

ISSN 2500-316X (Online)

<https://doi.org/10.32362/2500-316X-2021-9-1-7-17>



УДК 004.6

НАУЧНАЯ СТАТЬЯ

Способы обработки неструктурированных данных

В.С. Томашевская[@],
Д.А. Яковлев

МИРЭА – Российский технологический университет, Москва 119454, Россия

[@]Автор для переписки, e-mail: tomashevskaya@mirea.ru

Термин «неструктурированные данные» подразумевает данные, неупорядоченные и произвольные по форме, однако этот тип информации все же обладает определенной структурой. На сегодняшний день существует большое разнообразие данных и, как следствие, появляется необходимость их интерпретировать. Среди задач интерпретации можно выделить прогнозирование, классификацию, кластеризацию, ассоциацию, поиск последовательностей, визуализацию данных, анализ отклонений. Сложность обработки заключается в том, что сами данные могут различаться не только с точки зрения формата, но и с точки зрения своей структуры. Одной из ключевых задач при работе с неструктурированными данными является поиск и выявление закономерностей с целью их понимания и разработки шаблонов заполнения. В работе проводится анализ правил оформления библиографических источников с целью выявления общих закономерностей. Затрагиваются понятия структурированных и неструктурированных данных. Рассматриваются существующие направления работы с неструктурированными данными и способы их обработки, в частности, правила оформления библиографических списков литературных источников. На основании этих правил сформированы шаблоны, состоящие из смысловых групп, на основе примеров соответствующих списков библиографических источников. При итоговом сравнении полученных шаблонов выявлены как общие черты, объединяющие все рассмотренные шаблоны, так и черты, их разделяющие.

Ключевые слова: неструктурированные данные, текстовая аналитика, неструктурированная информация

Для цитирования: Томашевская В.С., Яковлев Д.А. Способы обработки неструктурированных данных. *Российский технологический журнал*. 2021;9(1):7–17. <https://doi.org/10.32362/2500-316X-2021-9-1-7-17>

RESEARCH ARTICLE

Research of unstructured data interpretation problems

Valeriya S. Tomashevskaya[@],
Dmitriy A. Yakovlev

MIREA – Russian Technological University, Moscow 119454, Russia

[@]Corresponding author, e-mail: tomashevskaya@mirea.ru

The term «unstructured data» means data that is unordered and arbitrary in shape. However, this type of information has a certain structure. Today there is a wide variety of data and, as a result, it is necessary to interpret them. Interpretation tasks include forecasting, classification, clustering, association, sequence search, data visualization, and variance analysis. The difficulty lies in the fact that the data itself can differ not only in terms of format, but also in terms of its structure. One of the key tasks when working with unstructured data is to find and identify patterns in order to understand them and develop filling patterns. The paper analyzes the rules for the design of bibliographic sources in order to identify common patterns. The concepts of structured and unstructured data are touched upon. The existing directions of work with unstructured data and methods of processing unstructured data, in particular, the rules for the design of bibliographic lists of literary sources, are considered. These rules were used to form templates consisting of semantic groups on the basis of examples of the corresponding lists of bibliographic sources. The final comparison of the obtained templates revealed both common features that unite all the considered templates and features that separate them.

Keywords: unstructured data, text analytics, unstructured information

For citation: Tomashevskaya V.S., Yakovlev D.A. Research of unstructured data interpretation problems. *Rossiiskii tekhnologicheskii zhurnal = Russian Technological Journal*. 2021;9(1):7–17 (in Russ.). <https://doi.org/10.32362/2500316X-2021-9-1-7-17>

Введение

На сегодняшний день существует большое разнообразие данных и, как следствие, появляется необходимость их интерпретировать. Среди задач интерпретации можно выделить прогнозирование, классификацию, кластеризацию, ассоциацию, поиск последовательностей, визуализацию данных, анализ отклонений. Эти задачи находят применение в таких областях как розничная торговля, телекоммуникации, банковское дело, страхование, медицина, техника и т.д. Сложность обработки данных заключается в том, что они могут различаться как с точки зрения формата, так и с точки зрения структуры.

По своей структуре данные можно разделить на два типа – структурированные и неструктурированные. К первому типу относятся данные, упорядоченные определенным образом и организованные с целью обеспечения возможности применения к ним некоторых действий [1]. Второй

тип – это неупорядоченные данные, произвольные по форме, которые были собраны независимо от того, как они могут быть использованы [2].

Для работы со вторым типом данных применяются три направления: интеллектуальный анализ данных, обработка естественного языка и интеллектуальный анализ текста. Первое направление исследует и решает проблемы, связанные с поиском неочевидных закономерностей и причинно-следственных связей в системах, процессах и явлениях [3]. Здесь применяются такие методы как дескриптивный, корреляционный, регрессионный, факторный анализ и многие другие методы, требующие представления входной информации определенным образом [4, 5]. Данное направление применимо в любых сферах человеческой деятельности, где существуют какие-либо данные и есть необходимость принятия решений: в экономике, здравоохранении, страховании, торговле, в различных областях, связанных с контролем и прогнозированием состояния

сложных динамических систем [6]. Второе направление подразумевает под собой компьютерный анализ и синтез естественных языков. Естественным языком является любой язык, используемый для общения людей и не созданный целенаправленно. Главной целью является создание более удобной формы взаимодействия компьютера и человека [7]. Третье направление представляет собой автоматизированный аналитический процесс извлечения новых и потенциально полезных знаний из текстовых документов [8]. При этом главной задачей является нахождение в неструктурированной текстовой информации новых знаний. Схожесть третьего с первым направлением наблюдается в целях, подходах к переработке информации и сферах применения. Различие между направлениями выражается в используемых методах.

