НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2026 Том 26 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2026 Vol. 26 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-4-860-867
УДК 004.83

# Метод построения RAG для анализа гетерогенных графо-иерархических текстовых корпусов

**Максим Валерьевич Улизько[1]✉, Артем Дмитриевич Береснев[2], Вадим Витальевич Жуков[3], Петр Денисович Марков[4], Наталия Федоровна Гусарова[5]**

[1,2,3,5] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[4] ПАО «БАНК УРАЛСИБ», Москва, 119048, Российская Федерация

[1] ulizko@itmo.ru✉, https://orcid.org/0009-0001-2374-8025
[2] artem.beresnev@itmo.ru, https://orcid.org/0000-0002-4646-6856
[3] vvzhukov@itmo.ru, https://orcid.org/0009-0000-2082-7433
[4] pdf.markov@gmail.com, https://orcid.org/0009-0005-4176-7931
[5] natfed@list.ru, https://orcid.org/0000-0002-1361-6037

**Аннотация**

**Введение.** Анализ корпусов документов со сложными внутренними и междокументными связями требует не только поиска релевантных текстов, но и построения минимального, проверяемого и трассируемого набора фрагментов, достаточного для последующего вывода и цитирования. В работе предложен метод извлечения фрагментов для корпусов документов, в котором смысл фрагмента определяется его локальным содержанием, положением в структуре документа и системой ссылок на другие фрагменты. **Метод.** Разработанный метод основан на совместном представлении корпуса как дерева структурных единиц и ориентированного графа ссылок, а также на гибридном ранжировании, сочетающем лексический поиск, векторное сходство и ссылочный сигнал. Доказана монотонность и субмодулярность целевой функции, что позволяет применять жадные алгоритмы с известной аппроксимационной гарантией и выполнять бюджетированный отбор контекста для системы генерации с дополненным поиском (Retrieval-Augmented Generation, RAG). Дополнительно вводится протокол оценки, разделяющий качество извлечения на уровне документов, фрагментов и цитат. **Основные результаты.** Представленный метод формально валидирован на тестах устойчивости к лексическому несовпадению и эффективности бюджетного отбора по сравнению с простыми стратегиями. Для содержательной иллюстрации эффективности метода использованы примеры из юридического домена. **Обсуждение.** Метод может использоваться как слой извлечения для RAG-систем в задачах построения вопросно-ответных систем, поиска доказательств, нормативного комплаенса и анализа больших структурно связанных корпусов.

**Ключевые слова**

интеллектуальный анализ документов, граф ссылок, графо-иерархическая структура корпуса, RAG-система, бюджетированный отбор контекста, извлечение информации, поиск доказательств, субмодулярная оптимизация, юридические документы

**Благодарности**

Работа частично поддержана Министерством науки и высшего образования Российской Федерации, госзадание FSER-2025-0013.

**Ссылка для цитирования:** Улизько М.В., Береснев А.Д., Жуков В.В., Марков П.Д., Гусарова Н.Ф. Метод построения RAG для анализа гетерогенных графо-иерархических текстовых корпусов // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 4. С.860–867. doi: 10.17586/2226-1494-2026-26-4-860-867

# RAG construction method for heterogeneous graph-hierarchical text corpus analysis

**Maxim V. Ulizko[1]✉, Artem D. Beresnev[2], Vadim V. Zhukov[3], Petr D. Markov[4], Natalia F. Gusarova[5]**

[1,2,3,5] ITMO University, Saint Petersburg, 197101, Russian Federation
[4] PJSC "Bank Uralsib", Moscow, 119048, Russian Federation

[1] ulizko@itmo.ru✉, https://orcid.org/0009-0001-2374-8025
[2] artem.beresnev@itmo.ru, https://orcid.org/0000-0002-4646-6856
[3] vvzhukov@itmo.ru, https://orcid.org/0009-0000-2082-7433
[4] pdf.markov@gmail.com, https://orcid.org/0009-0005-4176-7931
[5] natfed@list.ru, https://orcid.org/0000-0002-1361-6037

**Abstract**
Analysis of document corpora with complex internal and inter-document links requires not only retrieval of relevant texts but also construction of a compact, verifiable and traceable set of fragments sufficient for downstream reasoning and citation. The purpose of this work is to propose a fragment retrieval method for document corpora in which the meaning of a fragment is determined by its local content, position in the document hierarchy and links to other fragments. The method is based on a joint representation of the corpus as a tree of structural units and a directed link graph as well as on hybrid ranking that combines lexical search, vector similarity, and a link signal. The monotonicity and submodularity of the objective function are shown, which makes it possible to use greedy algorithms with a known approximation guarantee and to perform budgeted context selection for a Retrieval-Augmented Generation (RAG) system. In addition, an evaluation protocol is introduced that separates retrieval quality at the document, fragment, and citation levels. The method is formally validated on tests of lexical mismatch robustness and budgeted selection efficiency compared with simple strategies. Examples from the legal domain are used to illustrate the method. The method can be used as a retrieval layer for RAG systems in question answering, evidence retrieval, regulatory compliance, and analysis of large structurally connected corpora.

**Keywords**
intelligent document analysis, link graph, graph-hierarchical corpus structure, RAG system, budgeted context selection, information extraction, evidence retrieval, submodular optimization, legal documents

**Acknowledgements**
This work was partly supported by the Ministry of Science and Higher Education of the Russian Federation, State Assignment FSER-2025-0013.

