Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

Метод выделения составных частей научного текста на основе анализа распределения пикселей в сканирующей строке

Студент	ИУ7-84Б		К. А. Рунов
	(группа)	(подпись)	(инициалы, фамилия)
Руководитель ВКР			Ю. В. Строганов
		(подпись)	(инициалы, фамилия)
Консультант			
		(подпись)	(инициалы, фамилия)
Консультант			
		(подпись)	(инициалы, фамилия)
Нормоконтролер			А. С. Кострицкий
		(подпись)	(инициалы, фамилия)

2025 год

РЕФЕРАТ

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1 Аналитический раздел

1.1 Анализ предметной области

1.1.1 Анализ структуры документов

Анализ структуры документов (Document layout analysis, DLA) — процесс сегментирования входного изображения документа на однородные компоненты, такие как блоки текста, рисунки, таблицы, графики и т.д., и их классификации [1].

В общем случае анализ структуры документа делится на два взаимосвязанных процесса: физический и логический анализ. Целью физического анализа является выявление структуры документа и определение границ его однородных областей. Целью логического анализа является разметка обнаруженных областей. Выявленные области классифицируются как элементы документа — рисунки, заголовки, абзацы, логотипы, подписи и другие. [2]

Процесс анализа структуры документов состоит из двух основных этапов — этапа предварительной обработки и этапа анализа макета документа [2, 3]. На рисунке ниже приведена схема процесса анализа структуры документов.

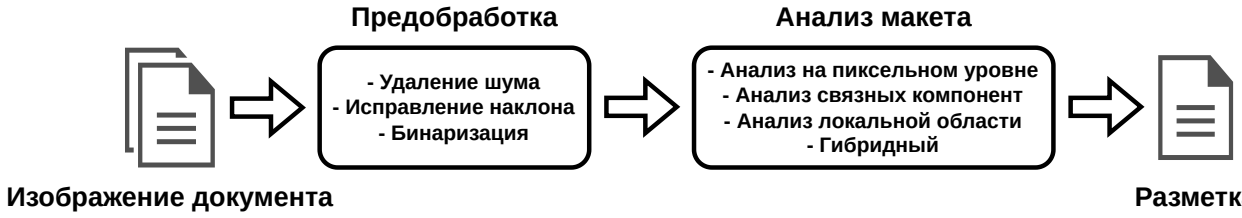

Рисунок 1 – Схема процесса анализа структуры документов [3]

Этап предварительной обработки

Этап анализа макета документа в любом методе анализа структуры документов (далее DLA) часто основывается на определённых предположениях о входных изображениях, таких как отсутствие шума, бинаризация, отсутствие наклона текста или все перечисленные факторы [2, 3].

Цель этапа предварительной обработки — преобразовать входное изоб-

ражение в соответствии с требованиями этапа анализа макета документа конкретного метода [2, 3].

В общем случае на этом этапе используются одна или несколько процедур предварительной обработки, таких как бинаризация, выравнивание и улучшение изображения [2, 4].

Этап анализа макета документа

Анализ макета документа включает в себя определение границ и типов составляющих областей входного изображения документа. Процесс определения границ областей документа называется сегментацией областей документа, а классификация найденных областей по их типу — классификацией областей документа. [3]

Существуют три типа стратегий анализа макета документа: снизу вверх (bottom-up), сверху вниз (top-down) и гибридная (hybrid).

По стратегии снизу вверх (bottom-up) параметры анализа часто вычисляются на основе исходных данных. Анализ макета документа начинается с небольших элементов, таких как пиксели или связанные компоненты. Затем однородные элементы объединяются, создавая более крупные области. Процесс продолжается, пока не будут достигнуты заранее определённые условия остановки.

По стратегии сверху вниз (top-down) анализ макета документа начинается с крупных областей, например, на уровне всего документа. Затем эта большая область разбивается на более мелкие, такие как колонки текста, на основе определённых правил однородности. Анализ сверху вниз прекращается, когда дальнейшее разбиение областей становится невозможным или достигаются условия остановки.

