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)

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

Студент: Рунов Константин Алексеевич, ИУ7-84Б
Научный руководитель: Строганов Юрий Владимирович

2025 г.

Область применения

Для задач нормоконтроля требуется выделять составные части текстов, такие как текст, таблицы, листинги, рисунки, графики, схемы.

Стандарт МГТУ для проведения нормоконтроля составляет 40 страниц за 8 часов.

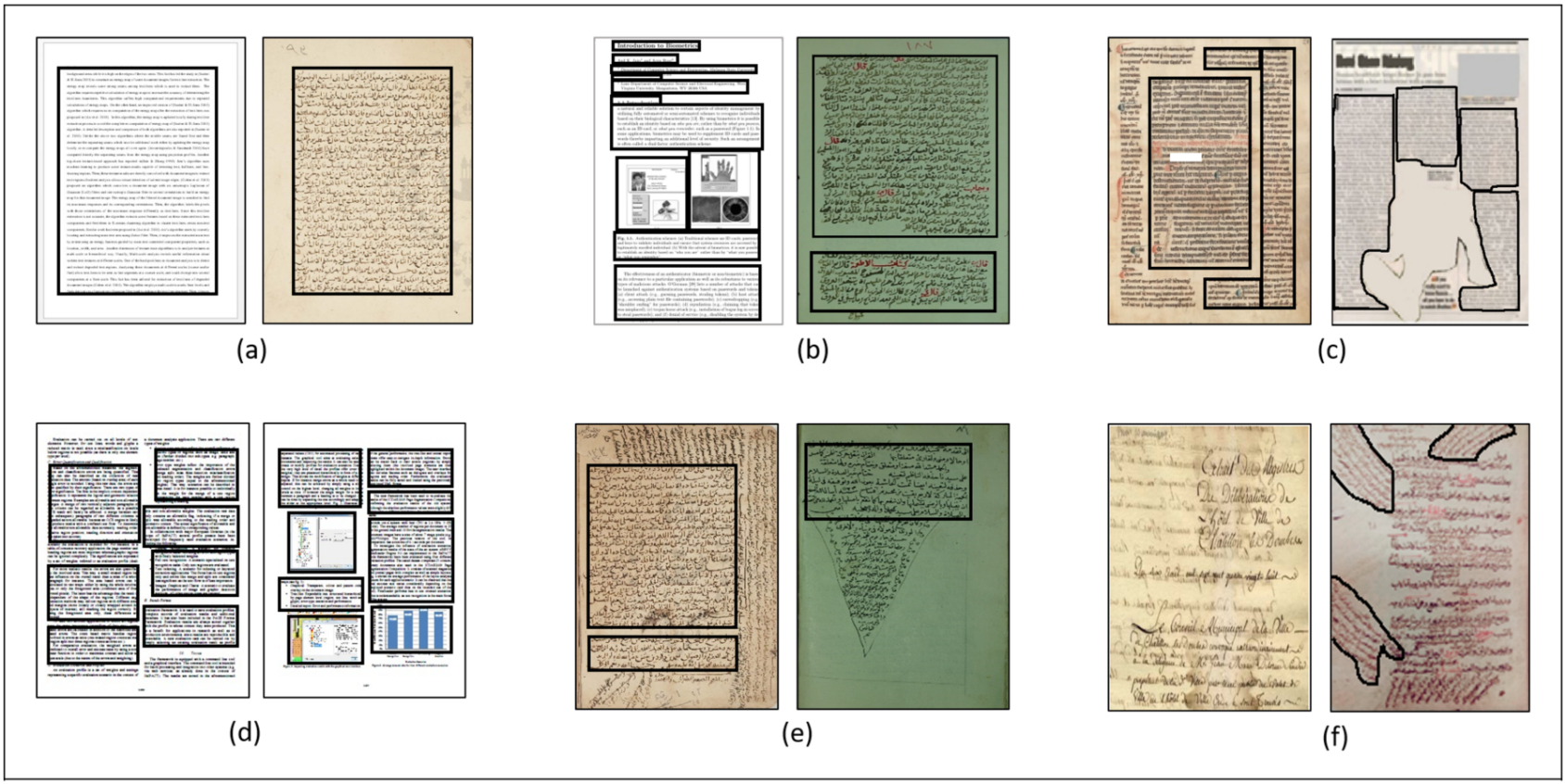
Количество документов, требующих обработки, постоянно увеличивается.

Цель и задачи

Целью работы является разработка метода для автоматического выделения составных частей научного текста, использующего только простые эвристики для классификации сегментов документа.
Задачи:

- Рассмотреть и сравнить существующие методы сегментации документов;
- Формализовать постановку задачи;
- Разработать описанный метод;
- Разработать программное обеспечение, реализующее данный метод;
- Исследовать скорость разметки и максимального объема используемой памяти в зависимости от количества процессов, участвующих в разметке.

Типы макетов документов



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

Классификация существующих методов сегментации

Метод \ Критерий	Скорость	Гибк.	Уст-ть	СпецТреб
Con. Comp. An.	2	2	3	Нет
Proj. Prof. An.	2	3	3	Нет
Run-Len. Sm. Alg.	1	3	3	Нет
Machine Learning	3	1	1	Да
PPA + CCA	2	3	2	Нет
Разраб. метод	1	3	3	Да

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 другие его составные части.

Формализация задачи

Документ $D = \{P_1, P_2, \ldots, P_n\}$ состоит из страниц $P_1, \ldots, P_n$, а каждая страница P_i состоит из множества сегментов $S_{i,1}, \ldots, S_{i,m}$.

Сегмент $S_{i,j}$ — 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 сегмента.

