УДК 004.853 DOI: 10.14529/ctcr210403

# ФОРМАЛИЗАЦИЯ БАЗОВЫХ ПРОЦЕССОВ И МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СИСТЕМЫ МОНИТОРИНГА И АНАЛИЗА ПУБЛИКАЦИЙ ЭЛЕКТРОННЫХ СМИ

***В.Н. Комаров[1], С.М. Рощин[2]***

*[1] Уральский колледж строительства, архитектуры и предпринимательства, г. Екатеринбург, Россия,*
*[2] Брянский государственный инженерно-технологический университет, г. Брянск, Россия*

В статье описан подход формализации базовых процессов и построения математической модели для системы сбора и анализа данных из электронных СМИ. Авторы, в рамках проведения научного исследования, занимаются созданием системы, в том числе ведется разработка новых алгоритмов, методов и подходов сбора и анализа текстовой информации из новостных интернет-источников. Основным направлением в исследовании рассматривается применение методов интеллектуального анализа текстовых данных на основе технологии искусственных нейронных сетей, методов обработки естественного языка, text mining, машинного обучения и обработки больших данных. **Цель исследования.** Разработать формализованное описание модели системы мониторинга и анализа текстовой информации электронных новостных СМИ методами математического моделирования. **Методы и инструментарий исследования.** Предложено использование инструментария методологии математического моделирования с методами системного анализа. Для исследования системы применены такие методы системного анализа, как абстрагирование, формализация, композиция и декомпозиция, структурирование и реструктурирование, моделирование, распознавание и идентификация. Система рассматривается как формализованная модель автоматического классификатора и кластеризатора набора текстовых документов на естественном языке в виде алгебраической системы. Для решения задач классификации и кластеризации текстов предложено применять методы машинного обучения на основе нейросетевых подходов. Структура системы и составляющие её процессы, а также процессы взаимодействующие с системой извне, представлены в виде формализованного математического описания. **Результаты.** Разработанное формализованное математическое описание модели системы наглядно показывает взаимосвязь компонентов системы между собой, а также внутренние процессы. Применяемый подход позволяет детализировать представление системы на основе ее декомпозиции на подсистемы и модули. Все это дает возможность упорядочить последовательности этапов создания системы и декомпозировать их на отдельные этапы работ. **Заключение.** Результаты, полученные в ходе проведенного исследования, позволяют перейти к следующему этапу жизненного цикла разрабатываемой информационной системы – ее программной разработке.

*Ключевые слова: мониторинг информации СМИ, анализ данных, система мониторинга и анализа данных, анализ текста, математическая модель системы, интеллектуальный анализ данных, нейросетевые методы, системный анализ, классификация текстов, кластеризация текстов.*

## Введение

Ранее, в рамках диссертационного исследования, авторы в статьях [1, 2] исследовали вопросы воздействия современных электронных новостных интернет-источников на общество, в частности на оборонно-промышленные предприятия нашей страны. Были представлены обобщенный алгоритм работы и структура создаваемой автоматизированной системы мониторинга и анализа текстовой информации в электронных СМИ.

В статье [3] авторы показали моделирование системы методами модельно-ориентированного системного инжиниринга.

В настоящей статье показано формализованное представление базовых процессов системы и её математическая модель.

**1. Формализация базовых процессов системы**

Многообразие различных процессов, протекающих в любой системе или с которыми она взаимодействует из вне, затруднительно изучить без создания упрощенной модели. Однако упрощение должно включать процессы наиболее значимые для изучения [4]. Именно поэтому применительно к разрабатываемой авторами системе мониторинга и анализа текстовой информации из электронных СМИ, целесообразно рассмотреть лишь основные, базовые процессы, протекающие как в самой системе, так и взаимодействующие с ней извне.

Ключевыми методами в задаче анализа текстовой информации являются классификация и кластеризация, поскольку другие функции системы, например, определение эмоциональной окраски текста, также сводятся к классификации [5, 6].