**For citation:** Ulizko M.V., Beresnev A.D., Zhukov V.V., Markov P.D., Gusarova N.F. RAG construction method for heterogeneous graph-hierarchical text corpus analysis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 4, pp. 860–867 (in Russian). doi: 10.17586/2226-1494-2026-26-4-860-867

## Введение

Современные системы интеллектуального анализа документов все чаще применяются не к отдельным текстам, а к наборам взаимосвязанных материалов. В таких корпусах смысл фрагмента определяется не только его локальным содержанием, но и положением в иерархии документа, а также связями с другими фрагментами. Если значимая информация распределена между несколькими разделами, простой поиск по семантической близости может возвращать фрагмент, близкий по формулировке, но недостаточный для построения обоснованного ответа [1–3].

В работе рассматривается система генерации с дополненным поиском (Retrieval-Augmented Generation, RAG) — система генерации ответа с предварительным извлечением релевантного контекста из внешнего корпуса. Ее особенность состоит в том, что языковая модель не формирует ответ только на основании внутренних параметров: перед генерацией ей передается набор найденных фрагментов, по которым можно проверить вывод и корректность цитирования [1, 4].

Особую сложность представляют гетерогенные корпуса с графо-иерархической структурой: договорные комплекты, нормативные акты, технические регламенты и комплаенс-документы. В данных корпусах важны не только близость текста к запросу, но и структурное положение пункта, его связь с определениями, приложениями и внешними нормами. Исходя из этого, традиционный отбор первых $k$ фрагментов по релевантности часто приводит к зашумлению контекста или выбору фрагмента из нерелевантного документа [5–7]. Методы подготовки контекста для RAG можно условно разделить на три группы.

Первая группа связана с сегментацией документов: вместо разбиения по фиксированному окну используются смысловая, иерархическая и отложенная сегментации. Данный подход уменьшает риск разрыва логически связанных положений, однако не всегда учитывает междокументные ссылки [8–11].

Вторая группа основана на контекстном обогащении: в каждый фрагмент добавляются сведения о родительском разделе, соседних элементах или роли фрагмента в документе. Указанный подход повышает полноту контекста, но одновременно увеличивает объем передаваемых токенов [2, 12].

Третья группа направлена на оптимизацию поиска и ранжирования: применяются вероятностная функция ранжирования Okapi BM25 (Best Matching 25), векторное сходство, позднее взаимодействие представлений, переранжирование и расширение запросов. Эти методы увеличивают качество первичного поиска, однако сами

по себе не решают задачу отбора минимального доказательного набора [1, 2, 13, 14].

Для графо-иерархических корпусов сохраняются три нерешенные проблемы. Во-первых, большинство подходов использует иерархию документа или граф связей, но не объединяет их в единую формальную модель. Во-вторых, отбор фрагментов часто остается эвристическим и не учитывает одновременно ценность фрагмента, покрытие элементов запроса и ограничение контекстного окна. В-третьих, стандартные метрики качества извлечения фиксируют факт ошибки, но не показывают, возникла ли она на уровне выбора документа, локализации фрагмента или цитирования.

Настоящая работа направлена на устранение указанных недостатков. Предлагается метод построения слоя извлечения фрагментов для RAG-систем на гетерогенных корпусах с графо-иерархической структурой. Вклад работы состоит в следующем:

— предложено совместное представление корпуса как дерева структурных единиц и ориентированного графа ссылок, что позволяет учитывать локальное содержание фрагмента, его положение в документе и топологическую значимость;
— предложен алгоритм отбора фрагментов на основе максимизации монотонной субмодулярной функции при ограничении типа рюкзака, позволяющий выбирать компактный набор доказательств при фиксированном токенном бюджете;
— разработан протокол верификации качества извлечения, разделяющий метрики на уровни документа, фрагмента и цитаты и позволяющий диагностировать источник ошибки.

## Методология построения слоя извлечения фрагментов для RAG-системы

**Способ построения слоя извлечения фрагментов.** Для графо-иерархического корпуса способ включает структурную сегментацию документов и выделение доказательных узлов; извлечение и нормализацию внутридокументных и междокументных ссылок; нормализацию запроса и извлечение целевых элементов; гибридное ранжирование кандидатов; бюджетированный отбор итогового контекста для языковой модели. Схема способа приведена на рис. 1.

Дадим формальное описание способа. Пусть задан гетерогенный корпус $C = \{d1, d2, \ldots, dn\}$, состоящий из структурированных документов и связанных с ними внешних источников. Для каждого документа $d$ строится упорядоченное дерево структуры $T_d = (N_d, E_d)$, где вершины $N_d$ соответствуют логическим единицам документа, а дуги $E_d$ задают отношение родитель–потомок. Множество кандидатов доказательных узлов $U$ определяется как подмножество вершин, пригодных для автономного использования в качестве контекста.

Запрос $q$ трактуется не как простая строка, а как набор информационных потребностей. После нормализации он представляется множеством целевых элементов $Z_q = \{z_1, z_2, \ldots, z_m\}$, где каждый элемент соответствует существенному аспекту ответа: условию, сроку, санкции, исключению, порядку уведомления, субъекту, действию или внешнему основанию решения. Такое представление позволяет перейти от плоского поиска к задаче покрытия смысловых элементов запроса.

