 приведен график, показывающий зависимость $v$ от пороговых значений $t_{Lev}$ и $t_{JW}$.

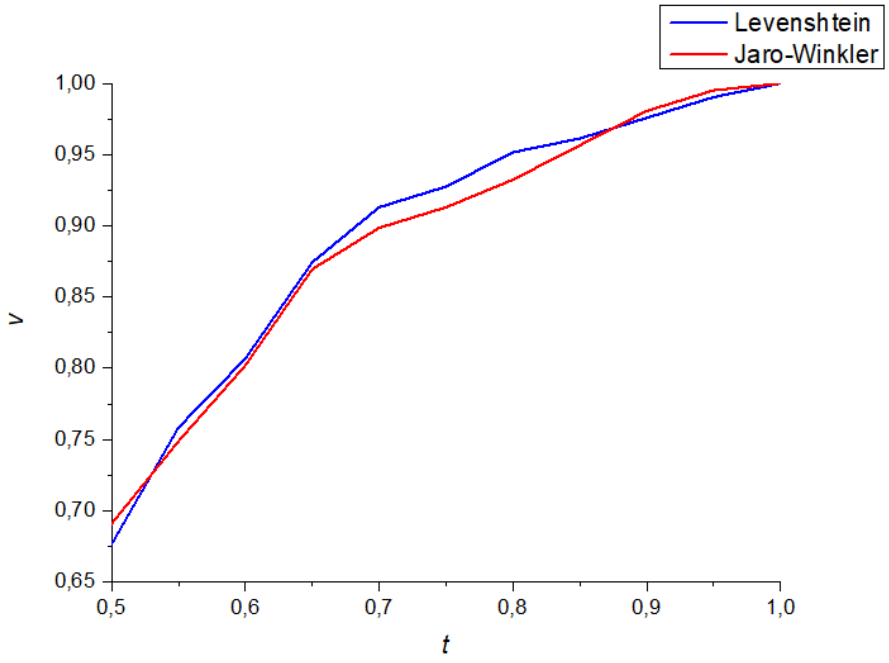


Рис. 1. Зависимость $v$ от пороговых значений $t_{Lev}$ для метрики сходства на основе расстояния Левенштейна и $t_{JW}$ для метрики сходства на основе расстояния Джаро–Винклера

Таким образом, по результатам определения порога были определены $t_{Lev} = 0.8$ и $t_{JW} = 0.85$. Метаданные, для которых значение указанной метрики

сходства превышает пороговое значение, считались корректно извлеченными при вычислении значений метрик качества извлечения.

В качестве метрик качества извлечения использовались полнота (recall), точность (precision) и $F_1$-мера. Для вычисления точности и полноты было использовано разбиение результатов с помощью матрицы спутанности, согласно которой результаты определения принадлежности к классу (в рассматриваемом случае результаты извлечения метаданных) делятся на положительные TP, ложноположительные FP, отрицательные TN и ложноотрицтальные FN. На основе этого полноту и точность можно выразить так:

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}},\ \text{precision} = \frac{\text{TP}}{\text{TP+FN}}.$$

Следует пояснить, что в рассмотренном случае представляют собой результаты по матрице ошибок. TP – это количество статей с корректно извлеченными метаданными, то есть такими, для которых значение метрики сходства превысило вычисленный порог $t_{\text{Lev}}$ или $t_{\text{JW}}$ соответственно. TP + FP – это количество всех статей, для которых результатом извлечения является непустой ответ, независимо от корректности. TP + FN – это количество статей, у которых образцовые данные являются непустыми, то есть общее количество исследуемых статей. F1-мера представляет собой среднее гармоническое для точности и полноты и определяется по формуле

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Определение значений по рассмотренным метрикам качества извлечения проводилось по отдельности для каждой представленной предметной области и каждого выбранного типа метаданных.

**4. РЕЗУЛЬТАТЫ**

На рис. 2–5 представлены рассчитанные на основании результатов эксперимента значения $F_1$-меры для четырех рассматриваемых типов метаданных в применении к 23 предметным областям для трех исследуемых решений.

Для названия статьи (рис. 2) CERMINE и GROBID показывают схожие результаты, без явного превосходства одной системы над другой, в то время как ScientificPdfParser несколько отстает от этих систем.

Для аннотаций (рис. 3) GROBID показывает существенно лучший результат, чем две другие системы. При этом GROBID показывает худший результат для даты публикации статьи (рис. 4).