В виде алгебраической системы формализованную модель автоматического классификатора текстовых данных, позволяющую применять методы классификации, применяемые в данной работе, можно описать как кортеж [7]:

$$R = < C, T, F, R_C^F, f >, \quad (1.1)$$

где $C$ – множество тематических рубрик; $T$ – выборка текстов из рубрик $C$; $F$ – множество описаний тематических рубрик; $R_C^F$ – отношение на $C \times F$, соотносящее тематические рубрики и соответствующие им описания; $f$ – операция классификации – отображение $T \to 2^C$, такое что $f(t) = \sigma$, где $t$ – текст из $T$, а $\sigma \in 2^C$ – элемент множества всех подмножеств $C$, т. е. множество тематических рубрик из $C$. Таким образом, отображение $f$ позволяет каждому документу множества $T$ поставить в соответствие некоторую тематическую рубрику из $C$.

Модель автоматического классификатора документов текстовых коллекций на естественном языке представляет собой систему $R$. Построение классификатора подразумевает частичное или полное формирование $C$, $F$, $R_C^F$, $f$ на основе некоторых априорных данных. На практике это означает, что экспертом формируется иерархия тематических рубрик. Описания тематических рубрик могут создаваться вручную, в виде правил соотнесения документов к тематическим рубрикам по определенным признакам или автоматически, на основе методов машинного обучения. Тогда в качестве обучающего множества выступает набор документов, ранее сопоставленных с категориями $T_0$ с помощью экспертных оценок.

Задачу кластеризации в общем случае можно выразить следующим образом [8].

Дано:

1) множество текстов $T = \{t_1, t_2, \ldots, t_N\}$;

2) количество кластеров $K$, предполагаемое или желательное в имеющемся наборе;

3) целевая функция, позволяющая оценить качество кластеризации.

Определяем соответствие $\gamma : T \to \{1, \ldots, K\}$, при котором будет достигаться минимум или максимум (экстремум) целевой функции. Целевая функция определяется в терминах сходства или расстояния между документами. Сходство документов выражается в виде одной из функций тематического сходства или в значениях на одних и тех же осях векторного пространства. Тематическое сходство определяется как косинусное или евклидово расстояния в векторном пространстве.

Получаемые при мониторинге новостные сообщения представляют собой большие объемы текстовых данных в неструктурированном или слабоструктурированном виде. Для обработки таких данных в настоящее время широкое распространение получило направление машинного обучения с применением нейросетевых методов [9]. Преимущество данных методов перед традиционными в следующем [10]:

1) обучаемость и самообучаемость;

2) быстродействие за счет распараллеливания вычислений;

3) устойчивость к шумам во входных данных.

Решение задач классификации и кластеризации текстов, целесообразно проводить с применением этих методов. Первостепенно для обучения системы необходимо произвести разметку имеющейся коллекции текстовых документов $T$ и каждому учебному тексту назначить метку класса, которому он соответствует. Кроме того, нужно определить способ формализации этих

данных, т. е. показать соответствие функции $f$ из множества текстовых документов $T$ в пространстве признаков $X$:

$$f : T \to X, \tag{1.2}$$

где $f$ – функция излечения признаков (feature extraction).

После определения $f$ и построения пространства признаков $X$, каждому тексту из $T$ ставится в соответствие точка из $X$, что дает возможность разделить все точки $X$ на подмножества.

Таким образом, поиск схожих текстовых документов сводится к задаче кластеризации точек из $X$, а распределение текстов по тематическим рубрикам сводиться к задаче классификации точек из $X$ [11].

Формализовано, требуется создать представление $g$ из множества вектор-признаков $X$ во множество меток $L$:

$$g : X \to L. \tag{1.3}$$

В результате, задача обработки текста сводится:

а) к извлечению признаков;

б) распределению пространства признаков на части.

