УДК 004.89
DOI:10.25729/ESI.2026.43.3.016

Иерархическая модель структуры документа для решения задачи генерации конспектов лекций

Исмагулов Милан Ерикович

Югорский государственный университет,
Россия, Ханты-Мансийск, *m_ismagulov@ugrasu.ru*

Аннотация. В статье предлагается модель описания структуры документа, основанная на методах анализа иерархической структуры документа (HDSA) и формализме теории графов. В рамках предложенного подхода документ рассматривается как иерархическая система взаимосвязанных структурных элементов, включающая заголовки различных уровней, текстовые блоки, таблицы, формулы, изображения и иные типы контента. Каждый элемент интерпретируется, как вершина графа, а отношения вложенности и логической зависимости между элементами описываются с использованием ориентированных ребер графовой структуры.

Применение предложенной модели к корпусу документов из различных предметных областей позволило выявить устойчивые закономерности организации текстового материала, характерные для конкретных типов документов. На основе полученных характеристик был сформирован набор JSON-шаблонов, описывающих типовые структуры документов для различных предметных областей, включая технические, юридические и научно-образовательные материалы.

Каждый шаблон содержит формализованное описание структуры документа, включая допустимые уровни заголовков, ограничения на типы содержимого, а также параметры стилевого и композиционного оформления. Дополнительно шаблон может включать сведения о предпочтительном формате представления данных, правилах генерации таблиц, формул и специализированных текстовых конструкций. Использование JSON-формата обеспечивает машинно-ориентированное представление структуры документа и возможность интеграции шаблонов в программные системы автоматической генерации текста.

Разработанные шаблоны предназначены для использования в процессе генерации текстов большими языковыми моделями (Large Language Models – LLM). В рамках предложенного подхода шаблон выступает как формализованное структурное ограничение, определяющее допустимую организацию итогового документа и последовательность формирования его разделов. Это позволяет снизить вариативность структуры генерируемого текста, повысить согласованность взаимосвязанных разделов и обеспечить соответствие результата требованиям конкретной предметной области.

Предложенный подход ориентирован на реализацию управляемой генерации документов, при которой большая языковая модель выполняет генерацию содержимого в рамках заранее заданной структурной схемы. За счет этого обеспечивается более стабильное качество формирования документа, уменьшается вероятность нарушения логики построения текста и повышается воспроизводимость результатов генерации. Кроме того, использование структурных шаблонов позволяет адаптировать процесс генерации под различные типы документов без изменения архитектуры самой языковой модели.

Ключевые слова: иерархическая структура документа, теория графов, JSON-шаблоны, большие языковые модели (LLM), структурированная генерация текста, структурные паттерны, модель представления документа

Цитирование: Исмагулов М.Е. Иерархическая модель структуры документа для решения задачи генерации конспектов лекций / М Е. Исмагулов // Информационные и математические технологии в науке и управлении, 2026. – № 3(43). – С. 197-210. – DOI:10.25729/ESI.2026.43.3.016.

Введение. В современных задачах генерации текста всё большую роль играет формирование структурированного вывода, согласованного с требованиями конкретной предметной области. Вместе с тем большинство существующих подходов опираются на универсальные шаблоны структуры документа и недостаточно учитывают реальные иерархические закономерности текстов. В настоящем исследовании предлагается модель оценки и использования иерархической структуры документа на основе метода HDSA и теории графов, позволяющая формировать JSON-шаблоны и обеспечивать контролируемую генерацию текста большими языковыми моделями [1, 2, 3].

1. Метод анализа иерархической структуры документа. Метод HDSA **(**Hierarchical Document Structure Analysis – иерархический анализ структуры документов) был разработан для решения задачи структурного разделения содержимого различных текстовых документов. Его основная цель заключается в выделении семантических блоков текста (заголовков, разделов, абзацев, параграфов с формулами и др.) с последующим построением иерархии и установлением связей между ними [4]. Наиболее простой и распространённый тип связи – отношение «родитель → ребёнок», отражающее вложенность блоков и обеспечивающее представление документа в виде дерева.

Формально документ можно представить, как дерево:

$$T = (V, E, r)$$

где V – множество вершин, соответствующих структурным блокам, $E \subseteq V \times V$ – множество рёбер, отражающих иерархические связи, а $r \in V$ – корень, интерпретируемый, как вся страница документа. Для каждой вершины $v \in V$ задаётся её тип $\tau(v)$ (заголовок, абзац, колонтитул и др.) и глубина $d(v)$, равная расстоянию до корня.

1.1. Связь между методами анализа иерархической структуры документа и анализом макета документа. Метод HDSA близок к Document Layout Analysis (DLA), однако между ними существует принципиальное различие. DLA направлен на анализ пространственного расположения текстовых блоков на странице и применяется преимущественно для восстановления визуальной структуры документа, сегментации и макетирования [5, 6]. HDSA же рассматривает документы с позиции логико-семантических связей между структурными единицами текста. Таким образом, DLA полезен для задач оптического распознавания символов и реконструкции макета [7], а HDSA – для анализа логической организации содержания документа и построения моделей его структурного представления [8, 9].