Помимо дерева структуры, над всеми доказательными узлами строится ориентированный граф ссылок $G = (U, E_{ref})$. Ребро $(u_i, u_j)$ создается, если фрагмент $u_i$ содержит явную или нормализуемую отсылку к фрагменту $u_j$. В граф включаются внутридокументные, междокументные и внешние нормативные ссылки, а также связи с определениями и иными элементами, без которых исходный фрагмент интерпретируется неполно. Здесь $E_{ref}$ — множество ориентированных ребер ссылочного графа.

Для ранжирования кандидатов используется гибридная функция, в которой ссылочный сигнал не заменяет лексический и векторный поиск, а корректирует его в случаях, когда поверхностно похожий текст находится в неправильном документе или вне релевантной логической цепочки.

$$\text{score}(q, u) = \lambda s_{lex(q,u)} + (1 - \lambda)\cos(e_q, e_u) + k_{ref} s_{ref}(q,u), \quad (1)$$

где $s_{lex(q,u)}$ — лексическая релевантность; $\cos(e_q, e_u)$ — сходство эмбеддингов запроса и фрагмента; $s_{ref}(q,u)$ — ссылочный сигнал, усиливающий фрагменты, связанные с релевантными узлами, определениями и извлеченными элементами запроса. Параметр $\lambda \in [0, 1]$ задает соотношение лексического и векторного компонентов, а $k_{ref}$ — вес ссылочного сигнала.

**Алгоритм бюджетного отбора доказательных фрагментов.** После первичного ранжирования необходимо выбрать итоговый набор фрагментов $S$ с учетом ограничения контекста. Пусть $C(u)$ — стоимость фрагмента $u$ в токенах; $B$ — допустимый бюджет контекста. Тогда задача бюджетного отбора принимает следующий вид:

$$\max F(q, S), \quad S \subseteq U, \quad \sum_{\{u \in S\}} C(u) \leq B. \quad (2)$$

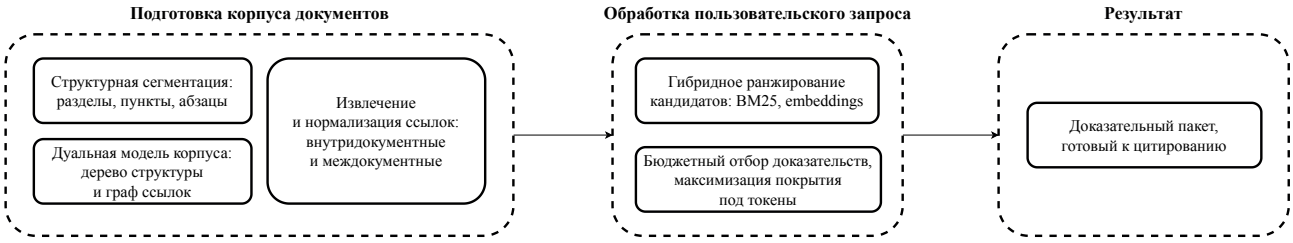


*Рис. 1.* Способ построения слоя извлечения фрагментов

*Fig. 1.* Fragment retrieval layer construction method

Целевая функция задается как сочетание полезности отдельных фрагментов и покрытия элементов запроса:

$$F(q, S) = \alpha \sum_{\{u \in S\}} \hat{s}(q,u) + (1 - \alpha)\text{Cover}(Z_q, S), \qquad (3)$$

где $\hat{s}(q, u)$ — нормализованный балл релевантности; $\text{Cover}(Z_q, S)$ — функция покрытия смысловых элементов запроса. Для ее вычисления каждый элемент $z \in Z_q$ снабжается фиксированным весом $w(z)$. В базовой шкале используются три класса важности: критические элементы получают вес 3, существенные — 2, вспомогательные — 1. Коэффициент $\alpha \in [0, 1]$ регулирует баланс между релевантностью и покрытием элементов запроса.

$$\text{Cover}(Z_q, S) = \sum_{\{z \in Z_q\}} w(z) \cdot 1[\exists u \in S : z \in E(u)].$$

При неотрицательных весах функция $F(q, S)$ является монотонной и субмодулярной: первая часть формулы (3) модульна, а вторая представляет собой взвешенную функцию покрытия [15, 16]. Следовательно, для задачи (2) применимы стандартные результаты для монотонной субмодулярной оптимизации с ограничением knapsack [15, 17].

Практически алгоритм состоит из 6 шагов: нормализовать запрос; получить пул кандидатов; вычислить $\text{score}(q, u)$; итеративно выбрать фрагмент с максимальным приростом $\Delta F/C$; обновить множество $S$ и покрытие $Z_q$; завершать отбор при исчерпании бюджета. Схема алгоритма приведена на рис. 2. Обозначение $u^*$ на рис. 2 означает конкретный фрагмент $u$, выбранный на текущей итерации как лучший кандидат по критерию максимального прироста целевой функции на единицу стоимости.

**Протокол верификации качества слоя извлечения фрагментов.** Протокол верификации построен так, чтобы различать ошибки разной гранулярности и не смешивать ошибку выбора документа с ошибкой локализации фрагмента или ошибкой цитирования. По этой причине качество слоя извлечения оценивается на трех уровнях: документном, фрагментном и цитатном.