**2. Математическая модель системы**

Математическая модель является математическим аналогом проектируемого объекта и описывает зависимость между исходными данными и искомыми величинами [12]. Её построение позволяет структурировать процессы, протекающие в системе при её функционировании, что даст возможность перейти к натурному физическому построению компонентов системы и обеспечению их взаимодействия. Для этого целесообразно применить методы математического моделирования.

Построение математической модели предполагает следующие этапы [13]:

1) составление описания, как в целом функционирует система;

2) описание, какие подсистемы и элементы входят в систему, как они взаимодействуют между собой, описание их функционирования и характеристик, а также начальных условий;

3) определение, какие внешние факторы перечня могут оказывать влияние на систему;

4) выбор характеристик системы, которые определяют степень ее соответствия заявленным требованиям;

5) составление формального математического описания системы;

6) построение математической модели системы.

Работа по первым четырем пунктам была описана авторами в предыдущих работах [14, 15]. Далее представлен процесс разработки формальной математической модели разрабатываемой системы.

Входными параметрами для работы системы являются информационные текстовые сообщения, получаемые из множества новостных источников в сети Интернет [16, 17].

Обозначим данное множество сообщений, получаемых из одного информационного источника – RSS-канала, как

$$M_n = \{m_1, m_2, \ldots, m_i\}, \tag{2.1}$$

где каждое сообщение $m_i$, представляет собой кортеж, состоящий из идентификатора $i_i$, заголовка $h_i$, основного текста сообщения $t_i$, ссылки на источник $l_i$, даты публикации $d_i$:

$$m_i = < i_i, h_i, t_i, l_i, d_i >; \tag{2.2}$$

множество источников сообщений:

$$L_n = \{l_1, l_2, \ldots, l_i\}; \tag{2.3}$$

множество тематических рубрик сообщений:

$$C_n = \{c_1, c_2, \ldots, c_i\}, \tag{2.4}$$

тогда модель собранной коллекции текстовых документов $T$ будет представлять собой кортеж:

$$T_n = < M_n, L_n, C_n >. \tag{2.5}$$

Совокупность компонентов (модулей и блоков) системы мониторинга и анализа публикаций можно описать в виде:

$$K_n = < B, A, S, V, E >, \quad (2.6)$$

где $B$ – модуль мониторинга с блоками парсинга RSS-каналов $b_{rss}$, сайтов $b_{site}$ и социальных сетей $b_{soc}$:

$$B = \{b_{rss}, b_{site}, b_{soc}\}; \quad (2.7)$$

модуль аналитики $A$ с блоками предварительной обработки текста $a_{pre\text{-}proc}$, извлечения информации $a_{extr}$, text mining $a_{tm}$, обработки естественного языка $a_{nlp}$, нейросетевых моделей $a_{neuro}$:

$$A = \{a_{pre\text{-}proc}, a_{extr}, a_{tm}, a_{nlp}, a_{neuro}\}; \quad (2.8)$$

модуль хранения $S$ с блоками необработанных данных $s_{rd}$, обработанных данных $s_{pd}$, метаданных $s_{md}$, исторических данных $s_{hd}$:

$$S = \{s_{rd}, s_{pd}, s_{md}, s_{hd}\}; \quad (2.9)$$

модуль вывода $V$ с блоками текстового представления $v_{txt}$, табличного $v_{tb}$ и графического $v_g$:

$$V = \{v_{txt}, v_b, v_g\}; \quad (2.10)$$

модуль управления системой $E$ с блоками административного $e_{adm}$, и экспертного уровня $e_{exp}$:

$$E = \{e_{adm}, e_{exp}\}. \quad (2.11)$$

Взаимодействие системы с источниками информации при сборе текстовых данных (процесс мониторинга) можно представить как функцию

$$f : B \to L. \quad (2.12)$$

Процесс анализа текстовых данных можно обозначить как взаимодействие модуля аналитики системы с полученными данными:

$$f : A \to M. \quad (2.13)$$

Процесс хранения данных в хранилище можно показать как

$$f : S \to M_{str}, \quad (2.14)$$

где $M_{str}$ – структурированные данные.