Визуальное подтверждение иерархической структуры документа обычно представлено в оглавлении. Содержание является линейным отображением HDSA-дерева: каждая вершина иерархии (глава, раздел, подраздел и т.д.) представлена в виде строки с соответствующим уровнем вложенности. Уровень определяется средствами визуального форматирования – нумерацией, отступами или шрифтовыми акцентами [10].

1.2. Основная идея анализа. Анализ страницы документа начинается с её разбиения на текстовые блоки с присвоением каждому из них уникального индекса. Страница интерпретируется, как корень дерева, содержащий как содержательные, так и служебные элементы (колонтитулы, номера страниц, сноски), которые образуют отдельные ветви. Далее формируется иерархия содержательных блоков – разделов, подразделов, абзацев и специализированных элементов, в результате чего документ представляется в виде дерева, отражающего его логико-семантическую структуру (рисунок 1).

Анализ иерархической структуры документа начинается с визуальной сегментации страницы на текстовые блоки по пробелам между ними. Далее строится иерархия блоков: для одно колоночных документов – линейно сверху вниз, для многоколоночных – с предварительным определением направления чтения. Для упорядочивания элементов задаётся весовое правило, согласно которому заголовки обладают наибольшим иерархическим приоритетом по отношению к остальным типам блоков. Особенностью анализа, представленного в данном исследовании, заключается представление HDSA в табличной форме, обеспечивающей формализованное представление HDSA-структуры (таблица 1).