Гибридная стратегия (hybrid) представляет собой комбинацию обеих стратегий (снизу вверх и сверху вниз). [2]

После сегментации областей происходит их классификация с помощью различных алгоритмов, в результате чего формируется логическая структура документа.

По завершении данного этапа извлеченные геометрическая и логическая структуры сохраняются для последующей реконструкции. Для этого, как правило, используется иерархическая древовидная структура данных. [3]

1.1.2 Типы макетов документов

Макеты документов могут иметь различные структуры. Печатные документы можно разделить на шесть типов [5]: прямоугольные, Манхэттенские, не-Манхэттенские, многоколоночные Манхэттенские, с горизонтальным наложением и с диагональным наложением.

Рисунок 2 – Макеты документов: (a) Стандартный (прямоугольный), (b) Манхэттенский, (c) Не-Манхэттенский, (d) Многоколоночный Манхэттенский, (e) Произвольный (сложный), (f) С горизонтальным и диагональным наложением. [2]

На рисунке выше показаны примеры описанных типов макетов документов:

— Стандартный макет характеризуется большими прямоугольными текстовыми блоками, расположенными в одной или нескольких колонках, при этом каждая колонка содержит по одному абзацу.

— Если документ содержит несколько абзацев в колонках, его можно отнести к Манхэттенскому макету. Примеры таких документов — научно-технические статьи, журналы и другие.

— Не-Манхэттенские макеты включают зоны непрямоугольной формы.

— Макеты с наложением содержат элементы, такие как текст, который перекрывает другие элементы документа. Наложение может возникать, например, из-за просвечивания (см. Рисунок 2(f)).

— Документы с произвольными (или сложными) макетами могут включать рукописный и/или печатный текст, содержащий различные стили, типы и размеры шрифтов.

Таким образом, документы, содержащие научно-технические тексты, обычно используют Манхэттенский макет.

1.1.3 Структура научно-технического текста

Научно-технический текст обычно [6, 7, 8] следует четко определенному шаблону и имеет следующую структуру:

1) Название;

2) Информация об авторах;

3) Аннотация и ключевые слова;

4) Введение;

5) Основная часть (кроме текста содержащая в том числе таблицы, рисунки, графики, листинги);

6) Заключение;

7) Ссылки на литературу.

Содержимое научного текста часто не ограничивается текстом, а содержит также следующие составные части:

1) таблицы,

2) листинги,

3) схемы алгоритмов,

4) рисунки,

5) графики.

Зная структуру научного текста и его основные части можно перейти к формализации задачи выделения составных частей научного текста.

1.2 Формализация предметной области

Пусть D — документ, представленный в виде набора изображений, содержащих текст, листинги, таблицы, рисунки и прочие структурные элементы.

Документ

$$D = \{P_1, P_2, \ldots, P_n\}$$

состоит из страниц $P_1, P_2, \ldots, P_n$, а каждая страница P_i в свою очередь содержит множество объектов $O_{i,1}, O_{i,2}, \ldots, O_{i,m}$.

Объект $O_{i,j}$ — кортеж $(x_{i,j}, y_{i,j}, w_{i,j}, h_{i,j})$, где $(x_{i,j}, y_{i,j})$ — координаты верхнего левого угла, $w_{i,j}$ — ширина, $h_{i,j}$ — высота объекта.

Требуется построить отображение

$$F : D \to \{(O_{i,j}, C_{i,j})\},$$

где каждому объекту $O_{i,j}$ ставится в соответствие класс

$$C_{i,j} = C_{i,j}(O_{i,j}),$$

область допустимых значений которого определяется исходя из требований к разметке.

Например, в случае задачи выделения составных частей научного текста, $C_{i,j} \subseteq \{$Фон, Текст, Таблица, Листинг, Схема алгоритма, Рисунок, График, Неопределенность$\}$; Объект классифицируется как «Неопределенность» в случае, когда не удалось распределить его ни в один из предыдущих классов.

Поставленную задачу можно решить, разбив на две подзадачи и решив каждую подзадачу соответственно: первая подзадача — нахождение объектов на страницах и выявление их геометрических свойств, вторая подзадача — классификация найденных объектов (определение $C_{i,j}$ для каждого объекта $O_{i,j}$).

