1 Обзор по представлениям логической структуры документов

Существует огромное количество документов самых разных форматов, доменов и имеющих различное устройство. Однако для всех документов можно определить три основных типа структуры, которая из них выделяется: физическая, логическая и семантическая [1].

1. Физическая структура документа описывает то, как выглядит документ. Другими словами, физическая структура оперирует терминами «символ», «набор символов», «строка», «блок текста», «страница». Помимо текстового содержимого, объектами физической структуры могут быть изображения, графики, таблицы и т. д. Кроме того, физическая структура описывает свойства символов (например, жирность или размер шрифта), строк (отступ от краев страницы) и текстовых блоков (расстояние между строками), описывается положение элементов на странице. Физическая структура не подразумевает выделения функций отдельных частей документа, их порядка чтения или смысла.

2. Логическая структура документа подразумевает некоторый анализ физических составляющих документа. Данный тип структуры связан с конкретным доменом, к которому документ относится, так как для разных доменов физические части документа выполняют разные функции и могут иметь разный семантический смысл (так, в научных статьях можно выделить введение, обзор существующих работ, заключение, список литературы). Выделение логической структуры документа подразумевает определение этих функций и семантического смысла конкретных составляющих документа, объединение в соответствии с этим физических частей документа (например, символы объединяются в слова и строки, строки объединяются в текстовые блоки). Кроме того, определяется взаимодействие частей друг с другом в физическом смысле, например, определяется порядок чтения или вложенность одного текстового блока в другой (многоуровневые списки или заголовки разных уровней). Логическая структура не анализирует смысл предложений, их взаимодействие друг с другом в семантическом смысле (например, в одном из предложений приводится пример, который описывается более подробно в следующем предложении). Логическая структура строится на основе известной физической структуры и правил, задаваемых конкретным доменом.

3. Семантическая структура связана с задачей понимания содержимого текста. Например, можно определить взаимодействие именованных сущностей, упомянутых в тексте. Задачами выявления семантики текста занимается отдельная область - обработка естественного языка.

В рамках решаемой задачи рассматривается логическая структура документа. Опишем некоторые из способов, которые применяются для представления такой структуры.

1.0.1 Представление структуры документа в виде дерева

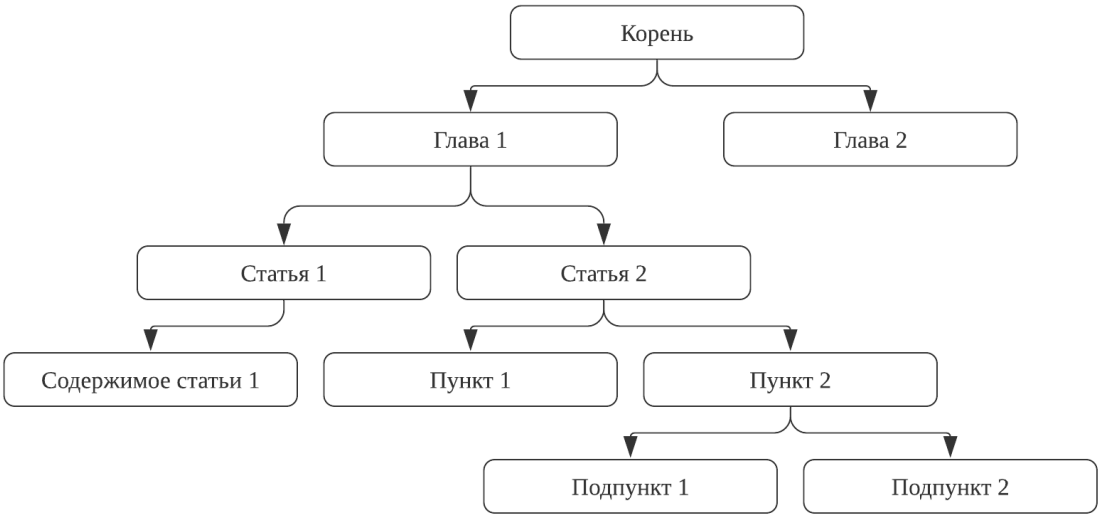

Рис. 1: Пример структуры документа в виде дерева

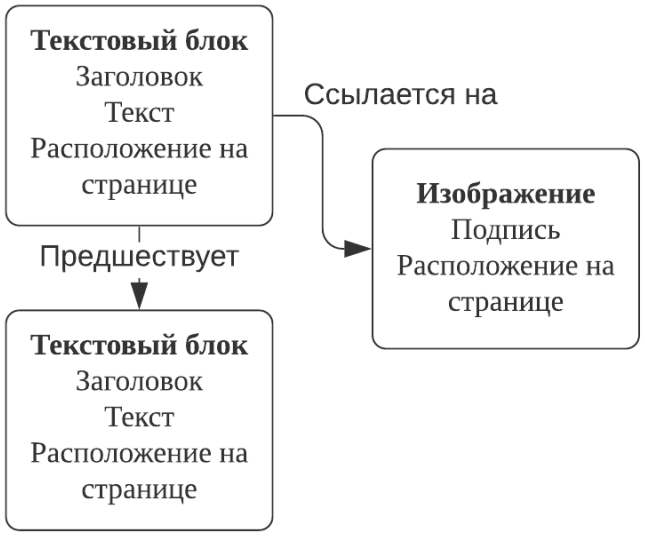

Рис. 2: Пример структуры документа в виде графа

Большое количество документов представляет из себя последовательность вложенных друг в друга частей. Например, статьи состоят из секций, которые, в свою очередь, могут состоять из подсекций и т. д., законы могут состоять из глав, статей, пунктов и

т. д. Дерево помогает получить представление документа в виде иерархической структуры, то есть документ разбивается на последовательность вложенных друг в друга элементов [2, 3] (рис. 1). На каждом уровне иерархии находятся элементы определенных типов. В листовых вершинах, как правило, располагается простой текстовый блок.