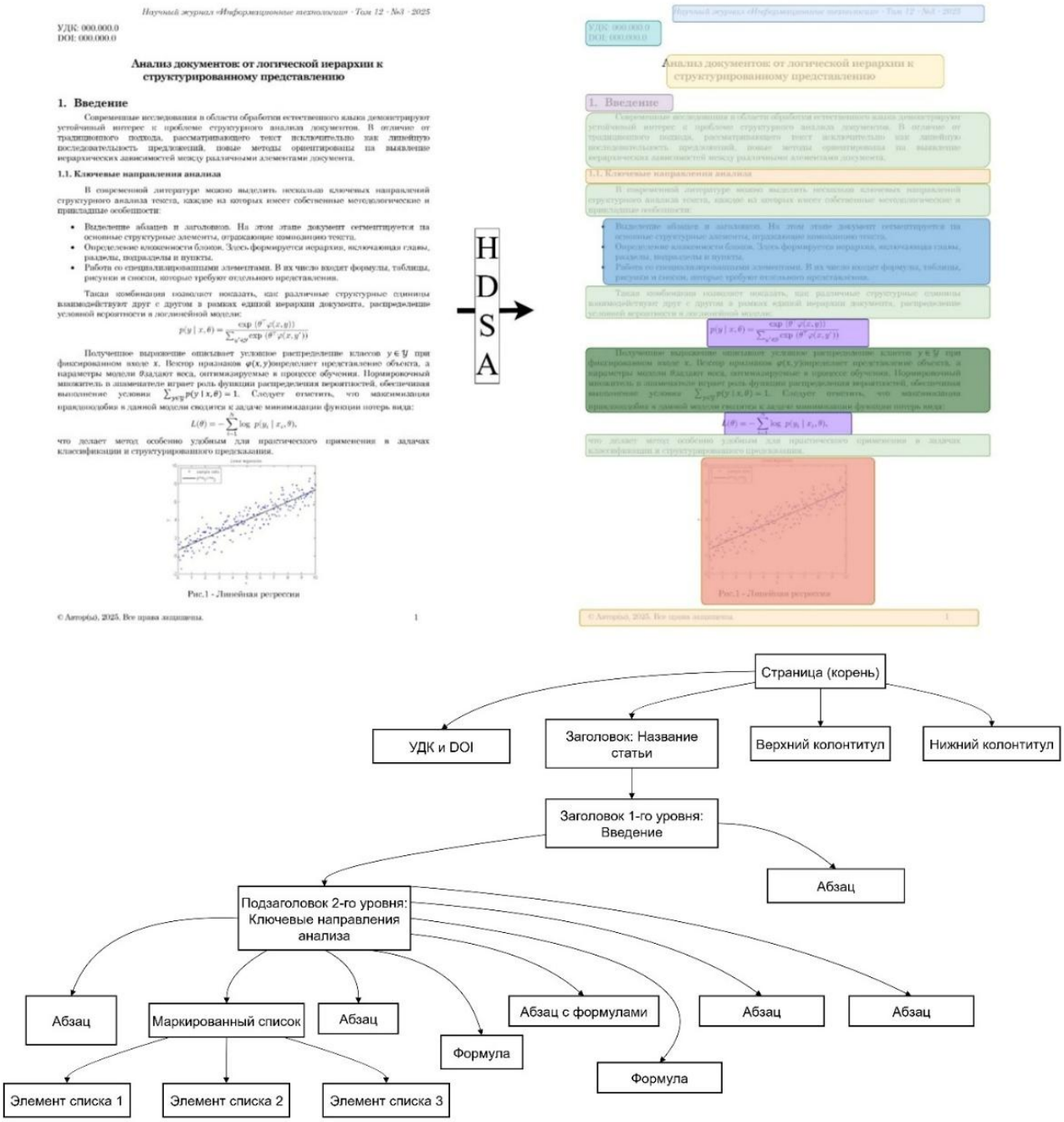

Рис. 1. Пример разбиения документа на блоки методом HDSA и построение дерева

Таблица 1. Пример HDSA таблицы анализа документа

Цвет элемента на схеме	Название элемента	Уровень вложенности	Горизонтальный вес (иерархический приоритет)	Родитель → Дочерние элементы
#e5eefe	Верхний колонтитул	0	0	Родитель (Корень) – Дочерний элемент отсутствует
#c3eaef	УДК и DOI	1	0	Родитель (Корень) – Дочерний элемент отсутствует
#fff6d9	Заголовок документа	2	0	Родитель (Корень) – Дочерний элемент заголовок 1-го уровня

Цвет элемента на схеме	Название элемента	Уровень вложенности	Горизонтальный вес (иерархический приоритет)	Родитель → Дочерние элементы
#e7dfee	Заголовок 1-го уровня	3	1	Родитель (Заголовок документа) – Дочерние элементы абзац и заголовок 2-го уровня
#e0eedf	Абзац	4	2	Родитель (Заголовок 1-го уровня) – Дочерний элемент отсутствует
#feedd9	Заголовок 2-го уровня	5	2	Родитель (Заголовок 1-го уровня) – Дочерний элемент отсутствует
#e0eedf	Абзац	6	3	Родитель (Заголовок 2-го уровня) – Дочерний элемент отсутствует
#9cc7e9	Маркированный список	7	3	Родитель (Заголовок 2-го уровня) – Дочерние элементы – элементы списка 1,2 и 3
#e0eedf	Абзац	11	3	Родитель (Заголовок 2-го уровня) – Дочерний элемент отсутствует
#d0b6ff	Формула блочная	12	3	Родитель (Заголовок 2-го уровня) – Дочерний элемент отсутствует
#84a882	Абзац со строчными формулами	13	3	Родитель (Заголовок 2-го уровня) – Дочерний элемент может быть представлен строчными формулами при дальнейшей декомпозиции
#d0b6ff	Формула блочная	14	3	Родитель (Заголовок 2-го уровня) – Дочерний элемент отсутствует
#e0eedf	Абзац	15	3	Родитель (Заголовок 2-го уровня) – Дочерний элемент отсутствует
#f4b4a8	Изображение с подписью	17	3	Родитель (Заголовок 2-го уровня) – Дочерний элемент отсутствует
#ffeedc	Нижний колонтитул	18	0	Родитель (Корень) – Дочерний элемент нумерация станицы

Подобное табличное представление удобно масштабируется на более сложные документы с большим числом элементов и связей. Каждому типу блока можно присвоить уникальный индекс, что обеспечивает компактное и формализованное описание структуры документа.

1.3. Предлагаемая модель. В отличие от традиционного применения метода HDSA для описания структуры, в настоящей работе он используется как основа для построения модели

оценки реальных документов различных предметных областей. На базе формализма теории графов вводятся количественные характеристики – глубина иерархии, показатели информативности блоков, меры посредничества и матрицы смежности. Эти характеристики позволяют выявить устойчивые структурные закономерности и использовать их для синтеза формализованных JSON-шаблонов [11, 12], интегрируемых в процессы генерации текста большими языковыми моделями [13, 14].

В настоящем исследовании модель, построенная по методу Hierarchical Document Structure Analysis (HDSA), основывается на ряде мер, заимствованных из теории графов:

1. Средняя глубина иерархии графа. Показатель глубины иерархии используется для количественной оценки степени вложенности структурных блоков текста и характеризует общую иерархическую сложность структуры документа. Средняя глубина графа не тождественна максимальному уровню заголовка, однако при отсутствии аномально глубоких ветвей может использоваться как устойчивая оценка эффективной глубины структуры документа.

Неясно, каким образом определяется принадлежность блока контента к той или иной макрогруппе при расчёте меры разнообразия блоков документа.

Средняя глубина иерархии для всего графа $G = (V, E)$ определяется формулой:

$$H(G) = \frac{1}{|V|} \sum_{v \in V} h(v)$$

где $h(v)$ – глубина вершины v, определяемая числом уровней вложенности до корневого элемента. $|V|$ – это мощность множества V, т. е. количество элементов этого множества.

2. Мера разнообразия блоков документа. Для описания общего состава документа D вводится вектор V_c присутствия типов контента в документе по макрогруппам:

$$V_c(D) = (n_{\text{math}}, n_{\text{struct}}, n_{\text{tabular}}, n_{\text{teorethic}}, n_{\text{visual}})$$

где каждый компонент n_i отражает наличие элементов соответствующего типа, присутствующих в документе. Принадлежность блоков к определенной макрогруппе определяется за счет семантики языка разметки, в данном исследовании языком разметки выступает LaTeX.