Существует несколько способов обработки неструктурированных данных:

- 1) выделение онтологии – описания-схемы предметной части, характеризующейся конкретной структурой;
- 2) поиск упоминаний, категоризация и извлечение фактов;
- 3) выделение эмоциональной окраски, оценки интересов, отношения;
- 4) выделение закономерностей – динамика и процесс изменения отношения, выделение общего, заимствования [9].

Целью данной статьи является анализ правил оформления библиографических источников и выявление возможных тенденций, закономерностей, зависимостей или аномалий с целью получения новых знаний для последующего их применения. Для достижения данной цели будет использоваться четвертый способ обработки неструктурированных данных.

Анализ правил оформления библиографических источников с целью выявления общих закономерностей

В рамках заданных условий в наличии имеется список правил оформления литературных источников. По ходу решения задачи необходимо выявить различные тенденции, закономерности, зависимости или аномалии, на основании которых можно будет

составить шаблоны определения типа литературного источника. Правила оформления следующие:

- оформление описаний монографий;
- оформление описаний статей или отдельных глав с указанием разных авторов из книги или сборника;
- оформление описаний диссертаций и авторефератов диссертаций;
- оформление описаний статей из газет или журналов;
- оформление описаний источников электронного ресурса удаленного доступа.

Первым шагом просматривается текст на естественном языке. Затем осуществляется его разбиение на одиночные токены. На третьем и последнем шаге токены объединяются в смысловые группы. Эта последовательность действий проиллюстрирована на рис. 1.

Тексты и документы представляют собой неструктурированные наборы данных, которые можно представить в виде $A = \{X_1, X_2, \dots, X_N\}$, где A – это предложение, содержащее текстовые сегменты, а X – текстовый сегмент, содержащий набор символов.

Вначале текст распределяется на образующие его элементы, как это показано на рис. 2.

Этот процесс подразумевает представление текста в виде последовательности слов. Для этого используется метод предварительной обработки – токенизация – разбиение потока текста на более мелкие части, которые называются токенами или лексическими элементами. К токенам относятся

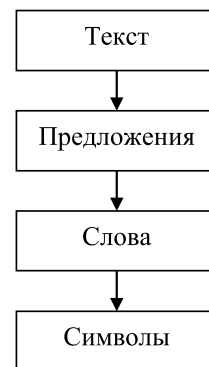


Рис. 2. Распределение текста на составные элементы.

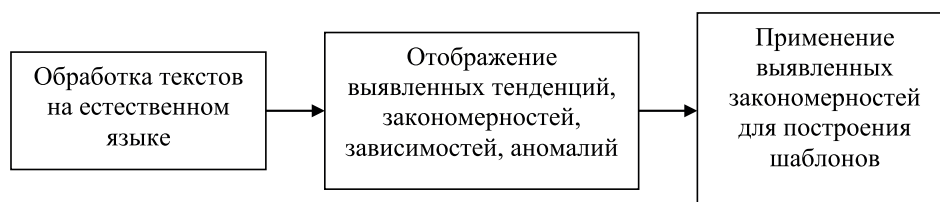


Рис. 1. Последовательность действий для построения шаблонов и выявления закономерностей при анализе списков литературных источников.

как слова, так и знаки препинания, границы абзацев и т.п. Как правило, после токенизации необходимо провести чистку информации на предмет знаков пунктуации и незначимых слов. Пример токенизации приведен на рис. 3.

Сам процесс не представляет трудности в языках, использующих символы-пробелы для разделения слов, но в языках, подобных китайскому, это сделать гораздо труднее, поскольку иероглифы могут обозначать как слоги, так и целые слова [10]. После разбиения текста происходит сборка токенов в группы, которые разграничиваются различными знаками препинания. Следующие за знаком препинания пробелы удаляются, поскольку они являются сугубо конвенциональными и не несут функциональной нагрузки [11]. Данные группы представляют собой смысловые части предложения, при этом важно отметить, что в роли такой части может выступать даже одиночный токен. Полученные группы можно представить в виде шаблона предложения.

Список требований по оформлению литературных источников основывается на стандартах, среди которых можно выделить стандарт ГОСТ Р 7.0.5-2008 СИБИД¹. Он подразумевает деление ссылок следующим образом:

– в зависимости от того, где они размещаются

в тексте: внутритекстовые, подстрочные и затекстовые;

– по составу элементов, находящихся в составе ссылки: полные и краткие.

В данной работе объектом рассмотрения являются краткие затекстовые ссылки, главное назначение которых заключается в оказании помощи при поиске объекта ссылки. Правила, утверждающие использование знаков препинания и представления элементов, должны осуществляться с использованием стандартов ГОСТ Р 7.0.100-2018 (СИБИД)² и ГОСТ 7.82-2001 СИБИД³. Данные правила являются независимыми от назначения ссылки.