Взаимодействие пользователя с системой можно показать как

$$f : E \to K_n. \quad (2.15)$$

Рассмотрим протекающие ключевые процессы в модулях системы при её функционировании.

Модуль мониторинга $B$ посылает сформированное пользователем через модуль управления $E$ множество запросов $Q_n$ множеству источников $L_n$.

$$f : E \to B \to Q_n \to L_n. \quad (2.16)$$

Результатом запросов является полученное множество сообщений $M_n$, которое передаётся в блок предварительной обработки $a_{pre\text{-}proc}$ модуля аналитики $A$, а затем в блок обработанных данных $s_{pd}$, модуля хранения $S$.

$$f : A \to M_n. \quad (2.17)$$

Полученные данные могут и без предварительной обработки помещаться в блок необработанных данных $s_{rd}$ модуля хранения с целью их накопления.

$$f : S \to M_n. \quad (2.18)$$

На следующем этапе данные выгружаются из блока необработанные данных в модуль аналитики, где проходят предварительную обработку в блоке $a_{pre\text{-}proc}$, затем, исходя из того, какой результат необходимо получить, предаются в блоки извлечения информации $a_{extr}$, text mining $a_{tm}$, обработки естественного языка $a_{nlp}$, нейросетевых моделей $a_{neuro}$. Задействование различных блоков определяется пользователем.

В блоке предварительной обработки $a_{pre\text{-}proc}$ применяются как по отдельности, так и комплексно, следующие методы:

– токенизация – разбивка текста на отдельные токены (абзацы, предложения, слова, символы, знаки пунктуации и т. д.), обозначим этот метод как $f_{token}(m_i)$;

– удаление неинформативных, малоинформативных слов (стоп-слов), обозначим этот метод как $f_{stop}(m_i)$;

– удаление числовых символов – $f_{num}(m_i)$;

– приведение регистра – преобразование всех символов слов к верхнему или нижнему регистру – $f_{registr}(m_i)$;

– стемминг – нахождение основы слов – $f_{stemm}(m_i)$;

– лемматизация – приведения словоформ текста к леммам – нормальной (словарной) форме – $f_{lemm}(m_i)$.

Таким образом, процесс предварительной обработки текста в системе можно показать, как кортеж применяемых методов:

$$f_{pre\text{-}proc}(m_i) = < f_{token}(m_i), f_{stop}(m_i), f_{num}(m_i), f_{registr}(m_i), f_{stemm}(m_i), f_{lemm}(m_i) > . \quad (2.19)$$

В блоке извлечения информации $a_{extr}$ применяются, как по отдельности, так и комплексно, следующие методы:

– метод извлечения именованных сущностей, таких как имена людей, названий организаций, событий и т. п.) – $f_{name}(m_i)$;

– метод разрешение кореференции $f_{co\text{-}ref}(m_i)$ – поиск связей нескольких разных отсылок в тексте к одному реальному объекту;

– метод извлечения фактов (события, мнения, отзывы, объявления, контактные данные и т. п.) – $f_{fact}(m_i)$;

– метод аннотирования текста $f_{abstr}(m_i)$ – преобразование текста с целью получения его краткого описания.

Таким образом, процесс извлечения информации из текста в системе можно показать как кортеж методов:

$$f_{extr}(m_i) = < f_{name}(m_i), f_{co\text{-}ref}(m_i), f_{fact}(m_i), f_{abstr}(m_i) > . \quad (2.20)$$

В блоке методов text mining $a_{tm}$ применяются как по отдельности, так и комплексно, следующие методы:

– классификация (категоризация) – $f_{class}(m_i)$;

– кластеризация – $f_{claster}(m_i)$.

Описанный ранее блок извлечения информации $a_{extr}$ также использует методы, относящиеся к text mining. В разрабатываемой системе целесообразно разделить описанные методы, оставив в блоке $a_{tm}$ только два ключевых – классификацию и кластеризацию, поскольку данные методы будут использоваться и в других блоках, таких как обработка естественного языка $a_{nlp}$ и нейросетевые модели $a_{neuro}$.