Для получения числовой меры разнообразия контента $M_{div}(D)$ вводится следующая функция:

$$M_{div}(D) = \frac{\textit{число присутствующих макрогрупп в документе}}{\textit{максимальное количество макрогрупп в векторе}}$$

- $M_{div}(D) \approx 0$ – документ однотипный (например, полностью текстовый);
- $M_{div}(D) \approx 1$ – документ содержит сбалансированное количество элементов разных типов, т.е. является разнообразным.

Согласно этой мере, документы условно можно разделить на 4 класса, представленные в таблице 2.

3. Мера объема текста в документе.

Для каждой вершины $v \in V$ определяются:

$$M_{total}(v) = \textit{общее количество токенов в элементе}$$

$$M_{text}(v) = \textit{количество текстовых токенов в элементе}$$

На их основе вводится отношение:

$$R_{text}(v) = \frac{M_{text}(v)}{M_{total}(v)},$$

которое характеризует долю текстовой информации в структуре конкретного элемента. Среднее значение этого отношения по всем вершинам может рассматриваться как показатель текстоцентричности документа:

$$\overline{R}_{text}(D) = \frac{1}{|V|}\sum_{v \in V} R_{text}(v)$$

Таблица 2. Примеры классов документа согласно мере разнообразия

Диапазон $M_{div}(D)$	Интерпретация
0.0 – 0.33	Низкое разнообразие
0.34 – 0.66	Среднее разнообразие
0.67 – 1	Высокое разнообразие

4. Мера центральности посредничества узлов (betweenness centrality). Эта мера необходима для оценки влияния отдельных типов вершин (заголовков, абзацев, формул и т.д.) на связанность графа документа. Она характеризует степень участия вершины v в кратчайших путях между другими вершинами и вычисляется по формуле:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

где σ_{st} – количество кратчайших путей между вершинами s и t, а $\sigma_{st}(v)$ – число таких путей, проходящих через вершину v.

В терминах анализа структуры документа высокая центральность посредничества означает, что элемент играет роль связующего звена между различными частями текста (например, раздел, объединяющий несколько подтем). Именно таким образом в рамках одного документа численно определяются элементы с высокой степенью посредничества, такие, как заголовки различных уровней, от элементов с низкой степенью посредничества, например, абзацы. В дополнение к основным количественным мерам оценки структуры документа рассматриваются вспомогательные показатели, в частности, показатель количества структурных блоков позволит оценить численно элементы графа.

2. Суть эксперимента. Экспериментальная часть исследования была направлена на проверку применимости модели и метода HDSA для автоматического извлечения структурных метрик документов и формирования на их основе формализованных JSON-шаблонов [15, 16]. Эксперименты проводились на корпусах объёмом 50 страниц для трёх предметных областей (математика, юриспруденция, экология), подготовленных в формате LaTeX, что позволило оценить устойчивость подхода при работе с текстами различной структурной организации. Выбор данных областей обеспечил анализ как универсальных, так и предметно-специфических структурных паттернов (рисунок 2).

В рамках исследования был сформирован специализированный корпус документов, содержащий 5 документов для каждой предметной области, участвующей в экспериментальном анализе.

Список документов по математике:

- Письменный Д. Т. «Конспект лекций по высшей математике»;
- Краснов М. Л., Киселёв А. И., Макаренко Г. И. и др. «Вся высшая математика»;
- Дейзенрот М. П., Фейзал А. А., Он Ч. С. «Математика для машинного обучения»;

- Демидович Б. П., Кудрявцев В. А. «Краткий курс высшей математики»;
- Будак Б. М., Фомин С. В. «Кратные интегралы и ряды».

Рис. 2. Алгоритм проведения эксперимента

Список документов по юриспруденции:

- Мамедова С.Н. «Курс лекций по дисциплине правоведение»;
- Мелехин А.В. «Теория государства и права: Учебник с учебно-методическими материалами»;
- Исакова В. Б. «Основы права: учебник для неюридических вузов и факультетов»;
- Клишаса А.А. «Теория государства и права: учебник – Российский университет дружбы народов, Юридический институт»;
- Сулейменов М. «Право как система».

Список документов по экологии:

- Вороного А.А., Ситникова С.В. «Экология конспект лекций»;
- Гашев С. Н., «Конспекты лекций по системной экологии: учебное пособие»;
- Курмаева Н. М., Смирнов Д.Г. «Краткий курс экологии»;
- Панин В.Ф., Сечин А.И., Федосова В.Д. «Экология общеэкологическая концепция биосферы и экономические рычаги преодоления глобального экологического кризиса; обзор современных принципов и методов защиты биосферы»;
- Криксунов Е.А., Пасечник В.В. «Экология 10, 11 класс».

Результаты оценки каждого документа, полученные с применением предложенной модели, представлены в таблице 3.