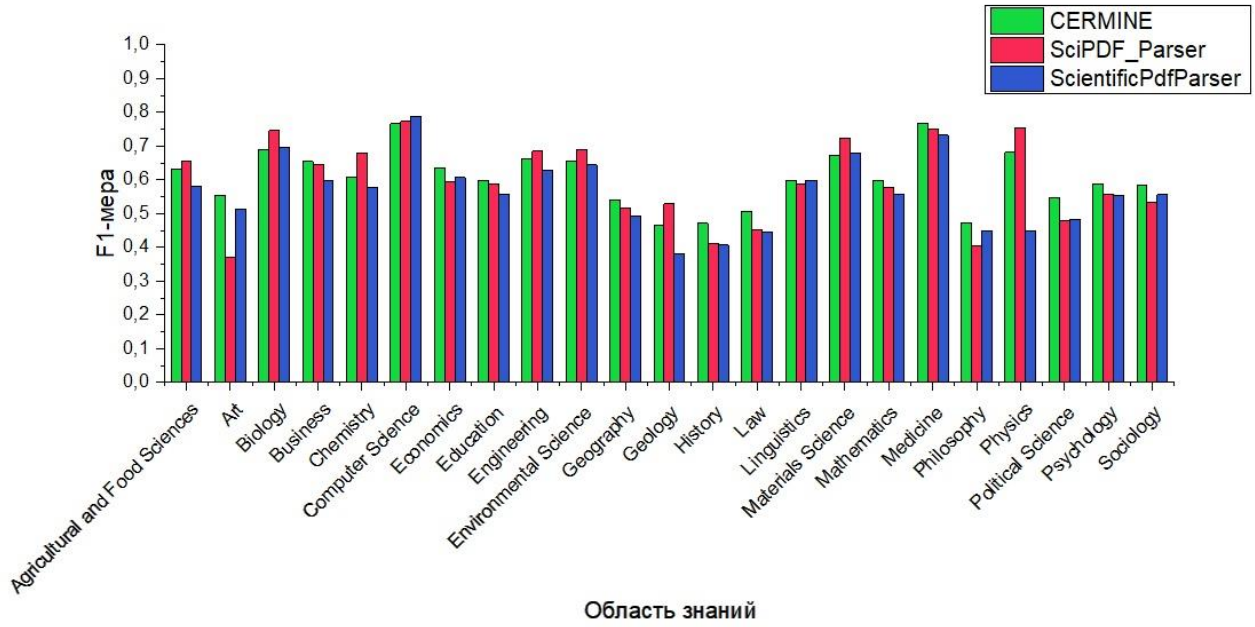


Рис. 2. Значения $F_1$-меры для извлечения названия статьи

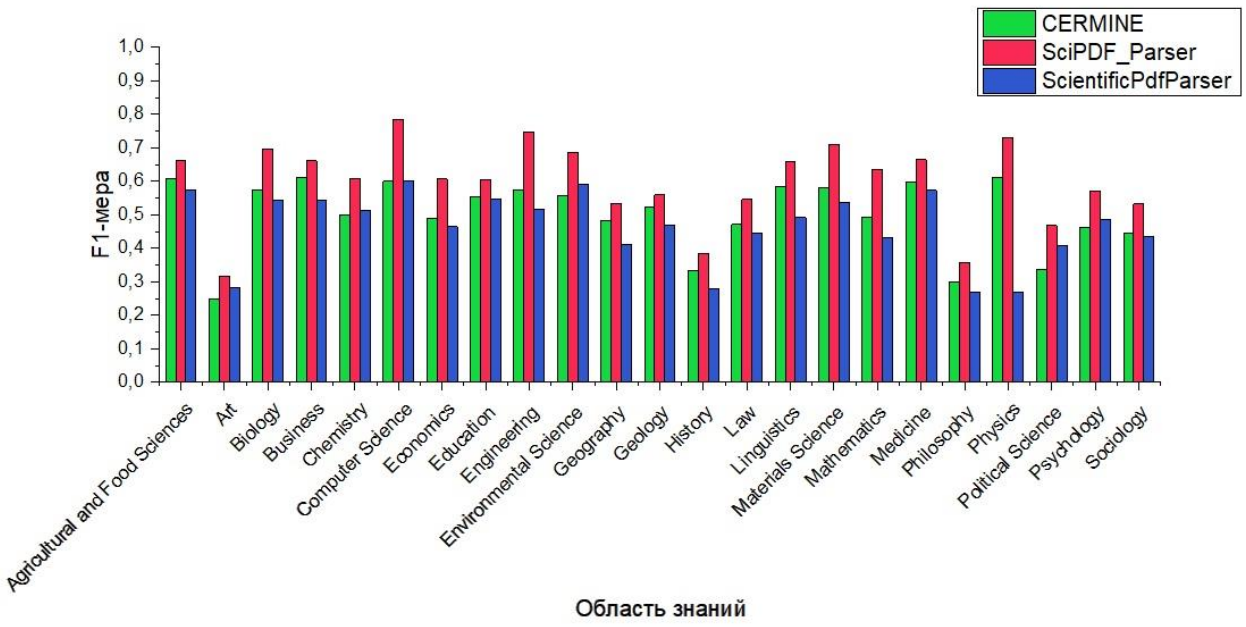


Рис. 3. Значения $F_1$-меры для извлечения аннотации статьи

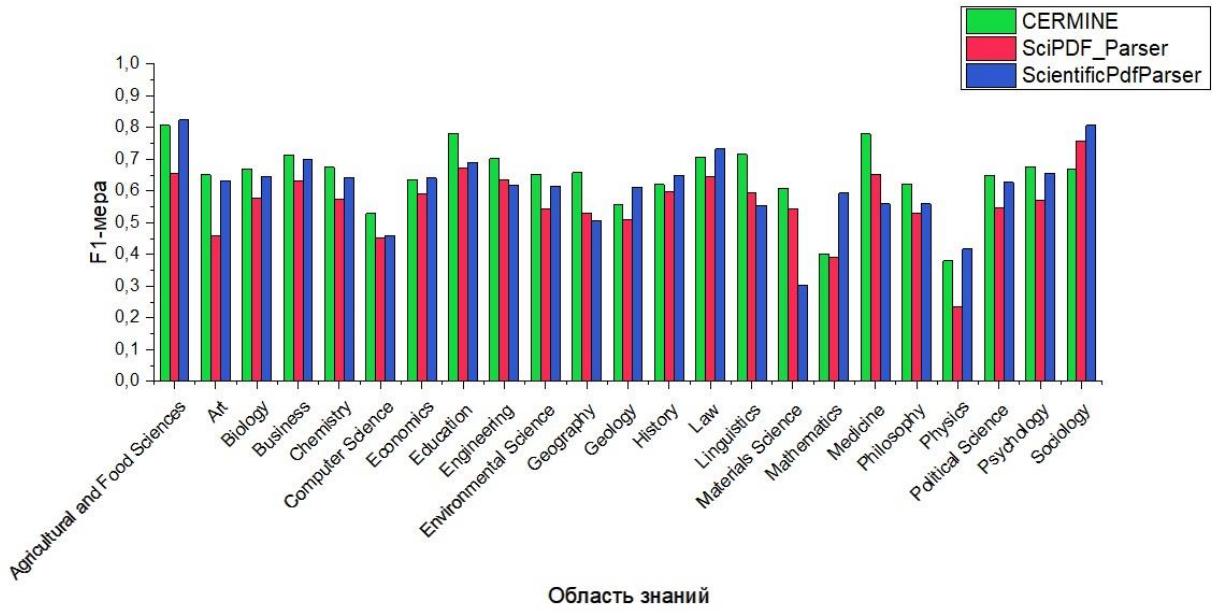


Рис. 4. Значения $F_1$-меры для извлечения даты публикации статьи

При извлечении имен авторов GROBID также показывает самый низкий результат, однако в этом случае результаты всех трех систем оказались относительно невысокими (рис. 5).

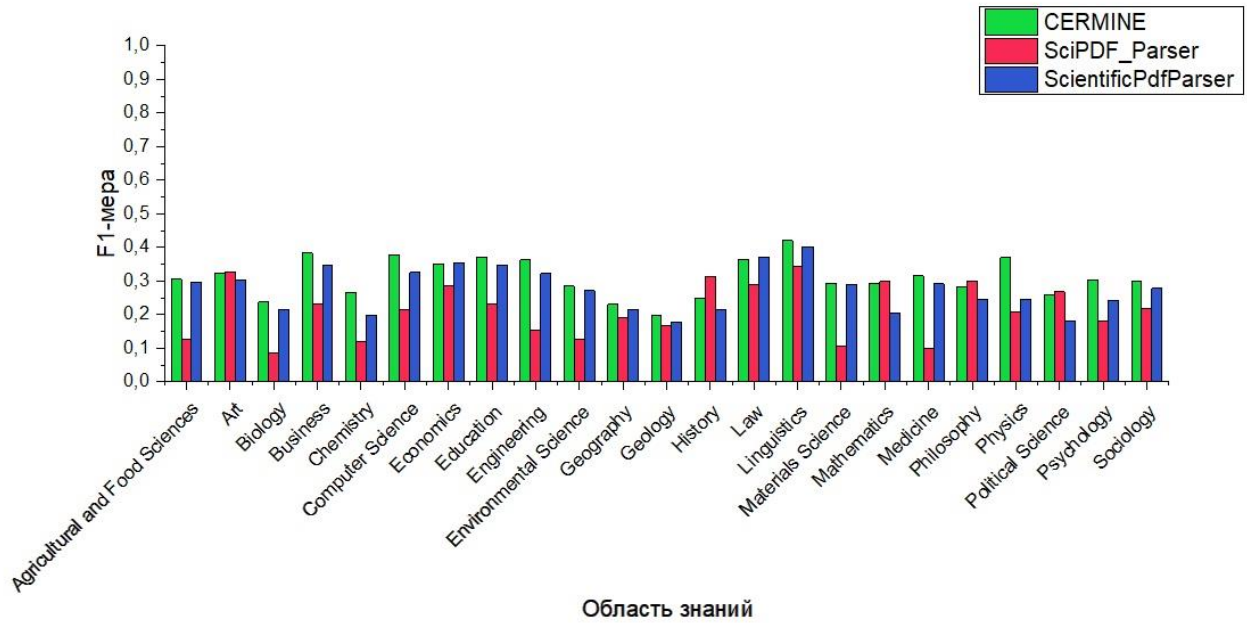


Рис. 5. Значения $F_1$-меры для извлечения имен авторов статьи

Одновременно с извлечением имен авторов была произведена оценка корректности их распознавания. Рассмотрение этого фактора обусловлено тем, что

в ряде случаев системам удается выделить не всех авторов из указанных в статье, или же, наоборот, система дает ряд ложноположительных распознаваний на основании неверного разбиения, определив как отдельных авторов инициалы или же указанные звания и аффилиации авторов. Результаты этой оценки, где метрики основаны на точном совпадении количества авторов, приведены на рис. 6.