Таким образом, процессы блока методов text mining $a_{tm}$ можно показать как

$$f_{tm}(m_i) = < f_{class}(m_i), f_{claster}(m_i) > . \quad (2.21)$$

В блоке методов обработки естественного языка $a_{nlp}$ применяются как по отдельности, так и комплексно, следующие методы:

– векторизация данных методом «Мешок слов» [18] (Bag of words (BOW)) – $f_{BOW}(m_i)$;

– векторизация данных методом TF-IDF [19] – $f_{TF\text{-}IDF}(m_i)$;

– оценка тональности текста (Sentiment Analysis) [20] – $f_{sent}(m_i)$.

Описанный ранее блок предварительной обработки $a_{pre\text{-}proc}$ также использует методы, относящиеся к обработке естественного языка. Разделение этих методов по двум блокам системы обусловлено тем, что предварительная обработка, с выделенными в ней методами, необходима для работы других блоков и модулей.

Таким образом, процессы блока $a_{nlp}$ можно показать как

$$f_{nlp}(m_i) = < f_{BOW}(m_i), f_{TF\text{-}IDF}(m_i), f_{sent}(m_i) > . \quad (2.22)$$

В блоке методов нейросетевых моделей $a_{neuro}$ применяются как по отдельности, так и комплексно, следующие методы на основе машинного обучения:

– классификация (категоризация) – $f_{class\_ML}(m_i)$;

– класстеризация – $f_{claster\_ML}(m_i)$;

– оценка тональности текста – $f_{sent\_ML}(m_i)$.

Таким образом, процессы блока $a_{neuro}$ можно показать как

$$f_{neuro}(m_i) = < f_{class\_ML}(m_i), f_{claster\_ML}(m_i), f_{sent\_ML}(m_i) > . \quad (2.23)$$

**Заключение**

Приведенное формализованное описание базовых процессов, протекающих в разрабатываемой системе мониторинга и анализа информации электронных СМИ и их математическое описание даёт четкое представление о ней и позволяет перейти к следующему этапу разработки – программной реализации системы.

***Литература***

*1. Комаров, В.Н. Мониторинг и системный анализ информации электронных СМИ для промышленных предприятий / В.Н. Комаров, С.М. Рощин // Наукоемкие технологии и интеллектуальные системы: сб. ст. по итогам Междунар. науч.-практ. конф. – Самара: ООО «Агентство международных исследований», 2018. – С. 36–40.*

*2. Комаров, В.Н. Разработка архитектуры системы мониторинга и анализа публикаций в сети интернет / В.Н. Комаров, С.М. Рощин // Передовые инновационные разработки. Перспективы и опыт использования, проблемы внедрения в производство: сб. науч. ст. по итогам девятой междунар. науч. конф. (31 октября 2019 г.). Ч. 2. – Казань: ООО «Конверт», 2019. – С. 27–29.*

*3. Комаров, В.Н. Моделирование системы мониторинга и анализа информации электронных СМИ методами модельно-ориентированного системного инжиниринга / В.Н. Комаров, С.М. Рощин // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2021 – Т. 21, № 1 – С. 12–22. DOI: 10.14529/ctcr210102*

*4. Зайцева, Н.А. Математическое моделирование: учеб. пособие / Н.А. Зайцева. – М.: РУТ (МИИТ), 2017. – 110 с.*

*5. Аверченков, В.И. Мониторинг и системный анализ информации в сети Интернет / В.И. Аверченков, С.М. Рощин. – Брянск: БГТУ, 2012. – 160 с.*

*6. Анализ данных и процессов / А.А. Барсегян, М.С. Куприянов, И.И. Холод и др. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.*

*7. Борисов, Е.С. Классификатор текстов на естественном языке / Е.С. Борисов. – http://mechanoid.kiev.ua/neural-net-classifier-text.html (дата обращения: 5.08.2021).*