2.1. Оценка документов из предметной области. На этапе предварительной визуальной оценки было выдвинуто предположение, что наибольшей структурной и содержательной

сложностью будут характеризоваться материалы по экологии вследствие их междисциплинарного характера и сочетания различных типов контента. Высокая структурная сложность также ожидается для математических материалов из-за частой смены текстовых и формализованных блоков. В то же время документы по юриспруденции предполагаются наиболее однотипными и текстоцентричными, что обусловлено преобладанием нормативного изложения.

Таблица 3. Результаты оценки документов

Предметная область	$H(G)$	$M_{div}(D)$	$R_{text}(D)$	Элементы с наибольшим $C_B(v)$	Количество блоков в графе
Математика (Письменный)	2,75	0,83	0,58	Заголовки 1 и 2 уровня	617
Математика (Краснов)	2,81	0,83	0,62	Заголовки 1 и 2 уровня	751
Математика (Дейзенрот)	3,1	0,83	0,7	Заголовки 1 и 2 уровня	433
Математика (Демидович)	2	0,83	0,51	Заголовки 1 и 2 уровня	303
Математика (Будак)	2,75	0,83	0,66	Заголовки 1 и 2 уровня	804
Юриспруденция (Мамедова)	2,81	0,33	1	Заголовки 1 и 2 уровня	239
Юриспруденция (Мелехин)	2,9	0,33	1	Заголовки 1 и 2 уровня	283
Юриспруденция (Клишаса)	2,97	0,33	1	Заголовки 1 и 2 уровня	237
Юриспруденция (Сулейменов)	2,55	0,33	1	Заголовки 1 и 2 уровня	172
Юриспруденция (Исакова)	2,67	0,33	1	Заголовки 1 и 2 уровня	258
Экология 1 (Вороной)	2,92	0,67	0,69	Заголовки 1 и 2 уровня	202
Экология 2 (Панин)	2,65	0,67	0,56	Заголовки 1 и 2 уровня	193
Экология 3 (Криксунов)	3,01	0,67	0,67	Заголовки 1 и 2 уровня	221
Экология 4 (Гашев)	2,89	0,67	0,72	Заголовки 1 и 2 уровня	284
Экология 5 (Курмаева)	3,41	0,67	0,84	Заголовки 1 и 2 уровня	306

Анализ данных таблицы 3 показывает, что для документов всех трёх предметных областей средняя глубина иерархического графа стремится к значению около 3, что согласуется с визуальной оценкой структуры и используемыми системами заголовков. Для всех документов также характерны высокие значения центральности узлов, соответствующих заголовкам. При этом наибольшая структурная и содержательная вариативность наблюдается в материалах по экологии и математике, что отражается в значениях показателя разнообразия и количестве блоков в графе. Документы по юриспруденции, напротив, отличаются выраженной текстоцентричностью и более однородной структурой.

2.3. Формирование шаблона на основе мер. В ходе эксперимента были сформулированы правила формирования JSON-шаблона, используемого при генерации структурированного документа. Шаблон включает функциональные блоки metrics, structure и style, а также параметр выходного формата output_format. Блок metrics содержит эталонные значения метрик и их интерпретации, structure задаёт иерархию заголовков и допустимые типы контентных блоков, а style определяет требования к стилю и профессиональному языку результирующего документа.

Листинг 1. Пример сокращенного шаблона, описывающий правила формирования на языке JSON

```
{"metrics": {
    "H(G)": {
      "value": 3.0,
      "interpretation": "Эффективная глубина иерархии документа"},
    "M_div": {
      "value": 0.67,
      "interpretation": "Разнообразие блоков контента"}},
    "R_text": {
      "value": 0.69,
      "interpretation": "Доля текстоцентричного контента"}},
  "structure": {
    "headings": [
      {"level": 1, "centrality": "high"},
      {"level": 2, "centrality": "medium"},
      {"level": 3, "centrality": "low"}],
...
    "blocks": {
      "text": "text text text",
      "inline_equation": "text $inline_equation$ text",
      "block_equation": "$$block_equation$$",
      "list": "- item"}},
  "style": {
    "domain_language": "professional",
    "tone": "academic",
    "verbosity": "moderate"
  },"output_format": "markdown"}
```

Анализ сформированных шаблонов показал, что их базовая структурная схема не зависит от предметной области и остаётся инвариантной. Различия между шаблонами проявляются не на уровне иерархии, а на уровне состава и разнообразия используемых контентных блоков. Для материалов по математике и экологии характерно расширение набора блоков и увеличение доли нетекстовых элементов, тогда как шаблоны для юриспруденции сохраняют преимущественно текстоцентричную структуру.

2.4. Результаты и обсуждение. Для обеспечения статистической устойчивости эксперимента на первом этапе было выполнено 10 генераций без использования шаблона, что позволило зафиксировать базовый уровень вариативности и структуры формируемого текста. На втором этапе были проведены 10 генераций с применением шаблона. В качестве большой языковой модели используется модель Mistral Large через API со стандартным значением температуры равное 0,3. Сводные результаты обеих серий экспериментов представлены в таблице 4, на которую далее даётся ссылка при анализе полученных различий.