На документном уровне используются Recall@k, MRR@k и nDCG@k. На фрагментном уровне считаются precision и recall по токенам или символам для эталонных фрагментов. На цитатном уровне оцениваются citation precision и citation recall, т. е. доля утверждений ответа, действительно поддержанных извлеченными спанами, и доля существенных утверждений, для которых такая поддержка найдена.

Дополнительно вводятся две специализированные характеристики: метрика несоответствия извлеченного документа (Document Retrieval Mismatch, DRM), фиксирующая случаи, когда верхний найденный фрагмент относится к неправильному документу, и мера избыточности (Excess Measure, EM), измеряющая избыточность выбранного набора доказательств по отношению к минимально достаточному эталонному набору.

$$DRM = \left(\frac{1}{|Q|}\right)\Sigma\{q \in Q\}1[\text{wrong}_{\text{doc(top}(q))}],$$

$$EM = \frac{(L_{used} - L_{\min})}{L_{\min}},$$

где $Q$ — множество оценочных запросов; $\text{wrong}_{\text{doc}}$ — индикатор выбора фрагмента из неправильного документа; $L_{used}$ — длина выбранного контекста; $L_{\min}$ — длина минимально достаточного эталонного контекста. Чем меньше метрика EM, тем компактнее и точнее собран контекст. Сводная схема протокола верификации приведена на рис. 3.

**Экспериментальная оценка и метрики.** Данные, использованные для экспериментальной оценки разработанного метода, представлены в табл. 1.

Экспериментальная оценка разделена на три группы: синтетические вычислительные сценарии для проверки математического поведения метода; содержательные юридические примеры для предметной иллюстрации; публичные наборы данных и корпуса как база для последующей корпусной валидации. Тесты 1 и 2 относятся к формальной проверке свойств алгоритма, тогда как примеры 1 и 2 служат содержательной иллюстрацией его применения в юридической предметной области. Такое разделение позволяет не смешивать формальную проверку алгоритма с предметной иллюстрацией и последующей оценкой на внешних данных.

В качестве основных метрик используются Recall@k, MRR@k и nDCG@k на документном уровне, precision/recall по спанам на фрагментном уровне, citation precision/citation recall на цитатном уровне, а

*Таблица 1*. Данные для экспериментальной оценки

*Table 1.* Data for experimental evaluation

| Данные | Назначение |
|---|---|
| Синтетические сценарии | Формальная валидация устойчивости к лексическому несовпадению и эффективности бюджетного отбора |
| Юридические примеры из договорного права Российской Федерации | Содержательная иллюстрация того, какие фрагменты должен собрать слой извлечения |
| Russian Law Open Data [18] | Русскоязычный нормативный слой и проверка поиска по связанным правовым документам |
| ContractNLI [19], MAUD [20], CUAD [21] | Договорные корпуса и постановки задач, ориентированные на доказательные фрагменты |
| LegalBench-RAG [5], бенчмарки надежного правового извлечения [6, 7] | Сопоставление с задачами извлечения по длинным и структурно-сложным юридическим документам |