1.0.2 Представление структуры документа в виде графа

Произвольный граф позволяет представить разбиение документа на части (каждая часть является вершиной графа), а также описать порядок чтения частей [4] (ребра графа могут быть помечены и описывать тип взаимоотношений между частями документа). Структура докумета при этом может получиться не обязательно иерархической (рис. 2). Так, документ можно разбить на текстовые блоки, изображения, таблицы и определить порядок чтения этих элементов в документе.

$[0.5] < START > \rightarrow < TITLE > < COLUMN > < COLUMN >$
$[0.5] < START > \rightarrow < TITLE > < COLUMN >$
$[1.0] < TITLE > \rightarrow < text\_line >$
$[1.0] < COLUMN > \rightarrow < TEXT\_BLOCKS >$
$[0.8] < TEXT\_BLOCKS > \rightarrow < TEXT\_BLOCK > < space > < TEXT\_BLOCKS >$
$[0.2] < TEXT\_BLOCKS > \rightarrow < TEXT\_BLOCK >$
$[1.0] < TEXT\_BLOCK > \rightarrow < TEXT\_LINES >$
$[0.9] < TEXT\_LINES > \rightarrow < text\_line > < newline > < TEXT\_LINES >$
$[0.1] < TEXT\_LINES > \rightarrow < text\_line >$

(a)

$< START > \rightarrow < TITLE > < COLUMN >$
$\rightarrow < text\_line > < COLUMN >$
$\rightarrow < text\_line > < TEXT\_BLOCKS >$
$\rightarrow < text\_line > < TEXT\_BLOCK > < space > < TEXT\_BLOCKS >$
$\rightarrow < text\_line > < text\_line > < newline > < TEXT\_BLOCK > < space > < TEXT\_BLOCKS >$
$\rightarrow < text\_line > < text\_line > < newline > < text\_line > < space > < TEXT\_BLOCKS >$
$\rightarrow < text\_line > < text\_line > < newline > < text\_line > < space > < TEXT\_BLOCK >$
$\rightarrow < text\_line > < text\_line > < newline > < text\_line > < space > < text\_line >$

(b)

Рис. 3: Пример структуры документа в виде формальной грамматики: (a) пример стохастической контекстно-свободной грамматики, которая выводит текстовый документ с заголовком. Заглавными буквами помечены нетерминальные символы, строчными буквами обозначены терминальные символы. (b) пример вывода документа, состоящего из заголовка и трех строк, из грамматики, представленной в (a).

1.0.3 Представление структуры документа с использованием формальных грамматик

Документ может быть представлен последовательностью правил, которые необходимо обработать с помощью специального парсера [5]. В результате такой обработки получа-

ется исходный документ (рис. 3).

1.0.4 Представление структуры документа в виде зон и логических меток

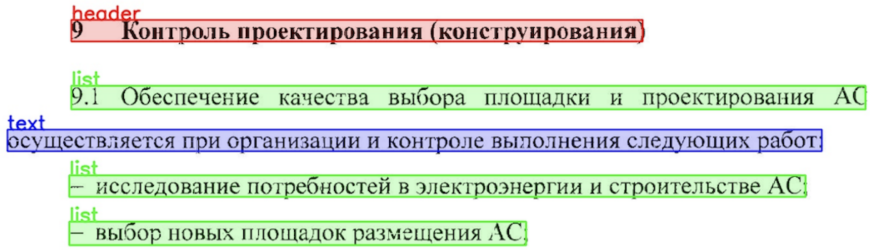

Рис. 4: Пример структуры документа в виде списка помеченных строк

Документ может быть представлен как плоская структура: последовательность частей какого-либо типа [6]. Такими частями могут быть страницы, текстовые блоки, строки, слова, символы и т. д. В зависимости от требований того или иного домена, каждому выделенному элементу назначается семантическая метка. Так, документ можно рассматривать построчно и каждой строке присваивать определенный тип, например, «заголовок», «элемент списка» и «простая текстовая строка» (рис. 4).

Помимо плоской структуры, семантические метки могут применяться и в более сложных структурах. Например, в иерархической структуре в виде дерева узлы могут быть типизированы и иметь некоторый семантический смысл.

Список литературы

[1] The representation of document structure: A generic object-process analysis / Dov Dori, David Doermann, Christian Shin et al. // Handbook of character recognition and document image analysis. — World Scientific, 1997. — Pp. 421–456.

[2] *Wexler, Michael C.* Structure extraction on electronic documents. — 2001. — 2. — US Patent 6,298,357.

[3] *Pembe, F Canan.* A Tree Learning Approach to Web Document Sectional Hierarchy Extraction. / F Canan Pembe, Tunga Güngör // ICAART (1). — 2010. — Pp. 447–450.

[4] *Paaß, Gerhard.* Machine learning for document structure recognition / Gerhard Paaß, Iuliu Konya // Modeling, Learning, and Processing of Text Technological Data Structures. — Springer, 2011. — Pp. 221–247.

[5] *Namboodiri, Anoop M.* Document structure and layout analysis / Anoop M Namboodiri, Anil K Jain // Digital Document Processing. — Springer, 2007. — Pp. 29–48.

[6] A Machine-Learning Based Approach for Extracting Logical Structure of a Styled Document. / Tae-young Kim, Suntae Kim, Sangchul Choi et al. // *TIIS*. — 2017. — Vol. 11, no. 2. — Pp. 1043–1056.