При оформлении библиографического описания в ссылке его дополняют заголовком библиографической записи в соответствии с ГОСТ 7.80-2000 СИБИД⁴. Требования к сокращениям слов устанавливаются в соответствии со стандартами ГОСТ Р 7.0.12-2011⁵ и ГОСТ 7.11-2004 (ИСО 832:1994) СИБИД⁶. Первый регламентирует правила для слов на русском языке, второй – для слов на иностранных европейских языках. Следование требованиям вышеуказанных пяти ГОСТов осуществляется с учетом особенностей, указанных в стандарте ГОСТ Р 7.0.5-2008 СИБИД. Библиографические ссылки, в зависимости от стиля, могут иметь различный набор правил оформления описа-

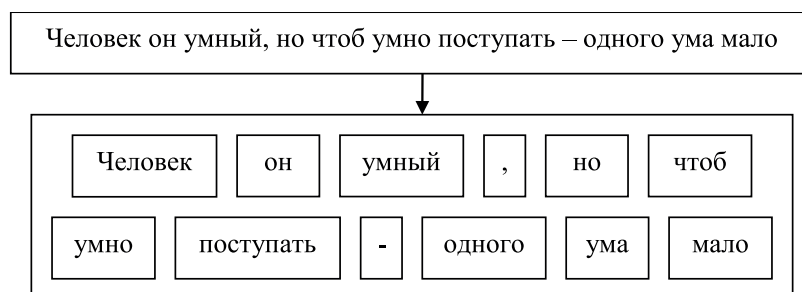


Рис. 3. Пример токенизации.

¹ Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс].

URL: <http://docs.cntd.ru/document/1200063713> [Elektronnyi fond pravovoi i normativno-tekhicheskoi dokumentatsii (Electronic fund of legal and normative-technical documentation). [Electronic resource]. URL: <http://docs.cntd.ru/document/1200063713>]

² Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс].

URL: <http://docs.cntd.ru/document/1200161674> [Elektronnyi fond pravovoi i normativno-tekhicheskoi dokumentatsii (Electronic fund of legal and normative-technical documentation). [Electronic resource]. URL: <http://docs.cntd.ru/document/1200161674>]

³ Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс].

URL: <http://docs.cntd.ru/document/1200025968> [Elektronnyi fond pravovoi i normativno-tekhicheskoi dokumentatsii (Electronic fund of legal and normative-technical documentation). [Electronic resource]. URL: <http://docs.cntd.ru/document/1200025968>]

⁴ Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс].

URL: <http://docs.cntd.ru/document/1200006960> [Elektronnyi fond pravovoi i normativno-tekhicheskoi dokumentatsii (Electronic fund of legal and normative-technical documentation). [Electronic resource]. URL: <http://docs.cntd.ru/document/1200006960>]

⁵ Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс].

URL: <http://docs.cntd.ru/document/1200093114> [Elektronnyi fond pravovoi i normativno-tekhicheskoi dokumentatsii (Electronic fund of legal and normative-technical documentation). [Electronic resource]. URL: <http://docs.cntd.ru/document/1200093114>]

⁶ Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс].

URL: <http://docs.cntd.ru/document/1200039536> [Elektronnyi fond pravovoi i normativno-tekhicheskoi dokumentatsii (Electronic fund of legal and normative-technical documentation). [Electronic resource]. URL: <http://docs.cntd.ru/document/1200039536>]

ний для монографий, статей, электронных ресурсов и т.д. В работе рассматриваются правила оформления согласно требованию ВАК РФ. При описании правил оформления указывается текстовое наполнение и сопутствующие знаки препинания.

Правило оформления описания монографий

Автор/ы: указывается фамилия и инициалы автора монографии, при наличии более одного автора они перечисляются через запятую. Заглавие: указываются сведения, относящиеся к заглавию, разделительный знак «/», указываются сведения об ответственности. Сведения об издании включают в себя номер издания и информацию о переиздании. Серия: указывается серия. Место издания: разделительный знак «—», сокращенное название издательства. Год издания: указывается дата, когда монография была опубликована. Объем: разделительный знак «—», указывается количество страниц монографии. Разбиение текста на соответствующие группы представлено на примере ниже (рис. 4).