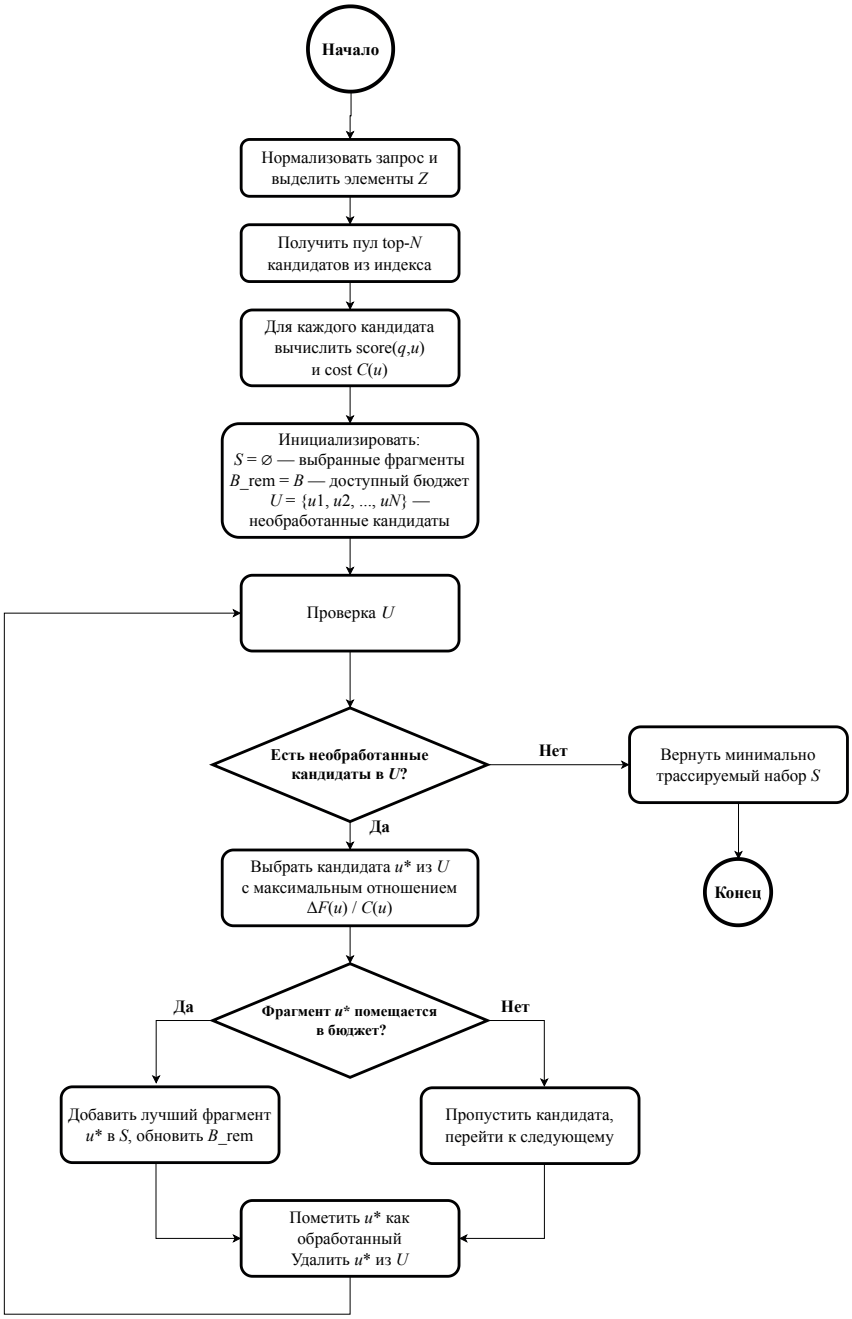


*Рис. 2.* Алгоритм бюджетного отбора доказательных фрагментов
*Fig. 2.* Budgeted selection algorithm for evidence fragments

также DRM и EM как специальные показатели ошибки неправильного документа и избыточности контекста.

## Результаты и обсуждение

**Формальная валидация.** Сначала проводится формальная валидация слоя извлечения. Задача валидации — не продемонстрировать итоговое качество на большом корпусе, а показать, что математическая постановка меняет поведение отбора в контролируемых сценариях.

**Тест 1. Устойчивость к лексическому несовпадению.** Запрос содержит условие: «письменное уведомление о просрочке в течение пяти рабочих дней». Рассматриваются два кандидата. Кандидат $u_A$ взят из другого договора: он содержит близкие слова, но не относится к проверяемому обязательству. Кандидат $u_B$ находится в нужном договоре, связан с определением уведомления и пунктом об ответственности, но использует иные формулировки. Для оценки применяются три компонента ранжирования: лексическое совпадение, векторное сходство и ссылочный сигнал. Результаты приведены в табл. 2.

По отдельному лексическому компоненту выше оценивается $u_A$. После учета векторного сходства и ссылочного сигнала итоговый балл становится выше

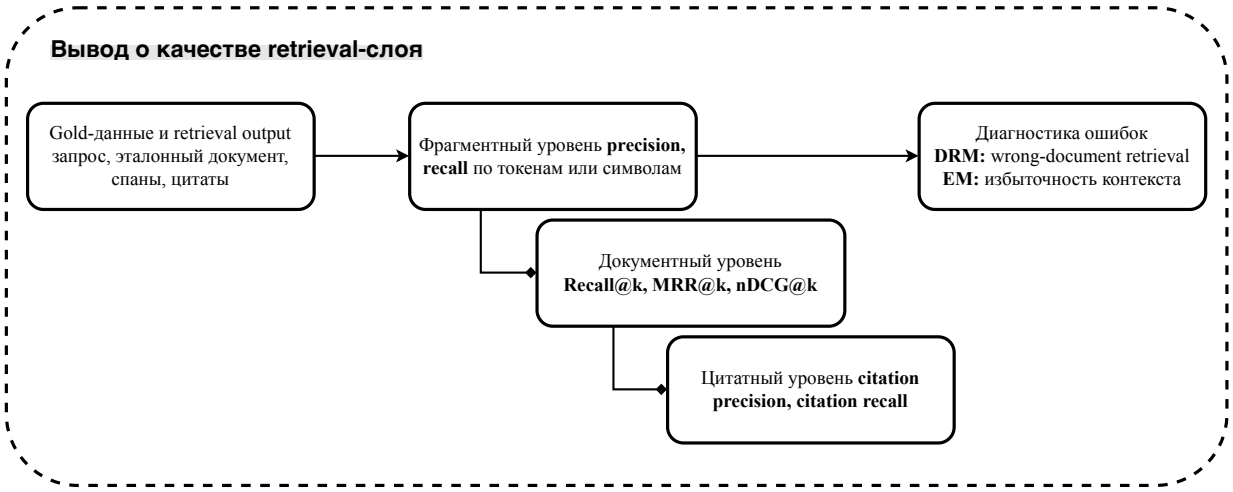


*Рис. 3.* Протокол верификации качества слоя извлечения фрагментов
*Fig. 3.* Verification protocol for fragment retrieval layer quality

*Таблица 2*. Результаты теста устойчивости к лексическому несовпадению
*Table 2.* Results of the lexical mismatch robustness test

| Кандидат | Компоненты ранжирования | | | Итоговый балл |
|---|---|---|---|---|
| | $s_{lex}$ | cos | $s_{ref}$ | |
| $u_A$ | 0,81 | 0,42 | 0,05 | 0,625 |
| $u_B$ | 0,46 | 0,74 | 0,91 | 0,782 |

*Примечание:* $s_{lex}$ — лексическое совпадение; cos — векторное сходство; $s_{ref}$ — ссылочный сигнал.