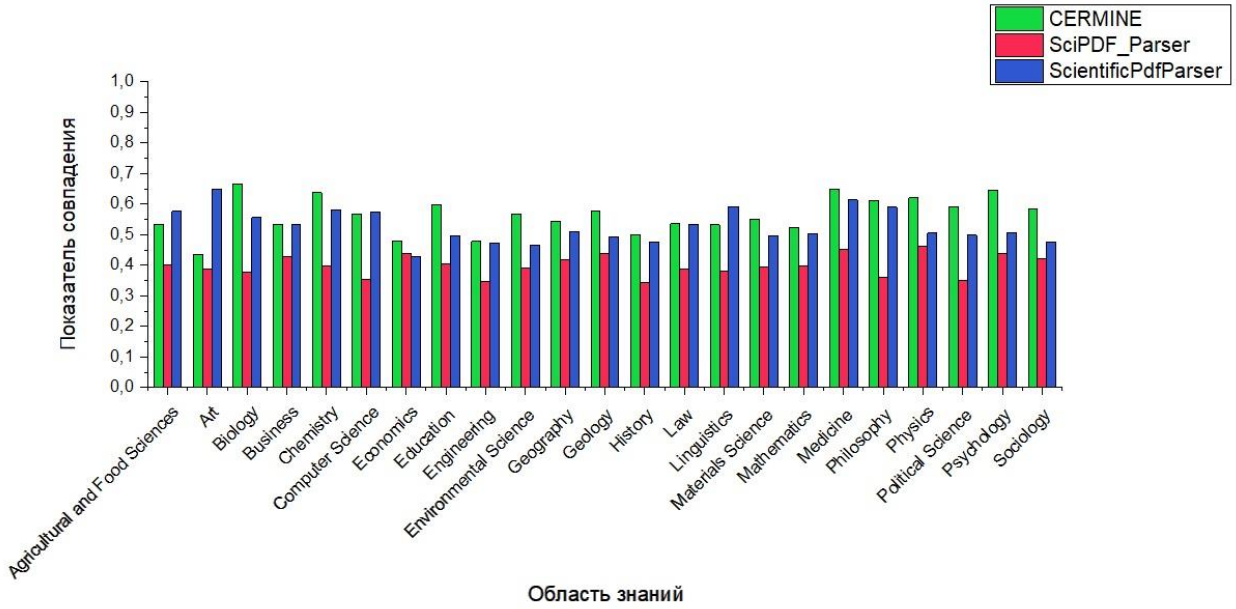


Рис. 6. Значения $F_1$-меры для корректности распознавания количества авторов

Все три решения показали относительно невысокий результат для правильного извлечения количества авторов. Это позволяет предполагать, что низкие показатели качества извлечения имен авторов обусловлены не только ошибками в извлечении настоящих имен как строк, но и большим количеством ошибок в распознавании элементов текста как настоящих имен (ложные распознавания и т. д.). Это говорит о необходимости дополнительной настройки систем в области извлечения имен авторов или же дополнительной постобработки извлеченных имен с целью удаления элементов, некорректно включенных в имена. Следует отметить, что эта связь является предполагаемой, поскольку в рамках настоящего исследования не проводились исследования, которые экспериментально выявили бы степень влияния некорректно выделенных и разделенных имен на общее число ошибок при извлечении имен авторов.

Средние значения $F_1$-меры для различных метаданных для всех предметных областей были рассчитаны и приведены в табл. 2.

Таблица 2. Среднее значение $F_1$-меры для всех рассмотренных предметных областей

| Тип метаданных | CERMINE | GROBID | ScientificPdfParser |
|---|---|---|---|
| Название | 0.6061 | 0.5954 | 0.5641 |
| Аннотация | 0.5011 | 0.5965 | 0.4644 |
| Дата публикации | 0.6461 | 0.5603 | 0.6278 |
| Имена авторов | 0.3102 | 0.212 | 0.2754 |

## 5. ВЫБОР ПРОГРАММНОГО РЕШЕНИЯ ИЗВЛЕЧЕНИЯ МЕТАДАННЫХ НА ОСНОВЕ ЭКСПЕРИМЕНТАЛЬНОЙ ОЦЕНКИ

Для определенных областей полученные оценки качества всех трех решений в среднем могут быть ниже, чем для других областей. Так, программные решения показывают наиболее низкие результаты для извлечения названия и аннотации в областях Art, History и Philosophy. Для Biology, Geology и Geography они показывают относительно низкие результаты для всех типов метаданных, в то время как для Economic, Education, Engineering, Law и Linguistics показывают стабильно более высокие результаты. Для Physics программные решения показывают высокие результаты для всех типов метаданных, кроме даты публикации статьи. Это позволяет говорить о том, что различие в результатах может быть не случайным, а обусловлено различными структурой и оформлением статей в различных предметных областях. В связи с этим при работе с конкретной предметной областью целесообразно ориентироваться на качество работы программного решения для нее, а не на средние показатели качества. Выбор программного решения, показывающего наиболее высокое качество извлечения, играет важную роль для решения любой конкретной задачи, одним из этапов которой является извлечение метаданных.

Для выбора программного решения в конкретных условиях предлагается использовать интегральную оценку качества на основе взвешенной суммы. Пусть дано $n$ наборов статей $A_i \subseteq X$, $i \in [1,n]$, полученных для соответствующих предметных областей, где $X$ – множество всех извлеченных статей, и $m$ типов метаданных $d_j \in D, j \in [1,m]$, где $D$ – множество типов извлекаемых метаданных.

Для каждого типа метаданных зададим метрику сходства и порог $t_j \in \{t_{\text{Lev}}, t_{\text{JW}}\}$, $j \in [1, m]$, превышение которого будет означать корректное распознавание. Пусть также даны веса $w_i$, отражающие значимость предметных областей, и веса $u_j$, отражающие значимость извлекаемых метаданных, такие, что $\sum_{i=1}^{n} w_i = 1$ и $\sum_{j=1}^{m} u_j = 1$.