Полученные группы можно представить в виде

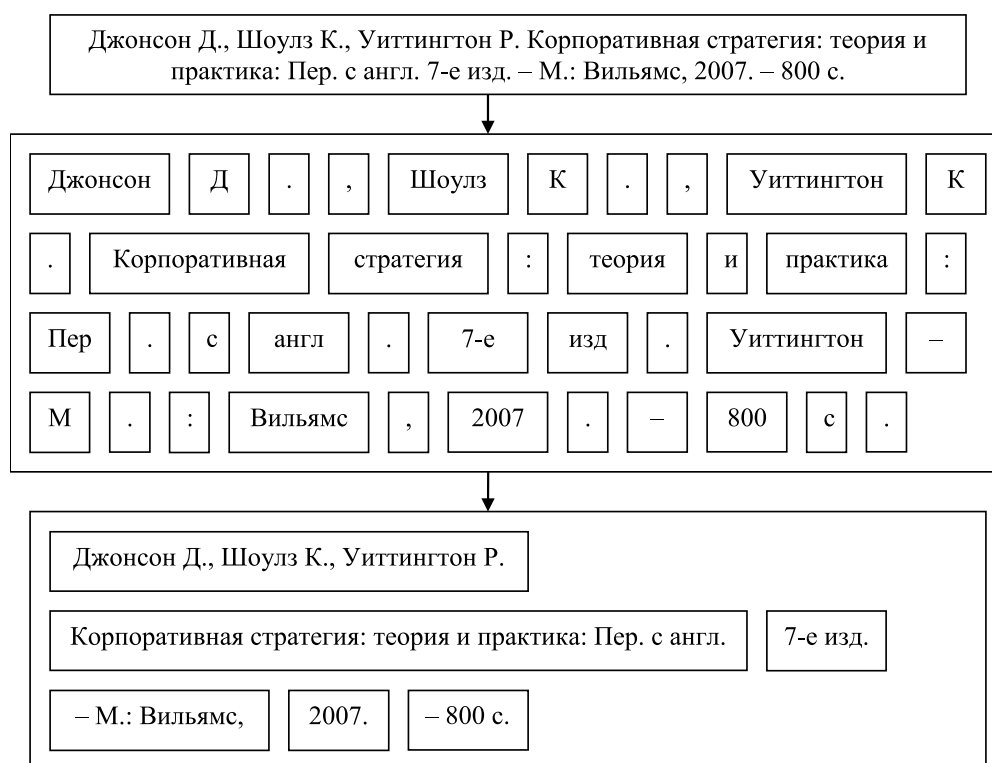


Рис. 4. Разбиение описания книг на смысловые группы.

Автор/ы	Заглавие	Сведения об издании	Место издания	Год издания	Объем
---------	----------	---------------------	---------------	-------------	-------

Рис. 5. Шаблон примера оформления описания монографий.

следующего шаблона (рис. 5).

Правило оформления описания статей или отдельных глав с указанием разных авторов из книги или сборника

Автор/ы: указывается фамилия и инициалы автора статьи, при наличии более одного автора перечисляются через запятую. Заглавие: указываются сведения, относящиеся к заглавию, разделительный знак «/», указываются сведения об ответственности. Заглавие книги: разделительный знак «//», указываются сведения, относящиеся к заглавию. Сведения об ответственности: разделительный знак «/», указываются сведения об ответственности, указываются последующие сведения об ответственности. Место издания: разделительный знак «—», сокращенное название издательства. Год издания: указывается дата, когда статья была опубликована. Объем: разделительный знак «—», указывается интервал страниц, обозначающий местоположение статьи. Разбиение текста на соответствующие группы представлено на примере ниже (рис. 6).

Полученные группы можно представить в виде следующего шаблона (рис. 7).

Правило оформления описаний диссертаций и авторефератов диссертаций

Автор: указываются фамилия и инициалы автора диссертации. Заглавие: указываются сведения, относящиеся к заглавию, разделительный знак «:», добавляется дисс. (автореф. дисс.), разделительный знак «...». Кандидат (доктор) наук: текстовая конструкция вида «канд.(док.) наук.» с указанием отрасли наук автора диссертации. Место издания (место написания): разделительный знак «—», название издательства (если указано). Год издания: указывается дата, когда диссертация была опубликована. Объем: разделительный знак «—», указывается количество страниц диссертации. Разбиение

текста на соответствующие группы представлено на примере ниже (рис. 8).

Полученные группы можно представить в виде следующего шаблона (рис. 9).

Правило оформления описаний статей из газет или журналов

Автор: указывается фамилия и инициалы автора статьи. Заглавие: указываются сведения, относящиеся к заглавию. Название журнала: разделительный знак «//» указывается название журнала. Год выпуска: разделительный знак «—», указывается дата, когда журнал был опубликован. Номер выпуска: разделительный знак «—», указывается номер выпуска. Местоположение статьи: разделительный знак «—», указывается интервал страниц, обозначающий местоположение статьи. Разбиение текста на