Решением первой подзадачи является построение отображений

$$P_i \to \{O_{i,j}\},$$

решением второй подзадачи является построение отображений

$$O_{i,j} \to C_{i,j}.$$

Далее будут рассмотрены существующие методы, позволяющие решить поставленную задачу.

1.3 Описание существующих методов

1.3.1 Анализ связных компонент

Методы на основе связных компонент (Connected Component Analysis, CCA) анализируют и объединяют связные компоненты для формирования однородных областей.

Определение начальных компонент, которые впоследствии объединяются, происходит, как правило, следующим образом. Изображение проходит стадию бинаризации (преобразование к черно-белому формату и назначение каждому пикселю интенсивности 0 или 1), после чего смежные пиксели объединяются на основе 4- или 8-связности. При 4-связности два пикселя считаются связными, если они расположены друг за другом по горизонтали или вертикали. При 8-связности два пикселя считаются связными, если они являются 4-связными, либо расположены друг за другом по диагонали.

После определения начальных компонент, компоненты объединяются в однородные области путем применения специальных алгоритмов. В качестве таких алгоритмов могут выступать, например, преобразование Хафа (Hough transform) или алгоритм K-ближайших соседей (K-nearest neighbor, KNN) [3].

Далее происходит классификация однородных областей. Для классфикации области могут использоваться эвристические алгоритмы (классификация на основе ширины штриха, размера или формы компонента) и алгоритмы на основе машинного обучения.

Методы на основе связных компонент могут работать в условиях скошенного текста при условии, что межстрочный интервал меньше пробела

между абзацами [3].

1.3.2 Анализ проекционного профиля

Суть методов на основе анализа проекционного профиля (Projection Profile Analysis, PPA) заключается в следующем. Пиксели документа проецируются на вертикальную и горизонтальную ось, после чего строятся гистограммы распределения пикселей. Далее анализируются пики и впадины на гистограммах. Впадины на вертикальном профиле указывают на пробелы между колонками текста. Впадины на горизонтальном профиле указывают на пробелы между строками или блоками текста.

На основе проведенного анализа можно разделить документ на логические компоненты — текстовые блоки, заголовки, таблицы, изображения.

Данные методы работают только с Манхэттенскими макетами документов и чувствительны ко скошенности текста в документе [3].

1.3.3 Алгоритм размазывания по длине серии

Методы на основе RLSA преобразуют бинарное изображение документа путем «размазывания» черных пикселей горизонтально и/или вертикально для формирования однородных областей.

На рисунке ниже можно видеть пример работы RLSA.