Пусть дано *o* программных решений $s_k \in S, k \in [1, o]$, где $S$ – множество рассматриваемых программных решений. Тогда эффективным будет программное решение $s^* \in S$, которое позволит максимизировать интегральную оценку качества:

$$s^* = \arg\max_{s_k \in S} \left( \sum_{j=1}^{m} \left( u_j \sum_{i=1}^{n} w_i f\left(s_k, A_i, d_j, \text{sim}_j, t_j, g\right) \right) \right),$$

где $f\left(s_k, A_i, d_j, \text{sim}_j, t_j, g\right)$ – функция, используемая для вычисления значения метрики качества извлечения $g$ (в данном случае, $F_1$-меры) для конкретных программного решения $s_k$, набора статей $A_i$ и типа метаданных $d_j$ при используемой метрике сходства $\text{sim}_j$ и заданном для нее пороговом значении $t_j$.

Рассмотрим пример определения эффективного программного решения с использованием интегральных оценок. Дано $n = 23$ набора статей $A_i, i \in [1, n]$, соответствующих рассмотренным ранее предметным областям, $m = 4$ рассматриваемых типа метаданных $d_j, j \in [1, m]$ (название $d_1$, аннотация $d_2$, дата публикации $d_3$, имена авторов $d_4$) и $o = 3$ программных решения $s_k, k \in [1, o]$ для оценки (CERMINE, GROBID, ScientificPdfParser).

Для каждого типа метаданных установим метрики сходства и вычисленные ранее пороговые значения. Для названия статьи и аннотации $\text{sim}_1 = \text{sim}_2 = \text{sim}_{\text{Lev}}$, $t_1 = t_2 = t_{\text{Lev}} = 0.8$; для даты публикации и имен авторов $\text{sim}_3 = \text{sim}_4 = \text{sim}_{\text{JW}}$, $t_3 = t_4 = t_{\text{JW}} = 0.85$.

Пусть веса для предметных областей распределяются следующим образом: Engineering, Mathematics, Physics $w_1 = w_2 = w_3 = 0.1$; Materials Science, Geology,

Chemistry, Medicine $w_4 = w_5 = w_6 = w_7 = 0.07$; Computer Science, Environmental Science $w_8 = w_9 = 0.05$; Biology $w_{10} = 0.04$; Agricultural and Food Science, Economics, Geography, Business, Education $w_{11} = w_{12} = w_{13} = w_{14} = w_{15} = 0.03$; Sociology, Psychology, Political Science, Law, Linguistics $w_{16} = w_{17} = w_{18} = w_{19} = w_{20} = 0.02$; Art, History, Philosophy $w_{21} = w_{22} = w_{23} = 0.01$. С использованием этих весов получим промежуточные оценки для четырех рассматриваемых типов метаданных (табл. 3) и программных решений.

Таблица 3. Промежуточные оценки качества программных решений для четырех типов метаданных

| Тип метаданных | CERMINE | GROBID | ScientificPdfParser |
|---|---|---|---|
| Название | 0.6318 | 0.6462 | 0.5801 |
| Аннотация | 0.5396 | 0.6486 | 0.4793 |
| Дата публикации | 0.6127 | 0.5286 | 0.5772 |
| Имена авторов | 0.3106 | 0.1879 | 0.2665 |

Пусть для типов метаданных веса распределяются следующим образом: имена авторов $u_1 = 0.4$; название $u_2 = 0.3$; дата публикации $u_3 = 0.2$; аннотация $u_4 = 0.1$. Тогда для итоговой оценки качества работы программных решений получим следующие значения: CERMINE – 0.4912; GROBID – 0.4396; ScientificPdfParser – 0.444, и соответственно, рейтинг программ выстраивается следующим образом:

$$\text{CERMINE} \to \text{ScientificPdfParser} \to \text{GROBID}.$$

CERMINE по интегральной оценке качества показывает эффективность выше, чем GROBID, на 10.5%, и выше, чем ScientificPdfParser, на 9.6%. Таким образом, получен рейтинг программных решений с учетом выбранных весов для предметных областей и типов метаданных для представленного примера.

**ЗАКЛЮЧЕНИЕ**

Представлена методика сравнения программных решений распознавания текстов научных публикаций в формате PDF по качеству извлечения метаданных. Для извлечения рассмотрены четыре типа метаданных: название, аннотация,

дата публикации, имена авторов. Проанализированы три программных решения для извлечения метаданных: CERMINE, GROBID и ScientificPdfParser.

Проведен эксперимент по сравнению решений на основе подготовленного набора данных и выбора программных решений в условиях заданных приоритетов для предметных областей и типов метаданных. В ходе эксперимента были установлены и сконфигурированы программные решения, примененные для извлечения метаданных. Получена оценка качества извлечения на основании сравнения результатов с выгрузкой метаданных из сервиса Semantic Scholar. При извлечении общий набор статей был разделен на 23 предметных области, что позволило в дальнейшем сравнить качество извлечения для этих областей и сделать предположение о наличии связи качества извлечения от специфики оформления статей, присущих той или иной предметной области.

При выборе программных решений предложено использовать интегральную оценку на основе взвешенной суммы. Рассмотрен пример присвоения приоритетов предметным областям и типам метаданных, получен рейтинг программных решений в рамках заданных условий.