Требуется построить отображение

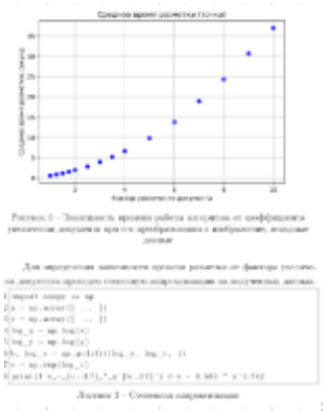
$$F : D \to \{(S_{i,j}, C_{i,j})\},$$

где каждому сегменту $S_{i,j}$ ставится в соответствие класс

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

область допустимых значений которого определяется согласно требованиям к разметке.

Предлагаемый метод

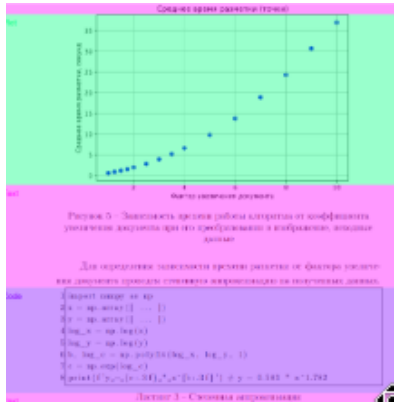


PDF-документ научного текста

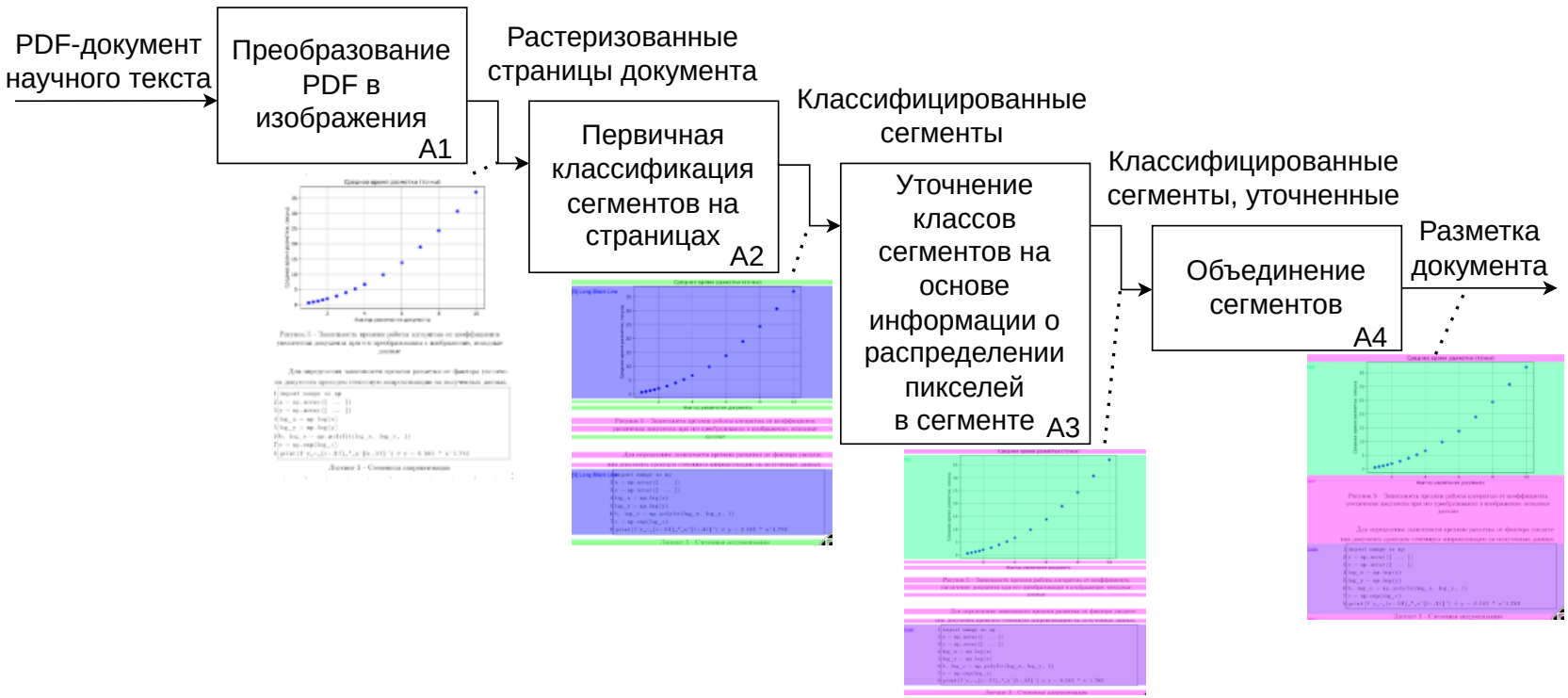
Разметка научного текста на основе анализа распределения пикселей