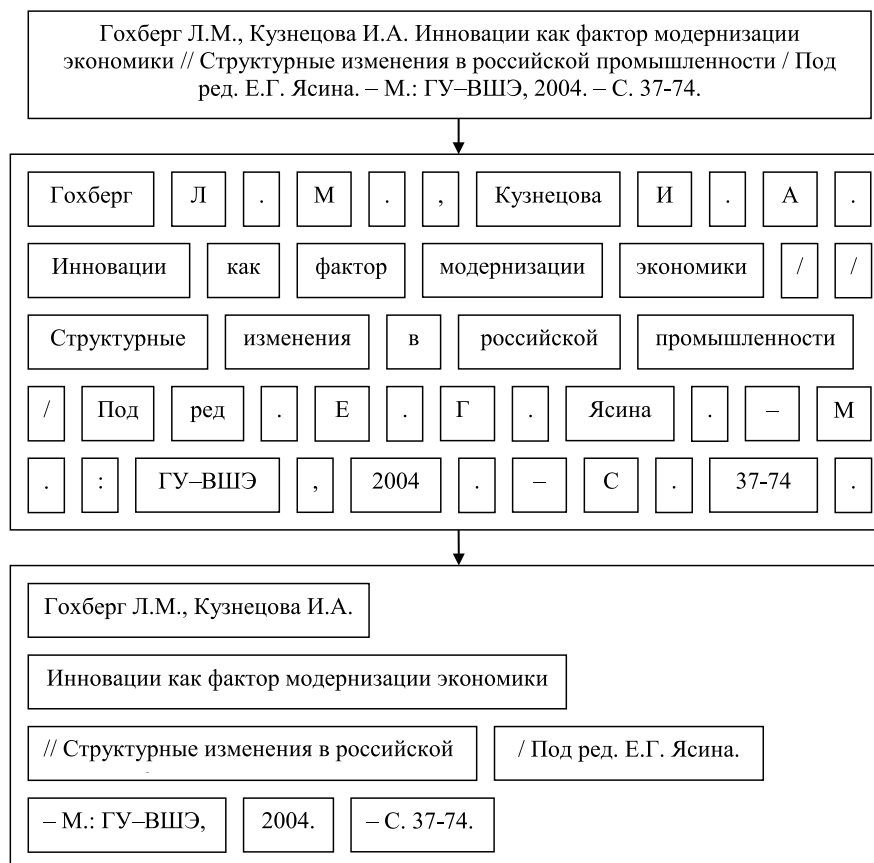


Рис. 6. Разбиение описания статей или отдельных глав с указанием разных авторов из книги или сборника на смысловые группы.

Автор/ы	Заглавие статьи	Заглавие книги	Сведения об ответственности	Место издания	Год издания
Объем					

Рис. 7. Шаблон примера оформления описания статей или отдельных глав с указанием разных авторов из книги или сборника.

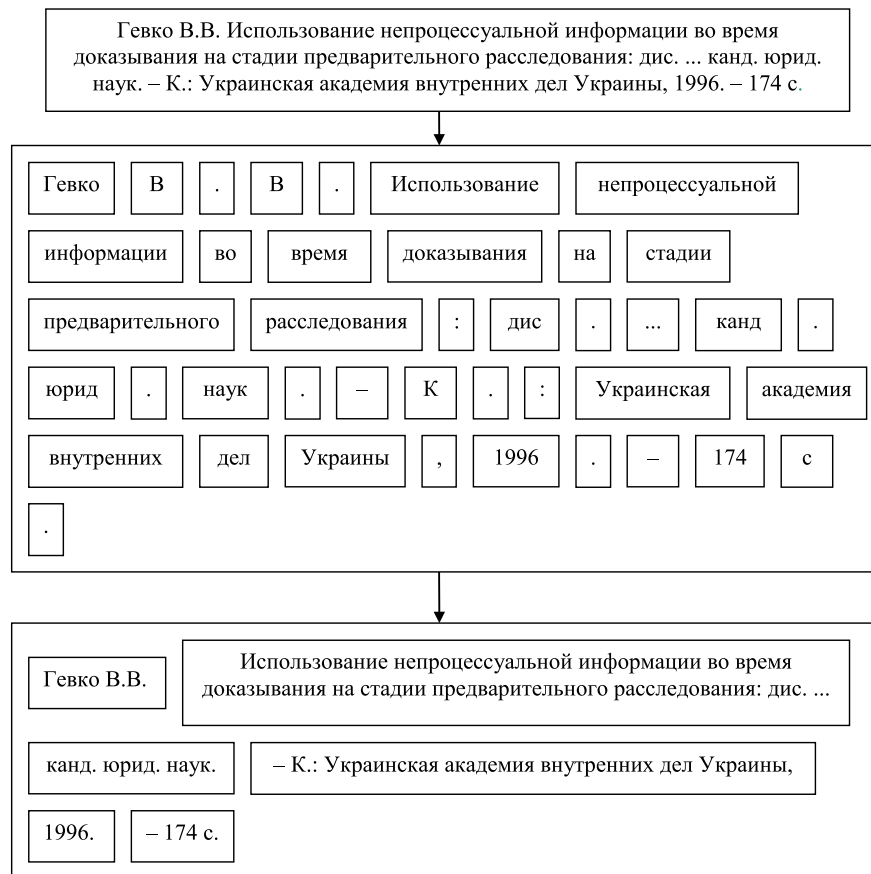


Рис. 8. Разбиение описаний диссертаций и авторефератов диссертаций на смысловые группы.

Автор	Заглавие	канд.(док.) наук.	Место написания	Год издания	Объем
-------	----------	-------------------	-----------------	-------------	-------

Рис. 9. Шаблон примера оформления описаний диссертаций и авторефератов диссертаций.

соответствующие группы представлено на примере ниже (рис. 10).

Полученные группы можно представить в виде следующего шаблона (рис. 11).

Пример оформления описания источников электронного ресурса удаленного доступа

Автор: указывается фамилия и инициалы автора. Заглавие: указываются сведения, относящиеся к заглавию. Название: разделительный знак «//», название источника, текстовая конструкция вида «[Электронный ресурс]. – Режим доступа:», указывается ссылка на ресурс, текстовая конструкция вида «(дата обращения: число, месяц, год)» с указанием числа, месяца и года. Разбиение текста на соответствующие группы представлено на примере ниже (рис. 12).