Разработанная методика имеет прикладное значение. Предложенную методику и полученные данные о качестве извлечения для различных программных средств можно использовать для выбора программного решения в условиях решения конкретной задачи.

Дальнейшие исследования могут быть посвящены особенностям функционирования программных решений, показателям их временной эффективности и вычислительной ресурсоемкости. Можно рассмотреть также построение ансамблей из программных решений извлечения метаданных с целью повышения качества извлечения.

**СПИСОК ЛИТЕРАТУРЫ**

1. *Qayyum F., Afzal M.T.* Identification of important citations by exploiting research articles' metadata and cue-terms from content // Scientometrics. 2019. Vol. 118. P. 21–43.

2. *Liu X., Zhang J., Guo C.* Full-text citation analysis: A new method to enhance scholarly networks // Journal of the American Society for Information Science and Technology. 2013. Vol. 64. No. 9. P. 1852–1863.

3. *Saier T., Färber M.* unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata // Scientometrics. 2020. Vol. 125. No. 3. P. 3085–3108.

4. *Safder I. et al.* Deep learning-based extraction of algorithmic metadata in full-text scholarly documents // Information processing and management. 2020. Vol. 57. No. 6. Article no. 102269.

5. *O'Leary N.A. et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets // Scientific data. 2024. Vol. 11. No. 1. Article no. 732.

6. *Safder I., Hassan S.U.* Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications // Scientometrics. 2019. Vol. 119. P. 257–277.

7. *Joshi B., Symeonidou A., Danish S.M., Hermsen F.* An End-to-End Pipeline for Bibliography Extraction from Scientific Articles // Proceedings of the Second Workshop on Information Extraction from Scientific Publications. 2023. P. 101–106.

8. *Ma A. et al.* A deep-learning based citation count prediction model with paper metadata semantic features // Scientometrics. 2021. Vol. 126. No. 8. P. 6803–6823.

9. *Lo K. et al.* PaperMage: A Unified Toolkit for Processing, Representing, and Manipulating Visually-Rich Scientific Documents // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2023. P. 495–507.

10. *Po D.K.* Similarity based information retrieval using Levenshtein distance algorithm // International Journal of Advances in Scientific Research and Engineering. 2020. Vol. 6. No. 04. P. 6–10.

11. *Nurcahyawati V., Mustaffa Z.* Online Media as a Price Monitor: Text Analysis using Text Extraction Technique and Jaro-Winkler Similarity Algorithm // 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE). IEEE, 2020. P. 1–6.

12. *Foppiano L. et al.* Automatic extraction of materials and properties from superconductors scientific literature // Science and Technology of Advanced Materials: Methods. 2023. Vol. 3. No. 1. Article no. 2153633.

13. *Petersen T. et al.* Geo-quantities: A framework for automatic extraction of measurements and spatial context from scientific documents // Proceedings of the 17th International Symposium on Spatial and Temporal Databases. 2021. P. 166–169.

14. *Chraibi A. et al.* Extraction of measurements from medical reports // 10ème conférence Francophone en Gestion et Ingénierie des Systèmes Hospitaliers, GISEH2020. 2020.

15. *Haviana S.F.C., Subroto I. M.I.* Obtaining Reference's Topic Congruity in Indonesian Publications using Machine Learning Approach // 2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). IEEE. 2019. P. 428–431.

16. *Ermakova L. Bordignon F., Turenne N., Noel M.* Is the Abstract a Mere Teaser? Evaluating generosity of article abstracts in the environmental sciences // Frontiers in Research Metrics and Analytics. 2018. Vol. 3. Article no. 16.

17. *El-Ebshihy A. et al.* A platform for argumentative zoning annotation and scientific summarization // Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022. P. 4843–4847.

18. *Choi W. et al.* Building an annotated corpus for automatic metadata extraction from multilingual journal article references // PloS one. 2023. Vol. 18. No. 1. Article no. E0280637.

19. *Krause J. et al.* Bootstrapping multilingual metadata extraction: a showcase in cyrillic // Proceedings of the Second Workshop on Scholarly Document Processing. 2021. P. 66–72.

20. *Shapiro I., Saier T., Färber M.* Sequence Labeling for Citation Field Extraction from Cyrillic Script References // Proceedings of the Workshop on Scientific Document Understanding; co-located with 36th AAAI Conference on Artificial Intelligence (AAAI 2022). 2022.

21. *Indrawati A., Yoganingrum A., Yuwono P.* Evaluating the quality of the indonesian scientific journal references using ParsCit, CERMINE and GROBID // Library Philosophy and Practice. 2019. P. 1–14.

22. *Meuschke N. et al.* A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents // International Conference on Information. Cham: Springer Nature Switzerland, 2023. P. 383–405.

23. *Guo Z., Jin H.* Reference metadata extraction from scientific papers // 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE. 2011. P. 45–49.

24. *Beel J., Langer S., Genzmehr M., M¨uller C.* Docear's PDF inspector: title extraction from PDF files // Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. New York, NY, USA: ACM, 2013. P. 443–444.

25. *Jensen Z. et al.* A machine learning approach to zeolite synthesis enabled by automatic literature data extraction // ACS central science. 2019. Vol. 5. No. 5. P. 892–899.

26. *Färber M., Albers A., Schüber F.* Identifying used methods and datasets in scientific publications // Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence (AAAI 2021). 2021.