A0

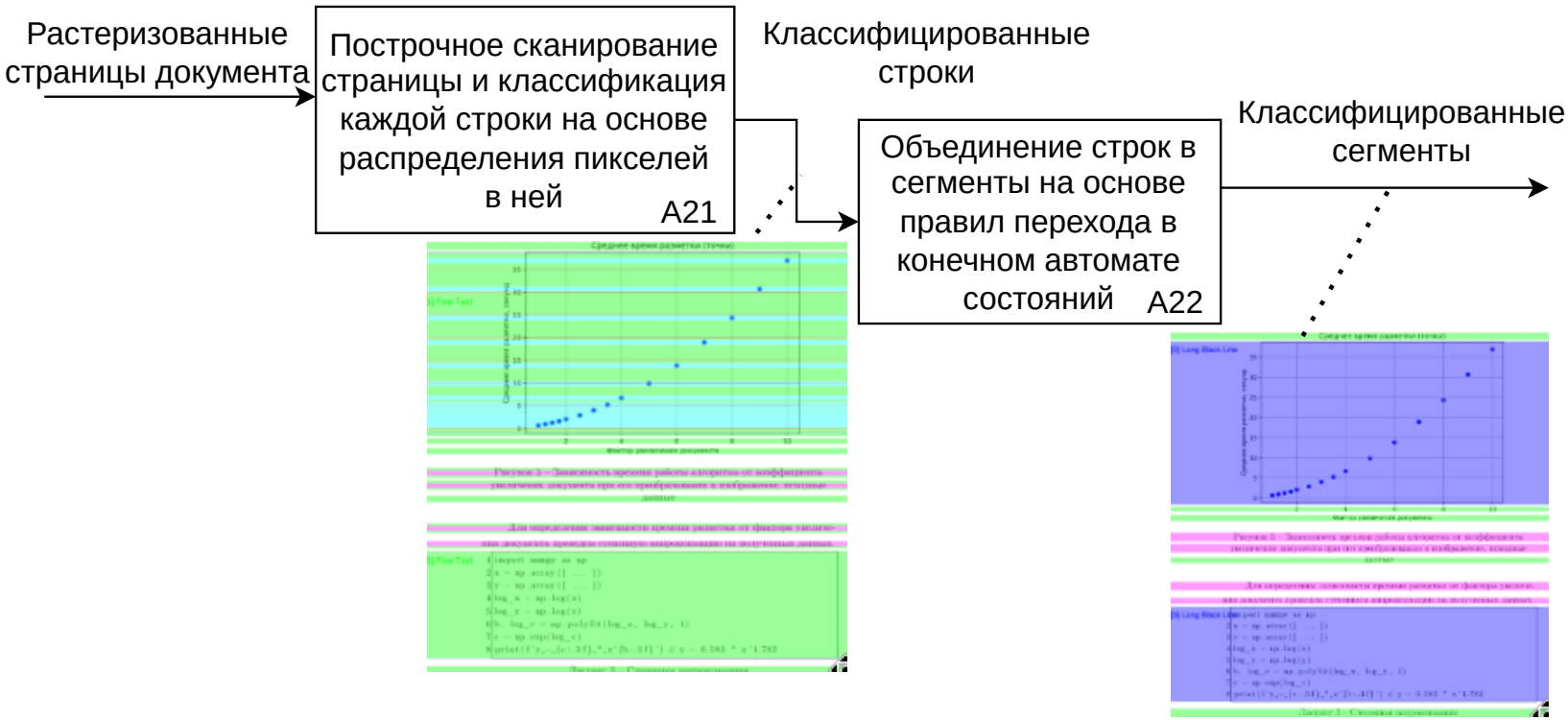
Разметка документа



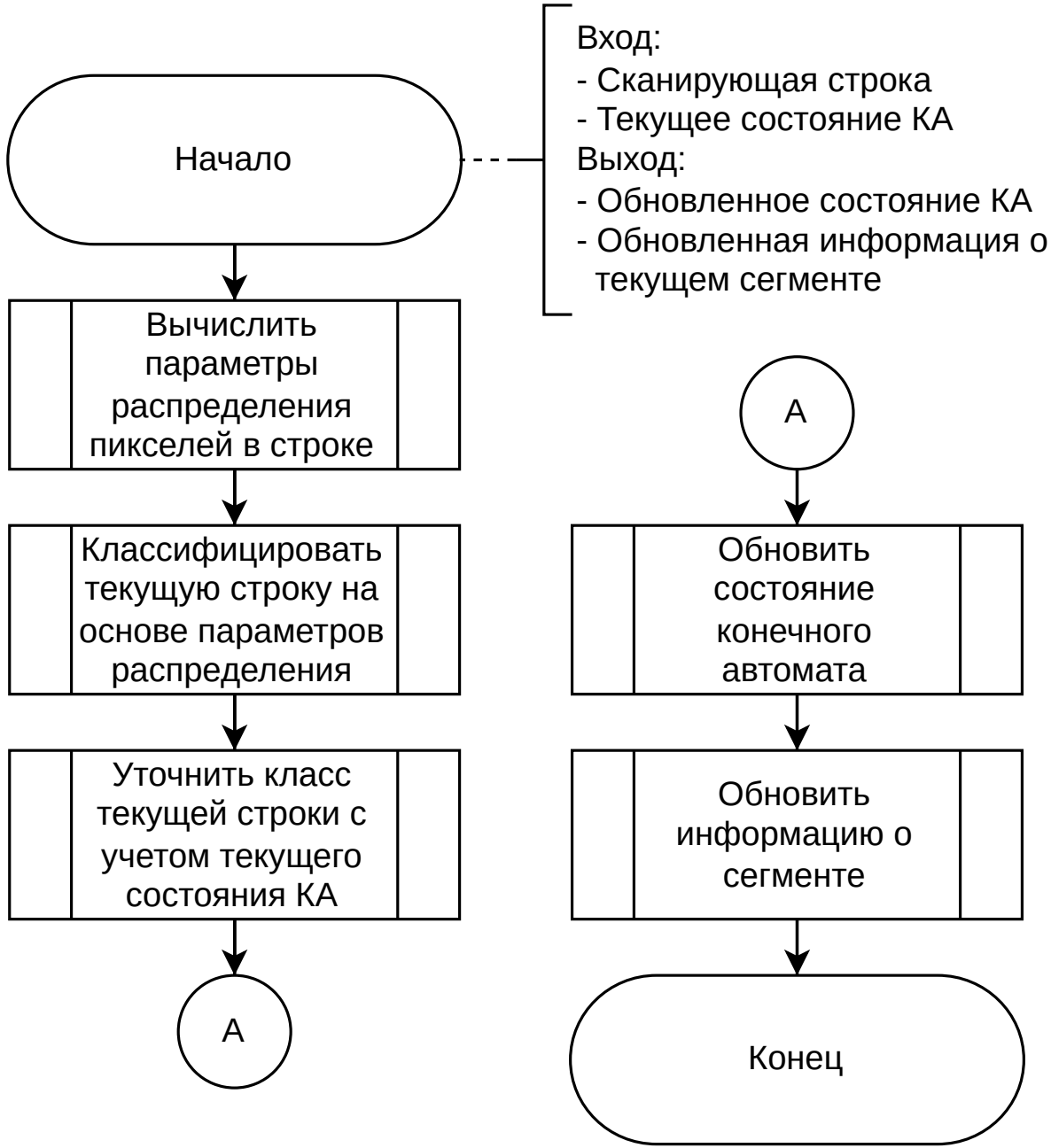