Полученные группы можно представить в виде следующего шаблона (рис. 13).

Полученные результаты можно представить в виде таблицы.

При рассмотрении таблицы можно выделить три типа полей. К первому можно отнести поля, которые идентичны по смысловому наполнению и по формулировке, ко второму – те поля, которые имеют разную формулировку, но схожее значение. Третий тип – это неидентичные поля.

Поля колонок 1 и 2 можно отнести к первому типу. Колонка под номером 1 отведена под имена авторов, а под номером 2 – предназначена для заглавия литературного источника. Далее можно выделить поле «Место издания» в строке «Оформление описаний монографий» и в строке «Оформление описаний статей или отдельных глав с указанием разных авторов из книги или сборника». Они предназначены для одинаковой смысловой информации, но занимают разное местоположение в своих строках.

Второй тип полей можно наблюдать при рассмотрении полей «Год издания» в строках

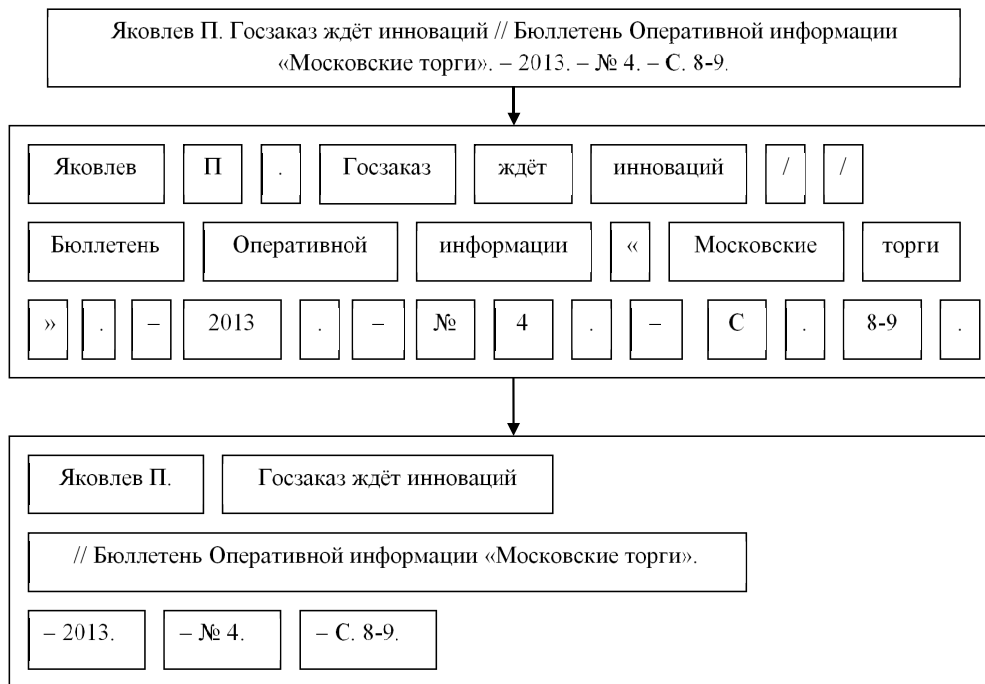


Рис. 10. Разбиение описания статьи из газеты или журнала на смысловые группы.

Автор	Заглавие	Название журнала	Год выпуска	Номер выпуска	Местоположение статьи
-------	----------	------------------	-------------	---------------	-----------------------

Рис. 11. Шаблон примера оформления описаний статьи из газеты или журнала.