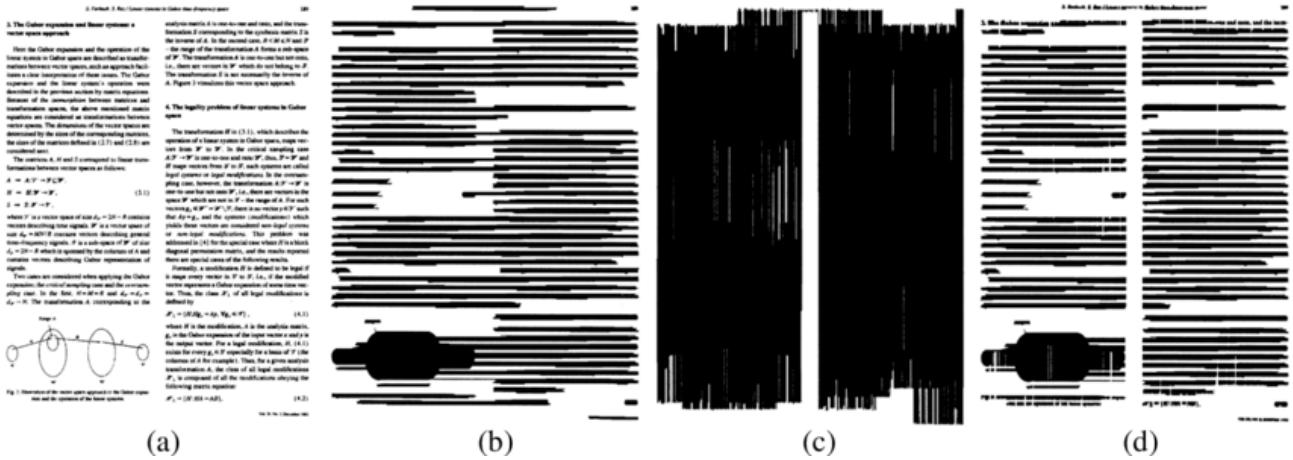

Рисунок 3 – Пример работы алгоритма RLSA. К начальному изображению документа (a) применяется горизонтальный (b) и вертикальный (c) RLSA, после чего в результате применения операции И к изображениям (b) и (c) формируется (d)

Для классификации полученных областей также используются эвристические алгоритмы и алгоритмы на основе машинного обучения.

Данные методы, как и методы на основе анализа проекционного профиля, работают преимущественно с Манхэттенскими макетами документов и чувствительны к скошенному тексту [3].

1.3.4 Методы на основе машинного обучения

Методы, не основанные на глубоком обучении, используют простые архитектуры нейросетей для обучения. Анализ с использованием нейросети происходит на трех уровнях: уровне пикселей, уровне блоков текста и уровне страниц.

Методы на основе машинного обучения в области анализа макетов документов страдают от несбалансированности данных и отсутствия контекстной информации. Если модель обучалась на документах, в наборе данных для обучения количество текстовой и фоновой информации сильно превосходит количество информации о рисунках и графиках. В связи с этим обученная модель может склоняться в сторону текстовых или фоновых пикселей. [2]

Обучение моделей лишь на основе информации о пикселях чревато потерей контекстной информации. Поэтому при обучении на уровнях блоков текста и страниц прибегают к использованию методов извлечения признаков для создания более надежных моделей. [2]

Методы на основе машинного обучения работают с любыми макетами документов и не чувствительны к скошенному тексту.

1.3.5 Гибридные методы на основе PPA и CCA

Гибридные методы на основе анализа проекционного профиля и связных компонент используют работают следующим образом. Начальные компоненты определяются применением метода из PPA, после чего происходит их уточнение применением методов из CCA.

Такой комбинированный подход позволяет лучше сегментировать текст, чем каждый метод по отдельности.

1.4 Классификация существующих методов

Для сравнения рассмотренных методов можно выделить следующие критерии:

— Стратегия анализа макета документа;

— Скорость работы — требования метода к вычислительным ресурсам;

— Гибкость — способность метода адаптироваться к различным типам макетов документов;

— Устойчивость — способность метода адаптироваться к шумам и искажениям текста.

— Специальное требование — позволяет сегментировать не только текст, но и такие составные части научного текста, как таблицы, листинги, схемы, рисунки, графики и прочее.

Ниже приведена таблица со сравнительным анализом рассмотренных методов.

Таблица 1 – Классификация методов DLA

Метод	**Стратегия**	**Скорость**	**Гибкость**	**Ус-ть**	**СпецТреб**
CCA	Снизу вверх	2	2	3	Нет
PPA	Сверху вниз	2	3	3	Нет
RLSA	Сверху вниз	1	3	3	Нет
ML	Снизу вверх	3	1	1	Да
PPA + CCA	Гибридный	2	3	2	Нет

Как можно видеть по сравнительной таблице, ни один из рассмотренных методов не позволяет «быстро», на основе лишь различных эвристик, произвести разметку документа, сегментирующую не только текст, но и другие составные части научного текста.

1.5 Формализованная постановка задачи

Целью данной работы является разработка метода, позволяющего выделять составные части научного текста на основе простых правил, без ис-

пользования нейросетей, а также разработка алгоритма, реализующего данный метод.

Формализованная в виде IDEF0 диаграммы постановка задачи представлена на рисунке 4 ниже.