Предлагаемый метод, детализация



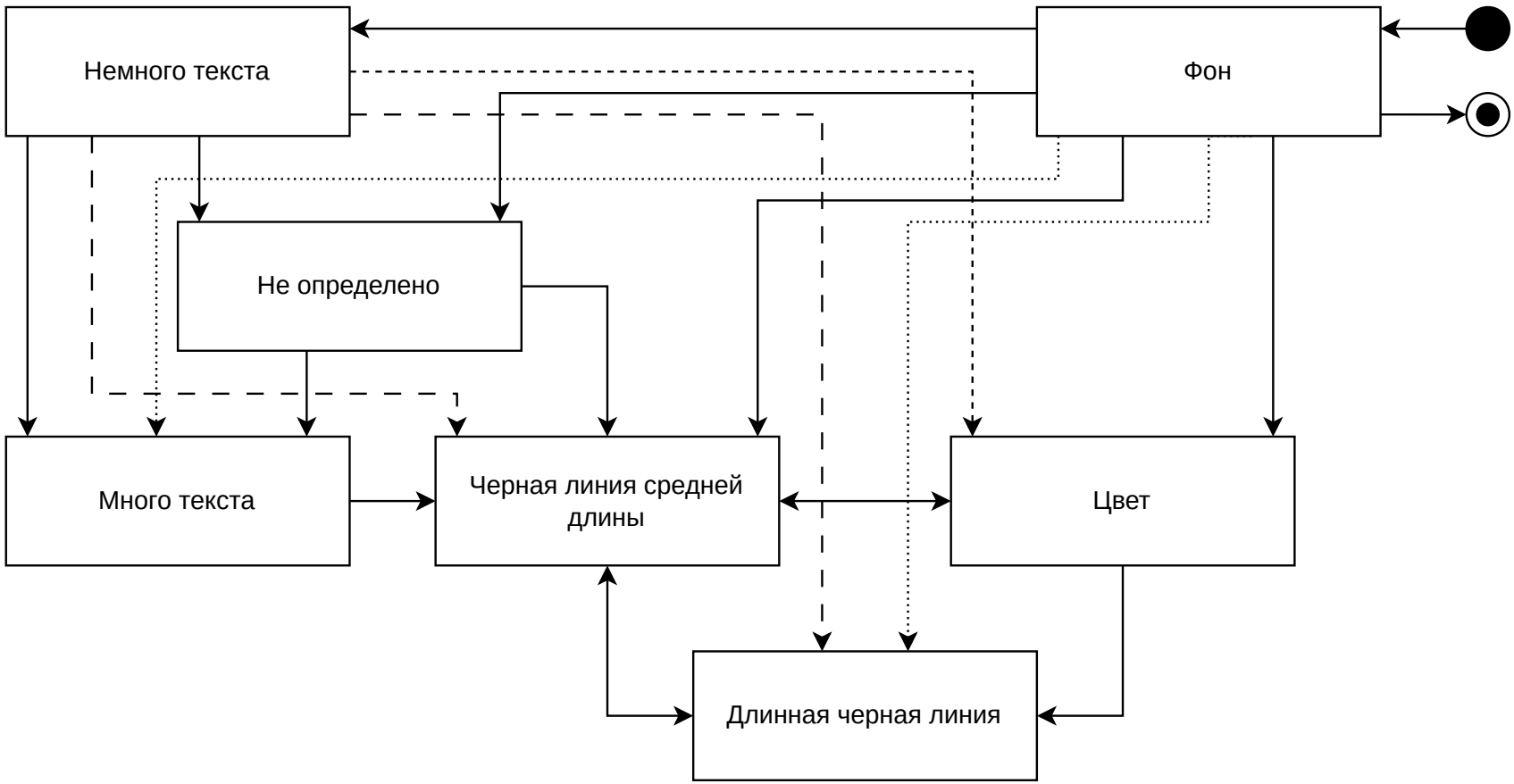
Предлагаемый метод, детализация первичной разметки