*8. Дюк, В.А. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях / В.А. Дюк, А.В. Флегонтов, И.К. Фомина // Известия российского государственного педагогического университета им. А.И. Герцена. – 2011. – № 138. – С. 77–87.*

*9. Акимов, Д.А. Подход к классификации интернет-страниц по степени их информативности / Д.А. Акимов, О.К. Редькин, И.В. Садыков // Вестник МГТУ МИРЭА. – 2015, № 4-1 (9). – С. 206–217.*

*10. Созыкин, А.В. Анализ текстов с помощью рекуррентных нейронных сетей / А.В. Созыкин. – https://www.youtube.com/watch?v=7Tx_cewjhGQ (дата обращения: 05.08.2021).*

*11. Архипенко, К. Рекуррентные нейронные сети в задачах анализа текстов / К. Архипенко. – https://docplayer.ru/42578505-Rekurrentnye-neyronnye-seti-v-zadachah-analiza-tekstov.html (дата обращения: 10.08.2021).*

*12. Трусов, П.В. Введение в математическое моделирование / П.В. Трусов. – М.: Университетская книга; Логос, 2007. – 440 с.*

*13. Русаков, А.М. Исследование и моделирование сложных систем / А.М. Русаков. – М.: Москов. гос. ун-т приборостроения и информатики, 2014. – 90 с.*

*14. Комаров, В.Н. Структура и обобщенный алгоритм работы системы мониторинга и анализа публикаций электронных СМИ / В.Н. Комаров, С.М. Рощин // XXI век: итоги прошлого и проблемы настоящего плюс. – 2019. – Т. 8, № 4 (48). – С. 61–66.*

*15. Комаров, В.Н. Мониторинг и системный анализ информации электронных СМИ для предприятий оборонно-промышленного комплекса России / В.Н. Комаров, С.М. Рощин // XXI век: итоги прошлого и проблемы настоящего плюс. – 2019. – Т. 8, № 2 (46). – С. 22–25.*

*16. Борисов, Е.С. Автоматизированная обработка текстов на естественном языке, с использованием инструментов языка Python / Е.С. Борисов. – http://mechanoid.kiev.ua/ml-text-proc.html (дата обращения: 10.08.2021).*

*17. Васильев, Ю. Обработка естественного языка. Python и spaCy на практике / Ю. Васильев. – СПб.: Питер, 2021. – 256 с.*

*18. Система формирования знаний в среде интернет: моногр. / В.И. Аверченков, А.В. Заболеева-Зотова, Ю.М. Казаков и др. – 3-е изд., стереотип. – М.: ФЛИНТА, 2016. – 181 с.*

*19. Вершинин, В.Е. Решение задач обработки естественного языка на основе нейросетевых моделей / В.Е. Вершинин, Е.В. Вершинин // Международная научно-практическая конференция НИЦ Аэтерна, 2018. – С. 54–59.*

*20. Витковский, А.В. Применение рекурсивных нейронных сетей для анализа тональности текста / А.В. Витковский, А.В. Жвакина // 54-я научная конференция аспирантов, магистрантов и студентов БГУИР, 2018. – С. 152–153.*

**Комаров Виталий Николаевич**, преподаватель, Уральский колледж строительства, архитектуры и предпринимательства, г. Екатеринбург; komaroffvn@mail.ru.

**Рощин Сергей Михайлович**, канд. техн. наук, доцент кафедры информационных технологий, Брянский государственный инженерно-технологический университет, г. Брянск; roschinsm@ya.ru.