27. *Suryawati E., Widyantoro D.H.* Combination of heuristic, rule-based and machine learning for bibliography extraction // 2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME). IEEE. 2017. P. 276–281.

28. *Tkaczyk D. et al.* CERMINE: automatic extraction of structured metadata from scientific literature // International Journal on Document Analysis and Recognition (IJDAR). 2015. Vol. 18. P. 317–335.

29. *Romary L., Lopez P.* Grobid-information extraction from scientific publications // ERCIM News. 2015. Vol. 100.

30. *Councill I.G., Giles C.L., Kan M.Y.* ParsCit: an Open-source CRF Reference String Parsing Package // Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. 2008. Vol. 8. P. 661–667.

31. *Prasad A., Kaur M., Kan M.Y.* Neural ParsCit: a deep learning-based reference string parser // International journal on digital libraries. 2018. Vol. 19. P. 323–337.

**32.** *Constantin A., Pettifer S., Voronkov A.* PDFX: fully-automated PDF-to-XML conversion of scientific literature // Proceedings of the 2013 ACM symposium on Document engineering. 2013. P. 177–180.

# PROCEDURE FOR COMPARING TEXT RECOGNITION SOFTWARE SOLUTIONS FOR SCIENTIFIC PUBLICATIONS BY THE QUALITY OF METADATA EXTRACTION

**I. I. Kuznetsov[1] [0009-0001-6287-8295], O. P. Novikov[2] [0009-0009-3494-3799], D. Y. Ilin[3] [0000-0002-0241-2733]**

*[1, 2]A.N. Kosygin Moscow State Textile University, Moscow, 115035 Russia;*
*[3]MIREA – Russian Technological University, Moscow, 119454 Russia*

[1]iliya-kuznetsov@mail.ru, [2]novikovop55@rambler.ru, [3]i@dmitryilin.com

***Abstract***

Metadata of scientific publications are used to build catalogs, determine the citation of publications, and perform other tasks. Automation of metadata extraction from PDF files provides means to speed up the execution of the designated tasks, while the possibility of further use of the obtained data depends on the quality of extraction. Existing software solutions were analyzed, after which three of them were selected: GROBID, CERMINE, ScientificPdfParser. A procedure for comparing software solutions for recognizing texts of scientific publications by the quality of metadata extraction is proposed. Based on the procedure, an experiment was conducted to extract 4 types of metadata (title, abstract, publication date, author names). To compare software solutions, a dataset of 112,457 publications divided into 23 subject areas formed on the basis of Semantic Scholar data was used. An example of choosing an effective software solution for metadata extraction under the conditions of specified priorities for subject areas and types of metadata using a weighted sum is given. It was determined that for the given example CERMINE shows efficiency 10.5% higher than GROBID and 9.6% higher than ScientificPdfParser.

***Keywords****: text recognition, scientific publications, metadata, data extraction quality, procedure.*

**REFERENCES**

1. *Qayyum F., Afzal M.T.* Identification of important citations by exploiting research articles' metadata and cue-terms from content // Scientometrics. 2019. Vol. 118. P. 21–43.

2. *Liu X., Zhang J., Guo C.* Full-text citation analysis: A new method to enhance scholarly networks // Journal of the American Society for Information Science and Technology. 2013. Vol. 64. No. 9. P. 1852–1863.

3. *Saier T., Färber M.* unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata // Scientometrics. 2020. Vol. 125. No. 3. P. 3085–3108.

4. *Safder I. et al.* Deep learning-based extraction of algorithmic metadata in full-text scholarly documents // Information processing and management. 2020. Vol. 57. No. 6. Article no. 102269.

5. *O'Leary N.A. et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets // Scientific data. 2024. Vol. 11. No. 1. Article no. 732.

6. *Safder I., Hassan S.U.* Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications // Scientometrics. 2019. Vol. 119. P. 257–277.

7. *Joshi B., Symeonidou A., Danish S.M., Hermsen F.* An End-to-End Pipeline for Bibliography Extraction from Scientific Articles // Proceedings of the Second Workshop on Information Extraction from Scientific Publications. 2023. P. 101–106.

8. *Ma A. et al.* A deep-learning based citation count prediction model with paper metadata semantic features // Scientometrics. 2021. Vol. 126. No. 8. P. 6803–6823.

9. *Lo K. et al.* PaperMage: A Unified Toolkit for Processing, Representing, and Manipulating Visually-Rich Scientific Documents // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2023. P. 495–507.

10. *Po D.K.* Similarity based information retrieval using Levenshtein distance algorithm // International Journal of Advances in Scientific Research and Engineering. 2020. Vol. 6. No. 04. P. 6–10.

11. *Nurcahyawati V., Mustaffa Z.* Online Media as a Price Monitor: Text Analysis using Text Extraction Technique and Jaro-Winkler Similarity Algorithm // 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE). IEEE, 2020. P. 1–6.

12. *Foppiano L. et al.* Automatic extraction of materials and properties from superconductors scientific literature // Science and Technology of Advanced Materials: Methods. 2023. Vol. 3. No. 1. Article no. 2153633.

13. *Petersen T. et al.* Geo-quantities: A framework for automatic extraction of measurements and spatial context from scientific documents // Proceedings of the 17th International Symposium on Spatial and Temporal Databases. 2021. P. 166–169.