Первичная разметка



Конечный автомат состояний сканирующей строки при первичной разметке



Первичная разметка. Примеры правил

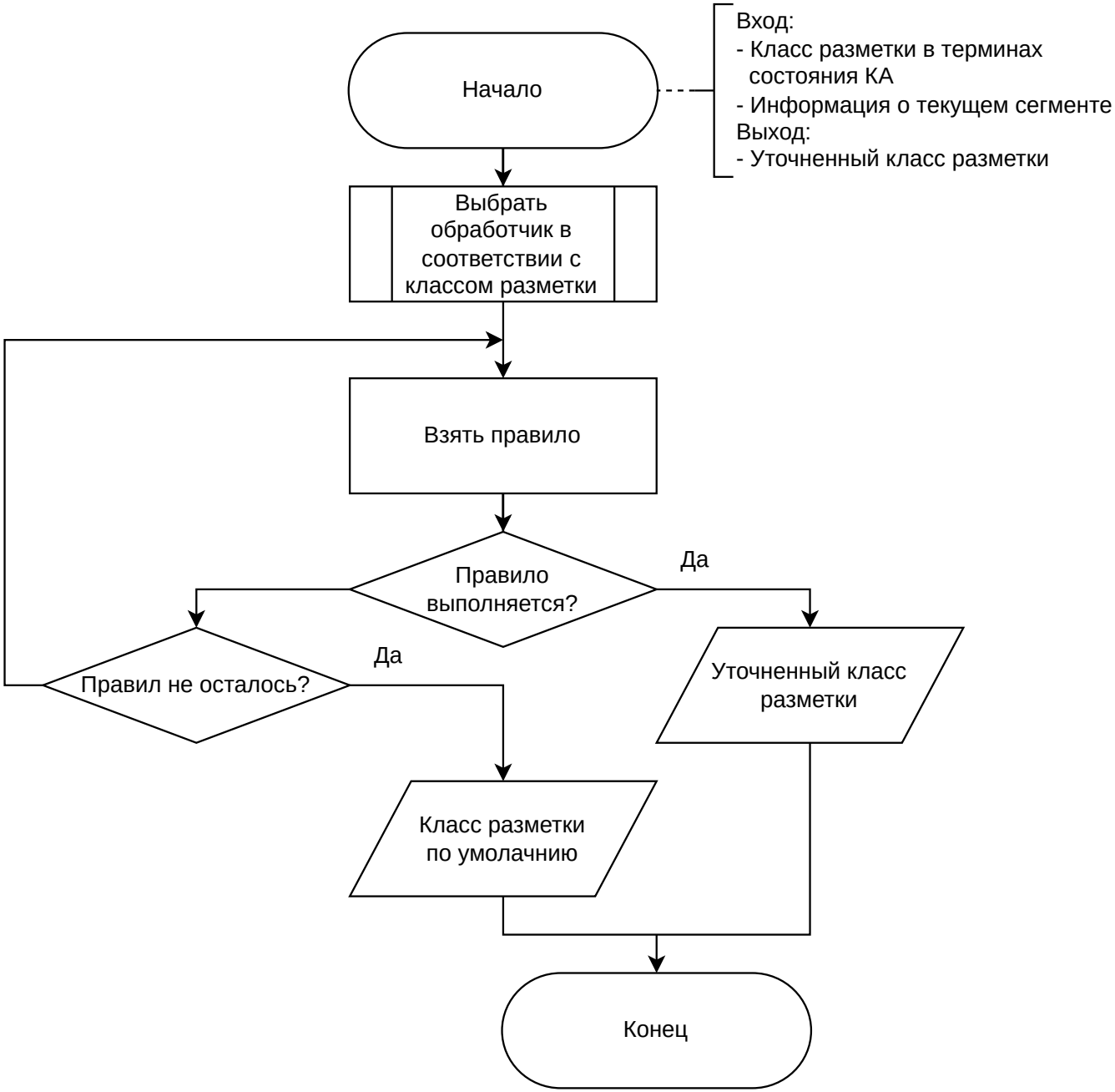
Классификация сканирующей строки:

- Если в сканирующей строке большое количество комнонент из смежных черных пикселей, строка относится к классу «Много текста»;
- Если сканирующая строка содержит цветные пиксели, она относится к классу «Цвет»;
- Если сканирующая строка содержит единственную компоненту из смежных черных пикселей длиной почти во всю ширину документа, она относится к классу «Длинная черная линия».

Обновление состояния конечного автомата:

- Если КА находится в состоянии «Фон» и встречает строку, содержащую черные пиксели, и их распределение не похоже ни на текст, ни на черные линии, КА переходит в состояние «Не определено».