у $u_B$. Это означает, что структура документа и граф ссылок позволяют компенсировать парафраз и снизить риск выбора фрагмента из неправильного документа.

**Тест 2. Эффективность бюджетного отбора доказательств.** Множество целевых элементов запроса задано как $Z$ = {срок, санкция, уведомление, исключение}. Рассматриваются 5 кандидатов: $u_1$: $C$ = 120, $s$ = 0,86, покрытие {срок, санкция}; $u_2$: $C$ = 80, $s$ = 0,55, покрытие {уведомление}; $u_3$: $C$ = 60, $s$ = 0,48, покрытие {исключение}; $u_4$: $C$ = 90, $s$ = 0,72, покрытие {срок}; $u_5$: $C$ = 110, $s$ = 0,70, покрытие {санкция, уведомление}. Здесь $C$ — стоимость фрагмента в токенах, $s$ — его нормализованный балл релевантности. При $B$ = 260 токенов, α = 0,6 и равных весах элементов 0,25 получены результаты, приведенные в табл. 3.

Результаты табл. 3 показывают, что предлагаемый критерий отбора достигает полного покрытия существенных элементов запроса при том же бюджете и дает наибольшее значение $F(q, S)$. Это показывает, что оптимизация по смешанной целевой функции лучше согласована со структурной природой задачи, чем простое ранжирование по релевантности или длине.

**Применение в юридических сценариях.** В экспериментах необходимо было показать, как предложенный слой извлечения ведет себя в юридически осмысленных сценариях, где ответ опирается не на один текстовый фрагмент, а на связанный набор договорных и нормативных оснований.

**Пример 1. Договор поставки.** Формулировка запроса: «Какие фрагменты договора и норм права необходимо извлечь для обоснования взыскания неустойки за просрочку поставки и оценки возможных возражений поставщика?» В таком сценарии существенными элементами запроса выступают срок исполнения обязательства, факт нарушения срока, условие о неустойке, правовое основание ответственности и возможные исключения ответственности. На нормативном уровне релевантными оказываются статьи 309, 330, 401, 506 и 516 ГК РФ[1]. Следовательно, модель извлечения

[1] Гражданский кодекс Российской Федерации (часть первая) от 30 ноября 1994 г. № 51-ФЗ, ст. 309, 330, 401; Гражданский кодекс Российской Федерации (часть вторая) от 26 января 1996 г. № 14-ФЗ, ст. 506, 516.

*Таблица 3*. Сравнение стратегий бюджетного отбора
*Table 3.* Comparison of budgeted selection strategies

| Стратегия | Выбранные фрагменты | Стоимость | Покрытие | $F(q, S)$ |
|---|---|---|---|---|
| По убыванию релевантности | $\{u_1, u_5\}$ | 230 | 0,75 | 1,236 |
| По минимальной стоимости | $\{u_2, u_3, u_4\}$ | 230 | 0,75 | 1,350 |
| Предлагаемый жадный отбор | $\{u_1, u_2, u_3\}$ | 260 | 1,00 | 1,534 |

должна предпочесть набор фрагментов, одновременно покрывающий срок, нарушение, санкцию и основания освобождения от ответственности, а не просто самый лексически похожий пункт договора.

**Пример 2.** Договор возмездного оказания услуг. Формулировка запроса: «Может ли заказчик в одностороннем порядке отказаться от договора и какие фрагменты необходимы для правового обоснования такого вывода?» Здесь целевые элементы включают наличие договорного условия об отказе, субъект отказа, допустимость одностороннего прекращения обязательства, последствия отказа и порядок оформления прекращения договора. На уровне норм права значимы статьи 310, 452 и 782 ГК РФ[1]. В этой постановке модель извлечения должна собрать связный минимум доказательств, в котором специальная норма о праве заказчика соотнесена с общим запретом одностороннего отказа и с договорным условием. Тем самым ссылочная компонента модели обеспечивает не просто тематическое совпадение, а правовую связность ответа.

## Заключение

В работе предложен метод построения слоя извлечения фрагментов для анализа гетерогенных графо-иерархических текстовых корпусов. В отличие от поиска по заранее нарезанным фрагментам, предложенный подход ориентирован на выбор минимального и проверяемого набора доказательных фрагментов, достаточного для последующего вывода и цитирования. Предложен способ представления корпуса как комбинации дерева структуры документа и графа ссылок. Представлен алгоритм бюджетного отбора фрагментов, основанный на монотонной субмодулярной целевой функции. Разработан протокол оценки качества извлечения на уровне документа, фрагмента и цитаты.

Валидация на тестовых сценариях показывает, что учет структуры и ссылок улучшает порядок отбора кандидатов при лексическом несовпадении, а бюджетный отбор по смешанной целевой функции обеспечивает более полное покрытие существенных элементов запроса при том же токенном бюджете.

[1] Гражданский кодекс Российской Федерации (часть первая) от 30 ноября 1994 г. № 51-ФЗ, ст. 310, 452; Гражданский кодекс Российской Федерации (часть вторая) от 26 января 1996 г. № 14-ФЗ, ст. 782.