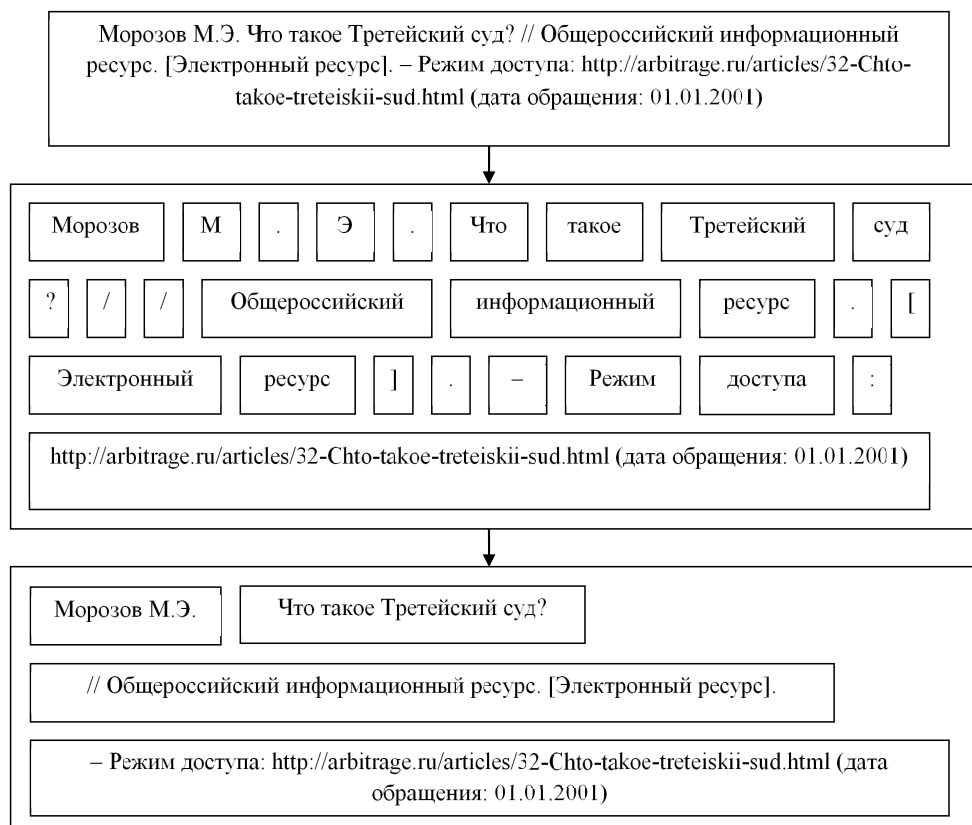


Рис. 12. Разбиение описаний источников электронного ресурса удаленного доступа на смысловые группы.

Автор	Заглавие статьи	Название источника	Режим доступа
-------	-----------------	--------------------	---------------

Рис. 13. Шаблон оформления описаний источников электронного ресурса удаленного доступа.

Сравнение шаблонов оформления литературных источников.

Название строк	1	2	3	4	5	6	7
Оформление описания монографий	Автор/ы	Заглавие	Сведения об издании	Место издания	Год издания	Объем	
Оформление описания статей или отдельных глав с указанием разных авторов из книги или сборника	Автор/ы	Заглавие	// Заглавие книги	/ сведения об ответственности	Место издания	год издания	Местоположение статьи
Оформление описания диссертаций и авторефератов диссертаций	Автор	Заглавие	канд.(док.) наук	Место написания	Год издания	Объем	
Оформление описания статей из газет или журналов	Автор.	Заглавие	// Название журнала	Год выпуска	Номер выпуска	Местоположение статьи	
Оформление описания источников электронного ресурса удаленного доступа	Автор	Заглавие	// Название источника. [Электронный ресурс].	Режим доступа: URL и дата обращения	Дата обращения		

«Оформление описания монографий», «Оформление описания статей или отдельных глав с указанием разных авторов из книги или сборника», «Оформление описания диссертаций и авторефератов диссертаций» и поля «Год выпуска» в строке «Оформление описания статей из газет или журналов». Несмотря на отличия в названии, все эти поля содержат информацию о годе выхода публикации.

К третьему типу можно отнести оставшиеся поля.

В результате при осуществлении анализа на определение принадлежности к литературному источнику необходимо начинать с третьей колонки, так как различия начинаются именно с нее.

Заключение

Несмотря на то, что термин «неструктурированные данные» подразумевает данные неупорядоченные и произвольные по форме, этот тип данных все же обладает определенной структурой. Одной из ключевых задач при осуществлении работы с неструктурированными данными является поиск и выявление закономерностей с целью их понимания и разработки шаблонов заполнения.

На основании правил оформления библиографических списков литературных источников сформированы шаблоны, состоящие из смысловых групп. Сравнение шаблонов выявило в них поля, как схожие по смыслу и значению, так и отличные друг от друга. Опираясь на эти сведения, можно будет определить тип библиографического источника и удостовериться в правильности его оформления.

Таким образом, на примере данной работы демонстрируется, что из потока данных можно отобрать образцы (шаблоны), анализ которых поможет выделить закономерности всего потока данных с целью получения новых знаний о связанных с ним процессах.

Вклад авторов

Все авторы в равной степени внесли свой вклад в исследовательскую работу.

Авторы заявляют об отсутствии конфликта интересов.

Authors' contribution

All authors equally contributed to the research work.

The authors declare no conflicts of interest.

Литература

1. Николаев А.А. Разнообразие структур данных в современной информации. *Молодой ученый*. 2019;23(261):21–23. URL: <https://moluch.ru/archive/261/60410/>
2. Антонов С.И., Редько С.Г. Автоматизация управления неструктурированными данными в рамках системы управления контентом на предприятии. *Научно-технические ведомости СПбГПУ. Инноватика*. 2009;5:277–282.
3. Татур М.М., Лукашевич М.М., Перцев Д.Ю., Искра Н.А. Интеллектуальный анализ данных и облачные вычисления. *Доклады БГУИР*. 2019;6(124):62–71. <https://doi.org/10.35596/1729-7648-2019-124-6-62-71>
4. Дядичев В.В., Ромашка Е.В., Голуб Т.В., Задачи и методы интеллектуального анализа данных. *Геополитика и экогеодинамика регионов*. 2015;1-11(3):23–29.
5. Амаева Л.А. Сравнительный анализ методов интеллектуального анализа данных. *Инновационная наука*. 2017;2(1):27–29.
6. Клишко Е.Г. Программно-алгоритмические средства интеллектуального анализа данных. *Радиоэлектроника и информатика*. 2001;3:64–67.
7. Юсков В.С., Баранникова И.В. Сравнительный анализ платформ обработки естественного языка. *Горный информационно-аналитический бюллетень*. 2017;3:272–278.
8. Цитульский А.М., Иванников А.В., Рогов И.С. Интеллектуальный анализ текста. *Научно-образовательный журнал для студентов и преподавателей «StudNet»*. 2020;6:476–483.
9. Укуев Б.Т. Особенности обработки неструктурированных данных в информационной базе научных исследований ВУЗа. *Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки*. 2018;03:75–76.
10. Цитульский А.М., Иванников А.В., Рогов И.С. NLP – Обработка естественных языков. *Научно-образовательный журнал для студентов и преподавателей «StudNet»*. 2020;6:467–475.
11. Кузнецов С.А., Скребцова Т.Г., Суворов С.Г., Клементьева А.В. Лингвистический анализатор: преобразование текста в метаязыковую структуру данных. СПб: Изд-во СПбГУ; 2019. 238 с. ISBN 978-5-288-05927-8