Анализ результатов, представленных в таблице 4, показывает, что использование шаблона не оказывает влияния на показатели $M_{div}(D)$ и $R_{text}(D)$, что обусловлено спецификой предметной области (юриспруденция), характеризующейся доминированием однородных текстовых блоков и отсутствием выраженного структурного разнообразия. В то же время

применение шаблона приводит к снижению вариативности графовой структуры документа, что выражается в стабилизации значения $H(G)$ и уменьшении разброса количества блоков.

Таблица 4. Результат генерации для конспекта по праву

$H(G)$	$M_{div}(D)$	$R_{text}(D)$	Количество блоков в графе
Эталонные значения			
2,8	0,33	1	В рамках эксперимента не используется
10 генераций без шаблона			
2,63	0,33	1	46
2,31	0,33	1	36
2,13	0,33	1	23
2,2	0,33	1	40
2,86	0,33	1	46
1,98	0,33	1	51
2,06	0,33	1	33
3,27	0,33	1	52
3,27	0,33	1	45
2,2	0,33	1	40
2,13	0,33	1	23
10 генераций с использованием шаблона			
3,07	0,33	1	28
3,04	0,33	1	21
3,0	0,167	1	20
3,08	0,33	1	23
3,0	0,33	1	20
3,0	0,167	1	20
3,06	0,33	1	33
2,7	0,33	1	32
3,09	0,33	1	22
3,03	0,33	1	21

В отличие от результатов, полученных для лекций по праву, эксперименты по математике (таблица 5) демонстрируют выраженное влияние шаблона на формирование структуры документа. Аналогично предыдущему эксперименту, применение шаблона приводит к снижению структурной вариативности, стабилизации значения $H(G)$ и сокращению разброса числа блоков, обеспечивая более компактную и воспроизводимую организацию документа. В генерациях без шаблона наблюдается значительный разброс характеристик и вырожденные случаи с утратой математической модальности. Дополнительно выявлена устойчивая тенденция к избыточному дроблению содержания на структурные блоки, что указывает на повышенную фрагментацию генерируемого документа.

Таблица 5. Результат генерации для конспекта по математике

$H(G)$	$M_{div}(D)$	$R_{text}(D)$	Количество блоков в графе
Эталонные значения			
2,75	0,83	0,58	В рамках эксперимента не используется
10 генераций без шаблона			
3,51	0,55	0,81	75
2,47	0,33	1	15
3,99	0,55	0,79	91
2	0,167	1	4
3,51	0,55	0,81	75
2,89	0,55	0,62	21
3,77	0,55	0,93	70
3,48	0,55	0,79	75
3,85	0,55	0,88	89
10 генераций с использованием шаблона			
3,51	0,67	0,76	52
3,404	0,55	0,77	52
3,36	0,55	0,86	58
3,39	0,67	0,89	57
3,45	0,55	0,89	49
2,84	0,55	0,91	51
2,75	0,67	0,73	41
3,67	0,55	0,65	54
3,68	0,55	0,67	54
3,63	0,55	0,69	35

Результаты экспериментов для экологической предметной области подтверждают тенденцию, ранее выявленную для математики: независимо от использования шаблона модель склонна к избыточной фрагментации документа, что проявляется в увеличении числа блоков и значений $H(G)$. При этом показатели текстоцентричности остаются близкими к единице, отражая текстоцентричный характер экологического материала и обусловливая низкие значения показателя разнообразия. Использование шаблона снижает структурную вариативность и устраняет вырожденные случаи, однако не устраняет саму тенденцию к фрагментации, что указывает на систематическую особенность генеративного механизма модели (таблица 6).

Заключение. В работе предложена и апробирована модель анализа и использования иерархической структуры документа на основе метода HDSA и теории графов, позволяющая формировать формализованные JSON-шаблоны. Эксперименты на материалах по юриспруденции, математике и экологии показали, что применение шаблонов снижает структурную вариативность генерации и стабилизирует графовую организацию документа без потери полноты содержания. Одновременно выявлена систематическая тенденция к избыточной фрагментации без использования шаблона в структурно насыщенных предметных областях, что определяет направления дальнейшей доработки шаблонов и метода.

Таблица 6. Результат генерации для конспекта по экологии

$H(G)$	$M_{div}(D)$	$R_{text}(D)$	Количество блоков в графе
Эталонные значения			
2,92	0,67	0,69	В рамках эксперимента не используется
10 генераций с использованием шаблона			
3,16	0,33	1	69
3,11	0,33	1	63
2,46	0,167	1	24
2,43	0,167	1	30
2,41	0,167	1	22
2,7	0,167	1	30
3,36	0,55	0,99	124
3,25	0,33	1	87
3,13	0,33	1	71
3,08	0,33	1	72
10 генераций без шаблона			
3,64	0,55	1	45
2,58	0,55	1	35
3,09	0,55	0,98	63
3,12	0,33	1	67
3,13	0,33	1	46
2,71	0,55	1	41
3,06	0,33	1	53
2,88	0,33	1	32
3,12	0,33	1	49
2,98	0,33	1	54