## Литература


1. Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y., et al. Retrieval-augmented generation for Large Language Models: a survey // arXiv. 2024. arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997
2. Barnett S., Kurniawan S., Thudumu S., Brannelly Z., Abdelrazek M. Seven failure points when engineering a retrieval augmented generation system // Proc. of the IEEE/ACM 3rd International Conference on AI Engineering — Software Engineering for AI. 2024. P. 194–199. doi: 10.1145/3644815.3644945
3. Jin B., Yoon J., Han J., Arik S.O. Long-context LLMs meet RAG: overcoming challenges for long inputs in RAG // Proc. of the International Conference on Learning Representations. 2025.
4. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks // Proc. of the 34th International Conference on Neural Information Processing Systems. 2020. P. 9459–9474.
5. Pipitone N., Alami G.H. LegalBench-RAG: a benchmark for retrieval-augmented generation in the legal domain // arXiv. 2024. arXiv:2408.10343. doi: 10.48550/arXiv.2408.10343
6. Reuter M., Lingenberg T., Liepina R., Lagioia F., Lippi M., Sartor G., et al. Towards reliable retrieval in RAG systems for large legal datasets // Proc. of the Natural Legal Language Processing Workshop. 2025. P. 17–30. doi: 10.18653/v1/2025.nllp-1.3
7. Zheng L., Guha N., Arifov J., Zhang S., Skreta M., Manning C.D., et al. A Reasoning-focused legal retrieval benchmark // Proc. of the 2025 Symposium on Computer Science and Law. 2025. P. 169–193. doi: 10.1145/3709025.3712219
8. Latif S., Ameer H., Akram M.H., Fatima M. The chunking paradigm: pecursive semantic for RAG optimization // Proc. of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025). 2025. P. 137–145.
9. Gunther M., Mohr I., Williams D.J., Wang B., Xiao H. Late chunking: contextual chunk embeddings using long-context embedding models // arXiv. 2024. arXiv:2409.04701. doi: 10.48550/arXiv.2409.04701
10. Qu R., Tu R., Bao F.S. Is semantic chunking worth the computational cost? // Findings of the Association for Computational Linguistics: NAACL. 2025. P. 2155–2177. doi: 10.18653/v1/2025.findings-naacl.114
11. Shin J., Park C., Park J., Seo J., Lim H. MultiDocFusion: hierarchical and multimodal chunking pipeline for enhanced RAG on long industrial documents // Proc. of the Conference on Empirical Methods


## References


1. Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y., et al. Retrieval-augmented generation for Large Language Models: a survey. *arXiv*, 2024. arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997
2. Barnett S., Kurniawan S., Thudumu S., Brannelly Z., Abdelrazek M. Seven failure points when engineering a retrieval augmented generation system. *Proc. of the IEEE/ACM 3rd International Conference on AI Engineering — Software Engineering for AI*, 2024, pp. 194–199. doi: 10.1145/3644815.3644945
3. Jin B., Yoon J., Han J., Arik S.O. Long-context LLMs meet RAG: overcoming challenges for long inputs in RAG. *Proc. of the International Conference on Learning Representations*, 2025.
4. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proc. of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 9459–9474.
5. Pipitone N., Alami G.H. LegalBench-RAG: a benchmark for retrieval-augmented generation in the legal domain. *arXiv*, 2024. arXiv:2408.10343. doi: 10.48550/arXiv.2408.10343
6. Reuter M., Lingenberg T., Liepina R., Lagioia F., Lippi M., Sartor G., et al. Towards reliable retrieval in RAG systems for large legal datasets. *Proc. of the Natural Legal Language Processing Workshop*, 2025, pp. 17–30. doi: 10.18653/v1/2025.nllp-1.3
7. Zheng L., Guha N., Arifov J., Zhang S., Skreta M., Manning C.D., et al. A Reasoning-focused legal retrieval benchmark. *Proc. of the 2025 Symposium on Computer Science and Law*, 2025, pp. 169–193. doi: 10.1145/3709025.3712219
8. Latif S., Ameer H., Akram M.H., Fatima M. The chunking paradigm: pecursive semantic for RAG optimization. *Proc. of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, 2025, pp. 137–145.
9. Gunther M., Mohr I., Williams D.J., Wang B., Xiao H. Late chunking: contextual chunk embeddings using long-context embedding models. *arXiv*, 2024. arXiv:2409.04701. doi: 10.48550/arXiv.2409.04701
10. Qu R., Tu R., Bao F.S. Is semantic chunking worth the computational cost? *Findings of the Association for Computational Linguistics: NAACL*, 2025, pp. 2155–2177. doi: 10.18653/v1/2025.findings-naacl.114
11. Shin J., Park C., Park J., Seo J., Lim H. MultiDocFusion: hierarchical and multimodal chunking pipeline for enhanced RAG on long industrial documents. *Proc. of the Conference on Empirical Methods*