References

1. Nikolaev A.A. Variety of data structures in modern information. *Molodoi uchenyi*. 2019;23(261):21–23 (in Russ.). URL: <https://moluch.ru/archive/261/60410/>
2. Antonov S.I., Redko S.G. Automation of unstructured data management as part of a content management system at an enterprise. *Nauchno-tehnicheskie vedomosti Sankt Peterburgskogo tehicheskogo universiteta. Innovatika = Scientific and technical statements of SPbSPU (Saint Petersburg State Polytechnic University)*. 2009;5:277–282 (in Russ.).
3. Tatur M.M., Lukashevich M.M., Pertsev D.Yu., Iskra N.A. Intelligent data analysis and cloud computing. *Doklady BGUIR*. 2019;6(124):62–71 (in Russ.). <https://doi.org/10.35596/1729-7648-2019-124-6-62-71>
4. Dyadichev V.V., Romashka E.V., Golub T.V. Objectives and methods of data mining. *Geopolitika i jekogeodinamika regionov = Geopolitics and ecogeodynamics of regions*. 2015;1-11(3):23–29 (in Russ.).
5. Amaeva L.A. Comparative analysis of data mining methods. *Innovacionnaja nauka = Innovation Science*. 2017;2(1):27–29 (in Russ.).
6. Klimko E.G. Software-algorithmic tools for data mining. *Radioelektronika i informatika = Radioelectronics & Informatics*. 2001;3:64–67 (in Russ.).
7. Yuskov V.S., Barannikova I.V. Comparison of platforms of natural language processing. *Gornyy informacionno-analiticheskij bjulleten' = Mining Information and Analytical Bulletin*. 2017;3:272–278 (in Russ.).
8. Tsitulsy A.M., Ivannikov A.V., Rogov I.S. Intelligent text analysis. *Nauchno-obrazovatel'nyj zhurnal dlja studentov i prepodavatelej «StudNet» = Scientific and educational journal for students and teachers «StudNet»*. 2020;6:476–483 (in Russ.).
9. Ukuev B.T. Features of processing of unstructured data in information base of scientific research of higher education institution. *Sovremennaja nauka: aktual'nye problemy teorii i praktiki. Serija: Estestvennye i tehicheskie nauki = Modern Science: actual problems of theory & practice. Series of Natural and technical sciences*. 2018;3:75–76 (in Russ.).
10. Tsitulsy A.M., Ivannikov A.V., Rogov I.S. NLP – Natural Language Processing. *Nauchno-obrazovatel'nyj zhurnal dlja studentov i prepodavatelej «StudNet» = Scientific and educational journal for students and teachers «StudNet»*. 2020;6:467–475 (in Russ.).
11. Kuznecov S.A., Skrebцова T.G., Suvorov S.G., Klement'eva A.V. *Lingvisticheskij analizator: preobrazovanie teksta v metajazykovuju strukturu dannyh (Linguistic analyzer: transformation of text into a metalanguage data structure)*. St. Petersburg: St. Petersburg State Polytechnic University Publishing house; 2019. 238 p. (in Russ.). ISBN 978-5-288-05927-8

Об авторах:

Томашевская Валерия Сергеевна, кандидат технических наук, доцент кафедры корпоративных информационных систем Института информационных технологий ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: tomashevskaya@mirea.ru. <http://orcid.org/0000-0002-6460-2866>.

Яковлев Дмитрий Андреевич, аспирант кафедры корпоративных информационных систем Института информационных технологий ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: abracadabraZzzz@yandex.ru. <http://orcid.org/0000-0002-3630-5389>.

About the authors:

Valeriya S. Tomashevskaya, Cand. Sci. (Engineering), Associate Professor of the Department of Corporate Information Systems, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow 119454, Russia). E-mail: tomashevskaya@mirea.ru. <http://orcid.org/0000-0002-6460-2866>.

Dmitriy A. Yakovlev, Postgraduate Student of the Department of Corporate Information Systems, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow 119454, Russia). E-mail: abracadabraZzzz@yandex.ru. <http://orcid.org/0000-0002-3630-5389>.

Поступила: 06.08.2020; получена после рецензирования: 26.10.2020; принята к опубликованию: 05.11.2020.