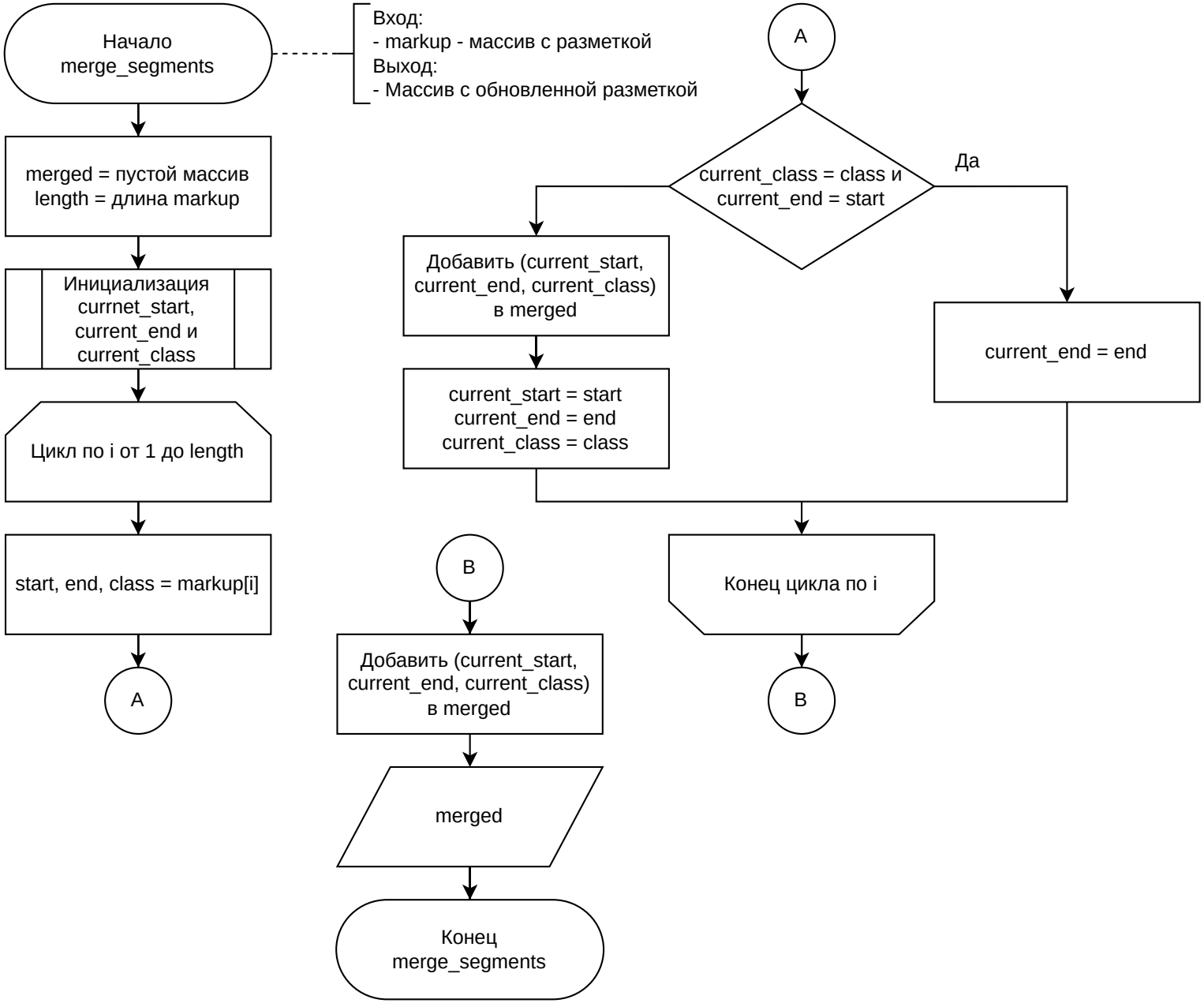
Алгоритм уточненной разметки



Уточненная разметка. Примеры правил

- Если сегмент был классифицирован, как «Не определено», при этом его высота небольшая ИЛИ много строк в сегменте было классифицировано, как «Немного текста», то уточненный класс сегмента будет «Текст»;
- Если сегмент был классифицирован, как «Много текста», но при этом из информации о сегменте видно, что в нем содержится больше двух столбцов черных пикселей высотой с сегмент, то его уточненный класс «Таблица»;
- Если сегмент был классифицирован, как «Цвет», и содержит одну вертикальную линию, а также количество белых пикселей в сегменте преобладает, то его уточненный класс «График».