Список источников

1. Liang X., Wang H., Wang Y. et al. Controllable text generation for large language models: a survey. Preprint arXiv:2408.12599, 2024, DOI: 10.48550/arXiv.2408.12599.
2. Liu Y., Li D., Wang K. et al. Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. Information Processing & Management, 2024, vol. 61, no. 5, art. no. 103809, DOI: 10.1016/j.ipm.2024.103809.
3. Lu Y., Li H., Cong X. et al. Learning to generate structured output with schema reinforcement learning. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 4905–4918.
4. Xing H., Cheng C., Gao F. et al. DocHieNet: a large and diverse dataset for document hierarchy parsing. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds.), 2024, pp. 1129–1142, DOI:10.18653/v1/2024.emnlp-main.65.
5. Pfitzmann B., Auer C., Dolfi M. et al. DocLayNet: a large human-annotated dataset for document-layout analysis. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), 2022, pp. 3530–3540, DOI: 10.1145/3534678.3539043.
6. Wang J., Krumdick M., Tong B. et al. A graphical approach to document layout analysis. Preprint arXiv:2308.02051, 2023, DOI: 10.48550/arXiv.2308.02051.
7. Jiovannini S., Marinai S. A survey on reading order, table of contents, and structure extraction in document analysis. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2025, pp. 7585–7594.

8. Rausch J., Martinez O., Bissig F. et al. DocParser: hierarchical document structure parsing from renderings. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, no. 5, pp. 4328–4338, DOI: 10.1609/aaai.v35i5.16558.
9. Wang J., Hu K., Zhong Z. et al. Detect-order-construct: a tree construction-based approach for hierarchical document structure analysis. Pattern Recognition, 2024, vol. 156, art. no. 110836, DOI: 10.1016/j.patcog.2024.110836.
10. Yang J., Jiang D., He L. et al. StructEval: benchmarking LLMs' capabilities to generate structural outputs. Preprint arXiv:2505.20139, 2026, DOI: 10.48550/arXiv.2505.20139
11. Chen Y., Liu R., Zheng J. et al. Graph-based document structure analysis. Preprint arXiv:2502.02501, 2025, DOI: 10.48550/arXiv.2502.02501
12. Kim D. Guided JSON with LLMs: from raw PDFs to structured intelligence. Available at: https://medium.com/@kimdoil1211/structured-output-with-guided-json-a-practical-guide-for-llm-developers-6577b2eee98a (accessed: 02/03/2025).
13. Geng S., Cooper H., Moskal M. et al. Generating structured outputs from language models: benchmark and studies. Preprint arXiv:2501.10868, 2025, DOI: 10.48550/arXiv.2501.10868.
14. Joshi I., Agarwal B., Rojkova V. Think inside the JSON: reinforcement strategy for strict LLM schema adherence. Preprint arXiv:2502.14905, 2025, DOI:10.48550/arXiv.2502.14905.
15. Balaskas G., Papadopoulos H., Pappa D. et al. A framework for domain-specific dataset creation and adaptation of large language models. Computers, 2025, vol. 14, no. 5, art. no. 172, DOI: 10.3390/computers14050172.
16. Taveekitworachai P., Dewantoro M.F., Xia Y. et al. BenchING: a benchmark for evaluating large language models in following structured output format instruction in text-based narrative game tasks. IEEE Transactions on Games, 2025, vol. 17, no. 3, pp. 665–675, DOI: 10.1109/TG.2025.3529117.

Исмагулов Милан Ерикович. *Аспирант по направлению 2.3.1. – Системный анализ, управление и обработка информации, статистика. Основные направления исследований автора: мультимодальное преобразование данных, машинное обучение, методы обработки данных моделями машинного обучения. AuthorID: 1301221, SPIN: 6808-8045, ORCID: 0009-0007-3280-5259, m_ismagulov@ugrasu.ru. 628012, Ханты-Мансийский автономный округ - Югра, г. Ханты-Мансийск, ул. Чехова, д.16.*

UDC 004.89

DOI:10.25729/ESI.2026.43.3.016

A Hierarchical model of document structure for the task of generating lecture summaries

Milan E. Ismagulov

Yugra State University, Russia, Khanty-Mansiysk, *m_ismagulov@ugrasu.ru*

Abstract. This article proposes a document structure description model based on Hierarchical Document Structure Analysis (HDSA) methods and the formalism of graph theory. Within the proposed approach, a document is considered as a hierarchical system of interconnected structural elements, including headings of different levels, text blocks, tables, formulas, images, and other types of content. Each element is interpreted as a graph vertex, while nesting and logical dependency relationships between elements are described using directed edges of the graph structure.

The application of the proposed model to a corpus of documents from various subject areas made it possible to identify stable patterns in the organization of textual material characteristic of specific document types. Based on the obtained characteristics, a set of JSON templates was developed to describe typical document structures for different subject areas, including technical, legal, and scientific and educational materials.

Each template contains a formalized description of the document structure, including permissible heading levels, constraints on content types, and parameters of stylistic and compositional formatting. Additionally, a template may include information on the preferred data presentation format, rules for generating tables and formulas, and specialized textual constructions. The use of the JSON format provides a machine-oriented representation of the document structure and enables the integration of templates into software systems for automatic text generation.