Рисунок 4 – Постановка задачи

Вывод

2 Конструкторский раздел

2.1 Требования и ограничения метода

Метод выделения составных частей научного текста на основе анализа распределения пикселей в сканирующей строке должен:

1) Работать с одноколоночными Манхэттенскими макетами документов;
2) Выделять текстовые блоки;
3) Выделять таблицы;
4) Выделять листинги;
5) Выделять схемы алгоритмов;
6) Выделять рисунки;
7) Выделять графики;
8) Работать на основе простых правил и эвристик, без использования нейросетей.

2.2 Описание разрабатываемого метода

Поставленная задача решается в четыре этапа:

1) Преобразование PDF документа в изображения;
2) Первичная разметка страниц;
3) Создание уточненной разметки на основе первичной;
4) Объединение уточненной разметку в более крупные блоки.

Разметка, ее уточнение и объединение происходят на основе определенных правил, которые будут описаны в данном разделе далее.

Основные этапы разрабатываемого метода представлены на IDEF0 диаграмме первого уровня (см. Рисунок 5).

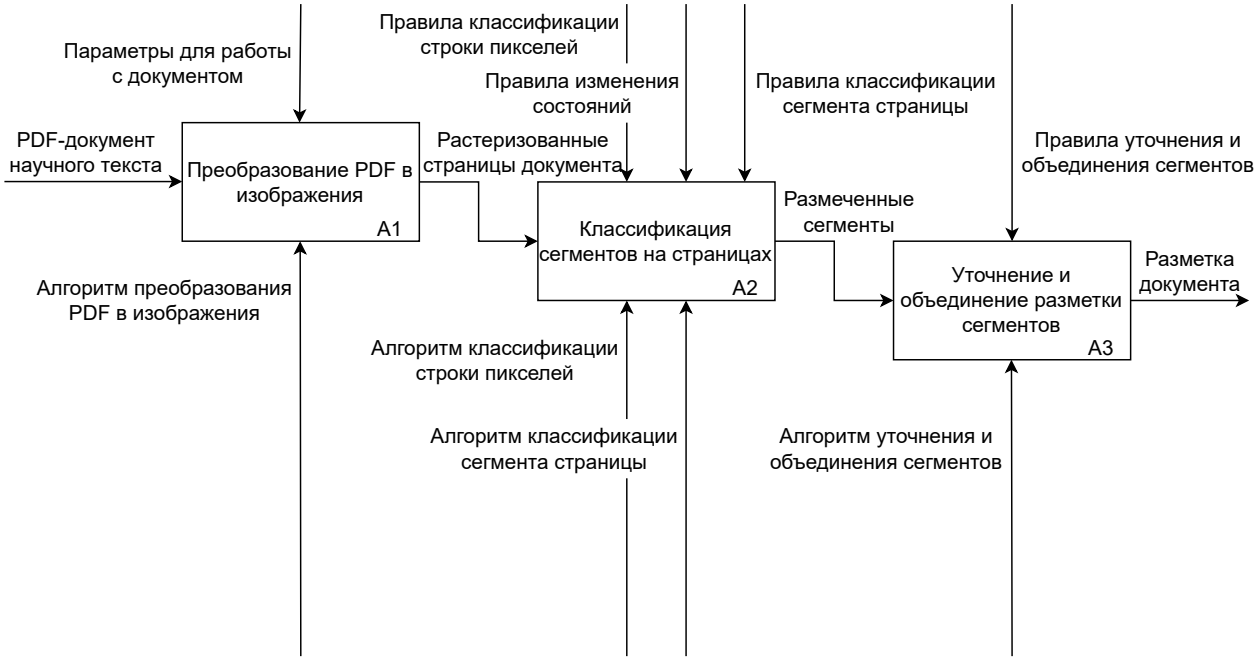

Рисунок 5 – IDEF0-диаграмма метода выделения составных частей научного текста на основе анализа распределения пикселей в сканирующей строке

2.2.1 Первичная разметка

Исходные данные — изображение страницы документа. Получаемый результат — разметка страницы и информация о каждом сегменте на ней.

Разметка страницы — массив кортежей типа

$$(y\_start, y\_end, C),$$

где y_start — y-координата начала сегмента в пространстве изображения, y_end — y-координата конца сегмента в пространстве изображения, C — класс сегмента, где $C \subseteq$ {Фон, Немного текста, Много текста, Цвет, Черная линия средней длины, Длинная черная линия, Не определено}.