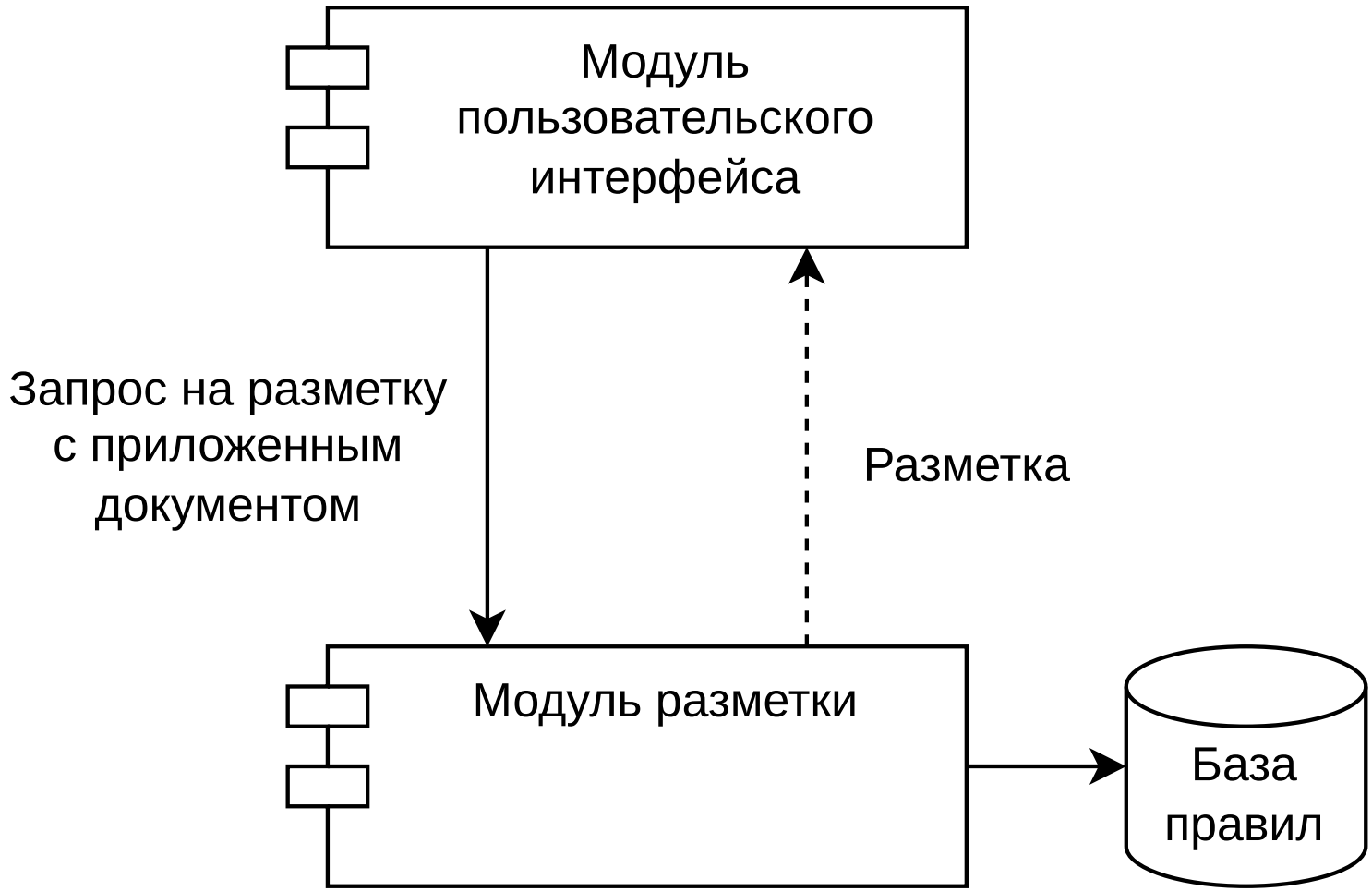
Алгоритм объединения сегментов



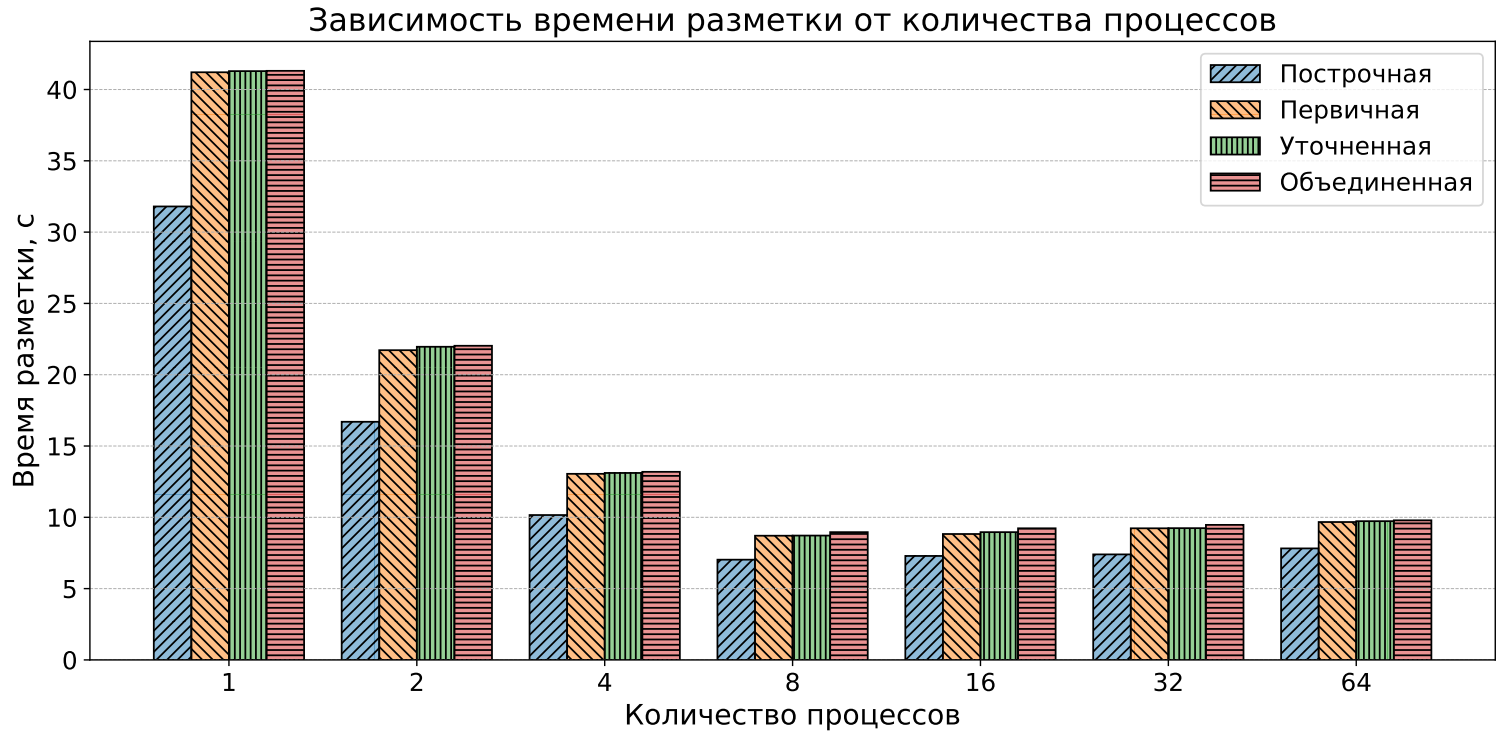
Объединенная разметка. Примеры правил

- Маленькие (меньше 30 пикселей) фоновые сегменты сливаются с наибольшим соседним;
- Фоновые сегменты сливаются с соседними, если у соседей одинаковый класс;
- Небольшие (меньше 200 пикселей) фоновые сегменты меняют класс на «Не определено»;
- Небольшие (меньше 200 пикселей) неопределенные сегменты сливаются с наибольшим соседним.