in Natural Language Processing. 2025. P. 20985–21004. doi: 10.18653/v1/2025.emnlp-main.1062
12. Sarthi P., Abdullah S., Tuli A., Khanna S., Goldie A., Manning C.D. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval // Proc. of the International Conference on Learning Representations. 2024.
13. Robertson S., Zaragoza H. The Probabilistic relevance framework: BM25 and beyond // Foundations and Trends in Information Retrieval. 2009. V. 3. N 4. P. 333–389. doi: 10.1561/1500000019
14. Santhanam K., Khattab O., Saad-Falcon J., Potts C., Zaharia M. ColBERTv2: Effective and efficient retrieval via lightweight late interaction // Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022. P. 3715–3734. doi: 10.18653/v1/2022.naacl-main.272
15. Khuller S., Moss A., Naor J. The budgeted maximum coverage problem // Information Processing Letters. 1999. V. 70. N 1. P. 39–45. doi: 10.1016/S0020-0190(99)00031-9
16. Lin H., Bilmes J. A class of submodular functions for document summarization // Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011. P. 510–520.
17. Sviridenko M. A note on maximizing a submodular set function subject to a knapsack constraint // Operations Research Letters. 2004. V. 32. N 1. P. 41–43. doi: 10.1016/S0167-6377(03)00062-2
18. Saveliev D., Kuchakov R. The Russian legislative corpus // arXiv. 2024. arXiv:2406.04855. doi: 10.48550/arXiv.2406.04855
19. Koreeda Y., Manning C. ContractNLI: a dataset for document-level natural language inference for contracts // Findings of the Association for Computational Linguistics: EMNLP. 2021. P. 1907–1919. doi: 10.18653/v1/2021.findings-emnlp.164
20. Wang S.H., Scardigli A., Tang L., Chen W., Levkin D., Chen A., et al. MAUD: an expert-annotated legal NLP dataset for merger agreement understanding // Proc. of the Conference on Empirical Methods in Natural Language Processing. 2023. P. 16369–16382. doi: 10.18653/v1/2023.emnlp-main.1019
21. Hendrycks D., Burns C., Chen A., Ball S. CUAD: an expert-annotated NLP dataset for legal contract review // arXiv. 2021. arXiv:2103.06268. doi: 10.48550/arXiv.2103.06268

*in Natural Language Processing*, 2025, pp. 20985–21004. doi: 10.18653/v1/2025.emnlp-main.1062
12. Sarthi P., Abdullah S., Tuli A., Khanna S., Goldie A., Manning C.D. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. *Proc. of the International Conference on Learning Representations*, 2024.
13. Robertson S., Zaragoza H. The Probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009. vol. 3, no. 4, pp. 333–389. doi: 10.1561/1500000019
14. Santhanam K., Khattab O., Saad-Falcon J., Potts C., Zaharia M. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 3715–3734. doi: 10.18653/v1/2022.naacl-main.272
15. Khuller S., Moss A., Naor J. The budgeted maximum coverage problem. *Information Processing Letters*, 1999, vol. 70, no. 1, pp. 39–45. doi: 10.1016/S0020-0190(99)00031-9
16. Lin H., Bilmes J. A class of submodular functions for document summarization. *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 510–520.
17. Sviridenko M. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 2004, vol. 32, no. 1, pp. 41–43. doi: 10.1016/S0167-6377(03)00062-2
18. Saveliev D., Kuchakov R. The Russian legislative corpus. *arXiv*, 2024. arXiv:2406.04855. doi: 10.48550/arXiv.2406.04855
19. Koreeda Y., Manning C. ContractNLI: a dataset for document-level natural language inference for contracts. *Findings of the Association for Computational Linguistics: EMNLP*, 2021, pp. 1907–1919. doi: 10.18653/v1/2021.findings-emnlp.164
20. Wang S.H., Scardigli A., Tang L., Chen W., Levkin D., Chen A., et al. MAUD: an expert-annotated legal NLP dataset for merger agreement understanding. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 16369–16382. doi: 10.18653/v1/2023.emnlp-main.1019
21. Hendrycks D., Burns C., Chen A., Ball S. CUAD: an expert-annotated NLP dataset for legal contract review. *arXiv*, 2021. arXiv:2103.06268. doi: 10.48550/arXiv.2103.06268

## Авторы

**Улизько Максим Валерьевич** — преподаватель, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 60149361200, https://orcid.org/0009-0001-2374-8025, ulizko@itmo.ru
**Береснев Артем Дмитриевич** — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57202210221, https://orcid.org/0000-0002-4646-6856, artem.beresnev@itmo.ru
**Жуков Вадим Витальевич** — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0000-2082-7433, vvzhukov@itmo.ru
**Марков Петр Денисович** — главный разработчик, ПАО «БАНК УРАЛСИБ», Москва, 119048, Российская Федерация, https://orcid.org/0009-0005-4176-7931, pdf.markov@gmail.com
**Гусарова Наталия Федоровна** — кандидат технических наук, старший научный сотрудник, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57162764200, https://orcid.org/0000-0002-1361-6037, natfed@list.ru

## Authors

**Maxim V. Ulizko** — Lecturer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 60149361200, https://orcid.org/0009-0001-2374-8025, ulizko@itmo.ru
**Artem D. Beresnev** — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57202210221, https://orcid.org/0000-0002-4646-6856, artem.beresnev@itmo.ru

**Vadim V. Zhukov** — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0000-2082-7433, vvzhukov@itmo.ru
**Petr D. Markov** — Leading Developer, PJSC "Bank Uralsib", Moscow, 119048, Russian Federation, https://orcid.org/0009-0005-4176-7931, pdf.markov@gmail.com
**Natalia F. Gusarova** — PhD, Senior Researcher, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57162764200, https://orcid.org/0000-0002-1361-6037, natfed@list.ru

*Статья поступила в редакцию 21.04.2026*
*Одобрена после рецензирования 06.07.2026*
*Принята к печати 24.07.2026*

*Received 21.04.2026*
*Approved after reviewing 06.07.2026*
*Accepted 24.07.2026*