***Поступила в редакцию 15 августа 2021 г.***

---

**DOI: 10.14529/ctcr210403**

# FORMALIZATION OF BASIC PROCESSES AND MATHEMATICAL MODEL OF THE SYSTEM FOR MONITORING AND ANALYSIS OF PUBLICATIONS OF ELECTRONIC MEDIA

***V.N. Komarov*[1], *komaroffvn@mail.ru*,**
***S.M. Roschin*[2], *roschinsm@ya.ru***

[1] *Ural College of Construction, Architecture and Business, Ekaterinburg, Russian Federation,*
[2] *Bryansk State Engineering Technological University, Bryansk, Russian Federation*

The article describes an approach to formalizing basic processes and building a mathematical model for a system for collecting and analyzing data from electronic media. The authors, as part of a scientific study, are creating a system, including the development of new algorithms, methods and approaches for collecting and analyzing textual information from Internet news sources. The main direction of the study is the application of methods for the mining of text data based on the technology of artificial neural networks, methods of natural language processing, text mining, machine learning and big data processing. **Purpose of the study.** To develop a formalized description of the model of the system for monitoring and analyzing the text information of electronic news media using the methods of mathematical modeling. **Research methods and tools.** The use of the toolkit of the methodology of mathematical modeling, with the methods of system analysis is proposed. To study the system, such methods of system analysis as abstraction, formalization, composition and decomposition, structuring and restructuring, modeling, recognition and identification were used. The system is considered as a formalized model of an automatic classifier and clusterizer for a set of text documents in a natural language in the form of an algebraic system. To solve the problems of classi-

fication and clustering of texts, it is proposed to apply machine learning methods based on neural network approaches. The structure of the system and its constituent processes, as well as processes interacting with the system from outside, are presented in the form of a formalized mathematical description. **Results.** The developed formalized mathematical description of the system model clearly shows the interconnection of the system components with each other, as well as internal processes. The applied approach makes it possible to detail the representation of the system based on its decomposition into subsystems and modules. All this makes it possible to streamline the sequence of stages of creating a system and decompose them into separate stages of work. **Conclusion.** The results obtained in the course of the study allow us to move on to the next stage of the life cycle of the information system being developed - its software development.

*Keywords: media information monitoring, data analysis, monitoring and data analysis system, text analysis, mathematical model of the system, data mining, neural network methods, system analysis, text classification, text clustering.*

## References

1. Komarov V.N., Roshchin S.M. [Monitoring and system analysis of electronic media information for industrial enterprises]. *Naukoemkie tekhnologii i intellektual'nye sistemy* [Science-intensive technologies and intelligent systems]. Samara, LLC “Agency for International Studies” Publ., 2018, pp. 36–40. (in Russ.)

2. Komarov V.N., Roshchin S.M. [Development of the architecture of the system for monitoring and analyzing publications on the Internet]. *Peredovyye innovatsionnyye razrabotki. Perspektivy i opyt ispol'zovaniya, problemy vnedreniya v proizvodstvo* [Advanced innovative developments. Prospects and experience of use, problems of implementation in production]. Kazan, LLC “Convert” Publ., 2019, pp. 27–29. (in Russ.)

3. Komarov V.N., Roshchin S.M. Modeling of the System of Monitoring and Analysis of Information of Electronic Media by Methods of Model Based System Engineering. Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics, 2021, vol. 21, no. 1, pp. 12–22. (in Russ.) DOI: 10.14529/ctcr210102

4. Zaitseva N.A. *Matematicheskoe modelirovanie* [Mathematical modeling]. Moscow, RUT (MIIT), 2017. 110 p.

5. Averchenkov V.I., Roshchin S.M. *Monitoring i sistemnyy analiz informatsii v seti Internet* [Monitoring and system analysis of information on the Internet]. Bryansk, BSTU Publ., 2012. 160 p.

6. Barsegyan A.A., Kupriyanov M.S., Holod I.I., Tess M.D., Elizarov S.I. *Analiz dannykh i protsessov* [Data and process analysis]. St. Petersburg, BHV-Peterburg Publ., 2009. 512 p.

7. Borisov E.S. *Klassifikator tekstov na estestvennom yazyke* [Classifier of texts in natural language]. Available at: http://mechanoid.kiev.ua/neural-net-classifier-text.html (accessed 08.05.2021).