Структура ПО

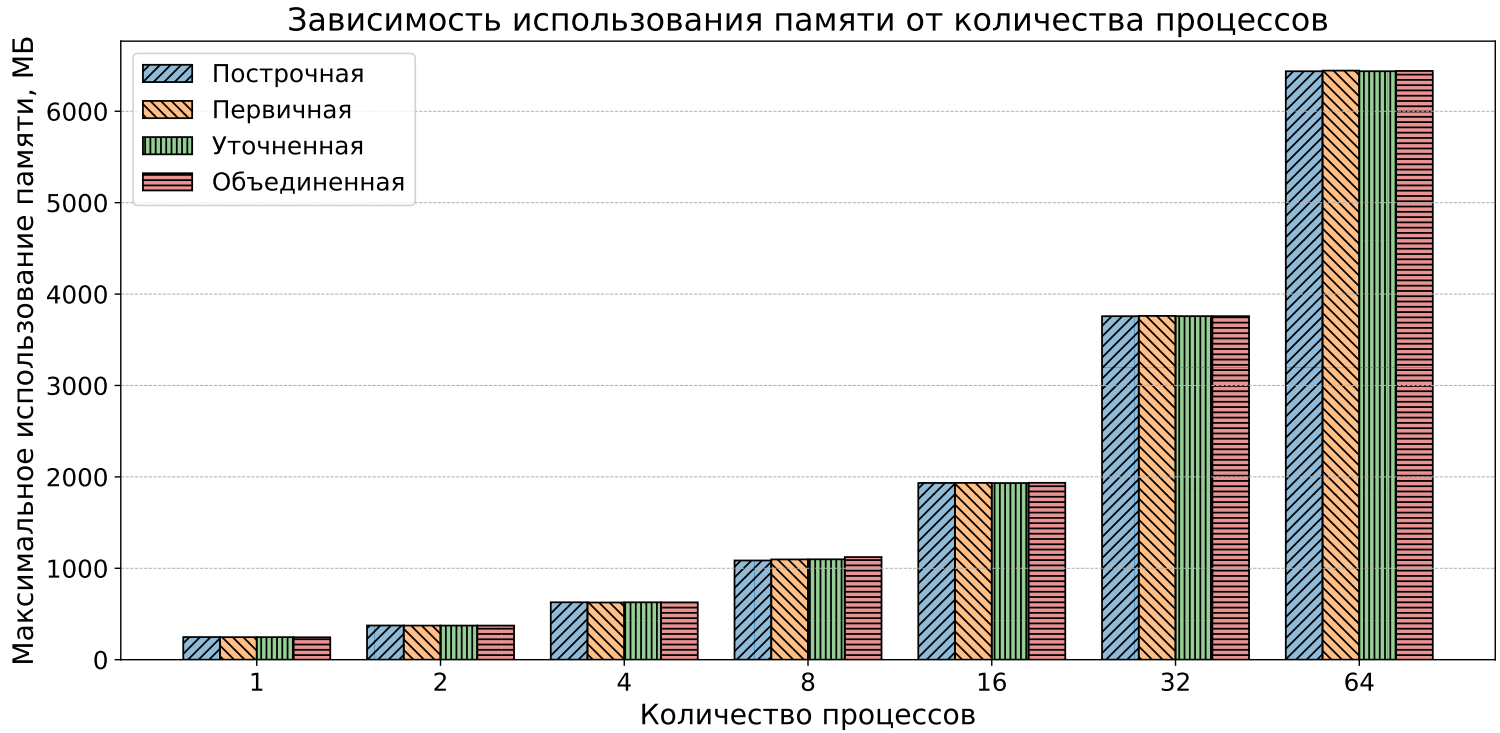


Зависимость времени разметки от количества рабочих процессов



Количество процессов	Построчн., с	Первичн., с	Уточн., с	Объед., с
1	31.80	41.21	41.29	41.31
2	16.70	21.72	21.96	22.03
4	10.16	13.05	13.11	13.19
8	7.03	8.71	8.72	8.96
16	7.29	8.83	8.96	9.23
32	7.40	9.23	9.24	9.47
64	7.82	9.67	9.73	9.79

Зависимость максимального объема используемой памяти от количества рабочих процессов



Количество процессов	Построчн., МБ	Первичн., МБ	Уточн., МБ	Объед., МБ
1	247.39	246.58	246.56	245.09
2	374.05	373.12	373.26	373.16
4	627.94	625.74	627.79	627.11
8	1085.41	1096.95	1098.18	1122.85
16	1933.34	1934.43	1932.78	1935.12
32	3757.83	3760.95	3758.53	3758.38
64	6437.75	6444.76	6438.00	6440.77

Точность и полнота качества работы

Класс \ Критерий	Верно	Л-П.	Л-Н.	Точность	Полнота
Фон	73	0	0	1.00	1.00
Текст	490	15	1	0.97	0.99
Схема	92	17	41	0.84	0.69
Рисунок	50	54	21	0.48	0.83
График	37	10	9	0.79	0.80
Таблица	90	0	12	1.00	0.88
Листинг	80	4	18	0.95	0.82
Не определено	20	0	0	1.00	1.00

Заключение

Цель достигнута: Разработан метод для автоматического выделения составных частей научного текста, использующего только простые эвристики для классификации сегментов документа.

Решены все задачи:

- Рассмотрены и сравнены существующие методы сегментации документов;
- Формализована постановка задачи;
- Разработан описанный метод;
- Разработано программное обеспечение, реализующее данный метод;
- Проведено исследование скорости разметки и максимального объема используемой памяти в зависимости от количества процессов, участвующих в разметке.

Дальнейшее развитие

- Добавление новых правил для учета дополнительных классов таблиц и рисунков в текстах;
- Поддержка классификации формул;
- Поддержка работы с двухколоночными документами.

Внедрение

- Планируется интеграция разработанного модуля разметки в систему проведения нормоконтроля МГТУ им. Н. Э. Баумана в осеннем семестре 2025 года.

Публикационная активность

- Опубликована статья «Рунов К.А. Проектирование базы данных для разметки параллельного корпуса технических текстов» в сборнике трудов XXI Международной научно-практической конференции «Инновационные, информационные и коммуникационные технологии» (РИНЦ).