14. *Chraibi A. et al.* Extraction of measurements from medical reports // 10ème conférence Francophone en Gestion et Ingénierie des Systèmes Hospitaliers, GISEH2020. 2020.

15. *Haviana S.F.C., Subroto I. M.I.* Obtaining Reference's Topic Congruity in Indonesian Publications using Machine Learning Approach // 2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). IEEE. 2019. P. 428–431.

16. *Ermakova L. Bordignon F., Turenne N., Noel M.* Is the Abstract a Mere Teaser? Evaluating generosity of article abstracts in the environmental sciences // Frontiers in Research Metrics and Analytics. 2018. Vol. 3. Article no. 16.

17. *El-Ebshihy A. et al.* A platform for argumentative zoning annotation and scientific summarization // Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022. P. 4843–4847.

18. *Choi W. et al.* Building an annotated corpus for automatic metadata extraction from multilingual journal article references // PloS one. 2023. Vol. 18. No. 1. Article no. E0280637.

19. *Krause J. et al.* Bootstrapping multilingual metadata extraction: a showcase in cyrillic // Proceedings of the Second Workshop on Scholarly Document Processing. 2021. P. 66–72.

20. *Shapiro I., Saier T., Färber M.* Sequence Labeling for Citation Field Extraction from Cyrillic Script References // Proceedings of the Workshop on Scientific Document Understanding; co-located with 36th AAAI Conference on Artificial Intelligence (AAAI 2022). 2022.

21. *Indrawati A., Yoganingrum A., Yuwono P.* Evaluating the quality of the indonesian scientific journal references using ParsCit, CERMINE and GROBID // Library Philosophy and Practice. 2019. P. 1–14.

22. *Meuschke N. et al.* A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents // International Conference on Information. Cham: Springer Nature Switzerland, 2023. P. 383–405.

23. *Guo Z., Jin H.* Reference metadata extraction from scientific papers // 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE. 2011. P. 45–49.

24. *Beel J., Langer S., Genzmehr M., Müller C.* Docear's PDF inspector: title extraction from PDF files // Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. New York, NY, USA: ACM, 2013. P. 443–444.

25. *Jensen Z. et al.* A machine learning approach to zeolite synthesis enabled by automatic literature data extraction // ACS central science. 2019. Vol. 5. No. 5. P. 892–899.

26. *Färber M., Albers A., Schüber F.* Identifying used methods and datasets in scientific publications // Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence (AAAI 2021). 2021.

27. *Suryawati E., Widyantoro D.H.* Combination of heuristic, rule-based and machine learning for bibliography extraction // 2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME). IEEE. 2017. P. 276–281.

28. *Tkaczyk D. et al.* CERMINE: automatic extraction of structured metadata from scientific literature // International Journal on Document Analysis and Recognition (IJDAR). 2015. Vol. 18. P. 317–335.

29. *Romary L., Lopez P.* Grobid-information extraction from scientific publications // ERCIM News. 2015. Vol. 100.

30. *Councill I.G., Giles C.L., Kan M.Y.* ParsCit: an Open-source CRF Reference String Parsing Package // Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. 2008. Vol. 8. P. 661–667.

31. *Prasad A., Kaur M., Kan M.Y.* Neural ParsCit: a deep learning-based reference string parser // International journal on digital libraries. 2018. Vol. 19. P. 323–337.

32. *Constantin A., Pettifer S., Voronkov A.* PDFX: fully-automated PDF-to-XML conversion of scientific literature // Proceedings of the 2013 ACM symposium on Document engineering. 2013. P. 177–180.

## СВЕДЕНИЯ ОБ АВТОРАХ

***КУЗНЕЦОВ Илия Игоревич*** – аспирант кафедры искусственного интеллекта, прикладной математики и программирования Института информационных технологий и цифровой трансформации, Российский государственный университет им. А.Н. Косыгина (Технологии. Дизайн. Искусство).

***Ilia Igorevich KUZNETSOV*** – postgraduate student of the Department of Artificial Intelligence, Applied Mathematics and Programming, Institute of Information Technologies and Digital Transformation, A. N. Kosygin Moscow State Textile University.

email: iliya-kuznetsov@mail.ru

ORCID: 0009-0001-6287-8295

***НОВИКОВ Олег Пантелеевич*** – доктор технических наук, профессор, профессор кафедры искусственного интеллекта, прикладной математики и программирования Института информационных технологий и цифровой трансформации, Российский государственный университет им. А.Н. Косыгина (Технологии. Дизайн. Искусство).

***Oleg Panteleevich NOVIKOV*** – Doctor of Engineering, professor, professor of the Department of Artificial Intelligence, Applied Mathematics and Programming of the Institute of Information Technology and Digital Transformation, A. N. Kosygin Moscow State Textile University.

email: novikovop55@rambler.ru

ORCID: 0009-0009-3494-3799

***ИЛЬИН Дмитрий Юрьевич*** – кандидат технических наук, доцент кафедры КБ-14 «Цифровые технологии обработки данных» Института кибербезопасности и цифровых технологий, МИРЭА – Российский технологический университет.

***Dmitry Yurievich ILIN*** – Candidate of Sciences (Engineering), associate professor at the department of Data Processing Digital Technologies, Institute of Cybersecurity and Digital Technologies, MIREA – Russian Technological University.

email: i@dmitryilin.com

ORCID: 0000-0002-0241-2733

*Материал поступил в редакцию 5 апреля 2025 года*