8. Dyuk V.A., Flegontov A.V., Fomina I.K. [Application of data mining technologies in natural science, technical and humanitarian fields]. Izvestia: Herzen University Journal of Humanities & Sciences, 2011, no. 138, pp. 77–87. (in Russ.)

9. Akimov D.A., Redkin O.K., Sadykov I.V. The approach to the web pages classification based on their informativity. *Bulletin of MSTU MIREA*, 2015, no. 4-1 (9), pp. 206–217. (in Russ.)

10. Sozykin A.V. *Analiz tekstov s pomoshch'yu rekurrentnykh neyronnykh setey* [Analysis of texts using recurrent neural networks]. Available at: https://www.youtube.com/wat·sh?v=7Tx_cewjhGQ (accessed 08.05.2021).

11. Arkhipenko K. *Rekurrentnyye neyronnyye seti v zadachakh analiza tekstov* [Recurrent neural networks in text analysis problems]. Available at: https://docplayer.ru/42578505-Rekurrentnye-neyronnye-seti-v-zadachah-analiza-tekstov.html (accessed 08.10.2021).

12. Trusov P.V. *Vvedeniye v matematicheskoye modelirovaniye* [Introduction to mathematical modeling]. Moscow, University book, Logos, 2007. 440 p.

13. Rusakov A.M. *Issledovaniye i modelirovaniye slozhnykh sistem* [Research and modeling of complex systems]. Moscow, Moscow State University of Instrument Engineering and Informatics, 2014. 90 p.

14. Komarov V.N., Roshchin S.M. Structure and generalized algorithm of the system of monitoring and analysis of electronic media publications. *XXI century: the results of the past and the problems of the present plus*, 2019, vol. 8, no. 4 (48), pp. 61–66. (in Russ.)

15. Komarov V.N., Roshchin S.M. Monitoring and system analysis of information electronic media for enterprises of the military-industrial complex of Russia. *XXI century: the results of the past and the problems of the present plus*, 2019, vol. 8, no. 2 (46), pp. 22–25. (in Russ.)

16. Borisov E.S. *Avtomatizirovannaya obrabotka tekstov na estestvennom yazyke, s ispol'zovaniyem instrumentov yazyka Python* [Automated processing of texts in natural language using Python tools]. Available at: http://mechanoid.kiev.ua/ml-text-proc.html (accessed 08.10.2021).

17. Vasil'yev Yu. *Obrabotka estestvennogo yazyka. Python i spaCy na praktike* [Natural language processing. Python and spaCy in practice]. St. Petersburg, Piter Publ., 2021. 256 p.

18. Averchenkov V.I., Zaboleyeva-Zotova A.V., Kazakov Yu.M., Leonov E.A., Roshchin S.M. *Sistema formirovaniya znaniy v srede internet* [The system of knowledge formation in the Internet environment]. Moscow, FLINTA Publ., 2016. 181 p.

19. Vershinin V.E., Vershinin E.V. [Solving natural language processing problems based on neural network models]. *International Scientific and Practical Conference of SIC Aeterna*, 2018, pp. 54–59. (in Russ.)

20. Vitkovsky A.V., Zhvakina A.V. [Application of recursive neural networks for text sentiment analysis]. *54th scientific conference of graduate students, undergraduates and students of BSUIR*, 2018, pp. 152–153. (in Russ.)

***Received 15 August 2021***

**ОБРАЗЕЦ ЦИТИРОВАНИЯ**

Комаров, В.Н. Формализация базовых процессов и математическая модель системы мониторинга и анализа публикаций электронных СМИ / В.Н. Комаров, С.М. Рощин // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2021. – Т. 21, № 4. – С. 28–36. DOI: 10.14529/ctcr210403

**FOR CITATION**

Komarov V.N., Roschin S.M. Formalization of Basic Processes and Mathematical Model of the System for Monitoring and Analysis of Publications of Electronic Media. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2021, vol. 21, no. 4, pp. 28–36. (in Russ.) DOI: 10.14529/ctcr210403