The developed templates are intended for use in text generation by large language models (LLMs). Within the proposed approach, a template serves as a formalized structural constraint that defines the permissible organization of the resulting document and the sequence in which its sections are generated. This makes it possible to reduce the structural variability of generated text, improve the consistency of interrelated sections, and ensure that the resulting document complies with the requirements of a specific subject area.

The proposed approach is aimed at implementing controlled document generation, in which a large language model generates content within a predefined structural scheme. This provides more stable document generation quality, reduces the likelihood of violations in the logical organization of the text, and improves the reproducibility of generation results. Furthermore, the use of structural templates makes it possible to adapt the generation process to different document types without modifying the architecture of the language model itself.

Keywords: hierarchical document structure, graph theory, JSON templates, large language models (LLMs), structured text generation, structural patterns, document representation model

References

1. Liang X., Wang H., Wang Y. et al. Controllable text generation for large language models: a survey. Preprint arXiv:2408.12599, 2024, DOI: 10.48550/arXiv.2408.12599.
2. Liu Y., Li D., Wang K. et al. Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. Information Processing & Management, 2024, vol. 61, no. 5, art. no. 103809, DOI: 10.1016/j.ipm.2024.103809.
3. Lu Y., Li H., Cong X. et al. Learning to generate structured output with schema reinforcement learning. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 4905–4918.
4. Xing H., Cheng C., Gao F. et al. DocHieNet: a large and diverse dataset for document hierarchy parsing. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds.), 2024, pp. 1129–1142, DOI:10.18653/v1/2024.emnlp-main.65.
5. Pfitzmann B., Auer C., Dolfi M. et al. DocLayNet: a large human-annotated dataset for document-layout analysis. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), 2022, pp. 3530–3540, DOI: 10.1145/3534678.3539043.
6. Wang J., Krumdick M., Tong B. et al. A graphical approach to document layout analysis. Preprint arXiv:2308.02051, 2023, DOI: 10.48550/arXiv.2308.02051.
7. Jiovannini S., Marinai S. A survey on reading order, table of contents, and structure extraction in document analysis. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2025, pp. 7585–7594.
8. Rausch J., Martinez O., Bissig F. et al. DocParser: hierarchical document structure parsing from renderings. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, no. 5, pp. 4328–4338, DOI: 10.1609/aaai.v35i5.16558.
9. Wang J., Hu K., Zhong Z. et al. Detect-order-construct: a tree construction-based approach for hierarchical document structure analysis. Pattern Recognition, 2024, vol. 156, art. no. 110836, DOI: 10.1016/j.patcog.2024.110836.
10. Yang J., Jiang D., He L. et al. StructEval: benchmarking LLMs' capabilities to generate structural outputs. Preprint arXiv:2505.20139, 2026, DOI: 10.48550/arXiv.2505.20139
11. Chen Y., Liu R., Zheng J. et al. Graph-based document structure analysis. Preprint arXiv:2502.02501, 2025, DOI: 10.48550/arXiv.2502.02501
12. Kim D. Guided JSON with LLMs: from raw PDFs to structured intelligence. Available at: https://medium.com/@kimdoil1211/structured-output-with-guided-json-a-practical-guide-for-llm-developers-6577b2eee98a (accessed: 02/03/2025).
13. Geng S., Cooper H., Moskal M. et al. Generating structured outputs from language models: benchmark and studies. Preprint arXiv:2501.10868, 2025, DOI: 10.48550/arXiv.2501.10868.
14. Joshi I., Agarwal B., Rojkova V. Think inside the JSON: reinforcement strategy for strict LLM schema adherence. Preprint arXiv:2502.14905, 2025, DOI:10.48550/arXiv.2502.14905.
15. Balaskas G., Papadopoulos H., Pappa D. et al. A framework for domain-specific dataset creation and adaptation of large language models. Computers, 2025, vol. 14, no. 5, art. no. 172, DOI: 10.3390/computers14050172.
16. Taveekitworachai P., Dewantoro M.F., Xia Y. et al. BenchING: a benchmark for evaluating large language models in following structured output format instruction in text-based narrative game tasks. IEEE Transactions on Games, 2025, vol. 17, no. 3, pp. 665–675, DOI: 10.1109/TG.2025.3529117.

Ismagulov Milan Erikovich. *Postgraduate student in the field 2.3.1 – System Analysis, Management and Information Processing, Statistics. The author's main research areas: multimodal data conversion, machine learning, data processing methods using machine learning models. AuthorID: 1301221, SPIN: 6808-8045, ORCID: 0009-0007-3280-5259. m_ismagulov@ugrasu.ru. 628012, Khanty-Mansi Autonomous Okrug – Yugra, Khanty-Mansiysk, Chekhov St., 16.*

Статья поступила в редакцию 27.03.2026; одобрена после рецензирования 01.06.2026; принята к публикации 05.08.2026.

The article was submitted 03/27/2026; approved after reviewing 06/01/2026; accepted for publication 08/05/2026.