Информация о сегменте содержит следующие данные:

1) start — ордината начала сегмента;

2) end — ордината конца сегмента;

3) count_long_black_line — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Длинная черная линия»;

4) count_single_long_black_line — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Длинная черная линия», считая несколько подряд идущих «Длинных черных линий» за одну;

5) count_medium_black_line — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Черная линия средней длины»;

6) count_single_medium_black_line — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Черная линия средней длины», считая несколько подряд идущих «Черных линий средней длины» за одну;

7) count_total_medium_black_line — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Черная линия средней длины», с учетом всех «Черных линий черной длины» если таких было зафиксировано несколько внутри одной сканирующей строки;

8) count_many_text — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Много текста»;

9) count_few_text — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Немного текста»;

10) count_color — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Цвет»;

11) count_undefined — количество раз, когда при разметке сегмента встретилась строка, идентифицированная, как «Не определено»;

12) count_white_px — количество белых пикселей в сегменте;

13) count_color_px — количество цветных пикселей в сегменте;

14) count_gray_px — количество черных пикселей в сегменте (сумма трех данных счетчиков дает общее количество пикселей в сегменте);

15) heatmap_black — массив, i-й элемент которого отражает количество черных пикселей в i-й колонке пикселей сегмента;

16) heatmap_color — массив, i-й элемент которого отражает количество цветных пикселей в i-й колонке пикселей сегмента.

Данная информация будет использоваться при уточнении разметки в следующем этапе.

2.2.2 Уточненная разметка

Исходные данные — разметка страницы и информация о каждом сегменте на ней. Получаемый результат — уточненная разметка страницы.

Уточненная разметка страницы — массив кортежей типа

$$(y\_start, y\_end, C),$$

где y_start — y-координата начала сегмента в пространстве изображения, y_end — y-координата конца сегмента в пространстве изображения, C — класс сегмента, где $C \subseteq$ {Фон, Текст, Таблица, Листинг, Схема алгоритма, Рисунок, График, Не определено}, причем координаты начала и конца сегментов совпадают с соответствующими координатами сегментов первичной разметки, уточняется только класс на основе информации о сегментах.

2.2.3 Объединенная разметка

Исходные данные — уточненная разметка страницы. Получаемый результат — объединенная разметка страницы.

2.3 Тестирование и классы эквивалентности

2.4 Структура разрабатываемого ПО

Вывод

3 Технологический раздел

3.1 Выбор средств реализации

3.2 Реализация программного обеспечения

3.3 Результаты тестирования

3.4 Пользовательский интерфейс

3.5 Руководство пользователя

Вывод

4 Исследовательский раздел

4.1 Описание исследования

4.2 Результаты исследования

Вывод

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Bhowmik et al. Text and non-text separation in offline document images: a survey // International Journal on Document Analysis and Recognition (IJDAR). 2018. Т. 21.

2. Binmakhashen G.M., Mahmoud S.A. Document Layout Analysis: A Comprehensive Survey // ACM Comput. Surv. 2019. Т. 52, № 6.

3. Bhowmik S. Document Layout Analysis. — Springer Singapore, 2023 — 86 с.

4. Kasturi R., O'Gorman L., Govindaraju V. Document image analysis: A primer // Sadhana — Academy Proceedings in Engineering Sciences. 2002. Т. 27. С. 3–22.

5. Kise K. Page Segmentation Techniques in Document Analysis // Doermann D., Tombre K. Handbook of Document Image Processing and Recognition. — Springer London, 2014. С. 135–175.

6. Бутенко Ю.И. Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. Т. 20, № 1. С. 5–13.

7. Романов Д.А. Кратко о структуре экспериментальной научной статьи на английском языке // Вестник Казанского технологического университета. 2014. Т. 17, № 6. С. 325–327.

8. Раицкая Л.К. Структура научной статьи по политологии и международным отношениям в контексте качества научной информации // Полис. Политические исследования. 2019. № 1. С. 167–181.

ПРИЛОЖЕНИЕ А