ISSN 2658-5014 (Print)
2022. Т. 4, № 1 (8). С. 37–43
СИИТ
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
ISSN 2686-7044 (Online)
http://siit.ugatu.su

УДК 004.82
doi 10.54708/26585014_2022_41837

СТРУКТУРНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ НАУЧНЫХ ПУБЛИКАЦИЙ ВЫДЕЛЕННОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

М. М. Гаянова [1], **А. М. Вульфин** [2]

[1] gayanova.mm@ugatu.su, [2] vulfin.am@ugatu.su

ФГБОУ ВО «Уфимский государственный авиационный технический университет» (УГАТУ)

Поступила в редакцию 17 января 2022 г.

Аннотация. Рассматриваются вопросы структурно-семантического анализа научных текстов; разработки системы для формирования корпуса текстов и организации хранения на основе единой информационной платформы. Платформа предназначена для широкого круга ученых, и ориентирована на реализацию не только поверхностного анализа (временной, географический), но и глубокого содержательного анализа, в результате которого могут быть выявлены наилучшие результаты в рассматриваемых предметных областях. Представлена структурно-функциональная организация платформы анализа и иерархическая организация корпуса научных текстов. Рассмотрена организация поиска и извлечения текстовых документов с тематических ресурсов в задаче построения корпуса текстов. Проанализированы требования к платформе о поддержке методов интеллектуального анализа с целью автоматизации процесса структурирования накопленных данных, выявления скрытых закономерностей и построения базы знаний предметной области. Рассмотрен пример ручного анализа научных текстов по подборке «Вихревые электромагнитные поля в инфокоммуникационных системах».

Ключевые слова: семантический анализ; обработка слабоструктурированных данных; WORM; корпус текстов; платформа семантического анализа; Text Mining; интеллектуальный анализ данных; нейросетевая модель; извлечение именованных сущностей; структурно-функциональная схема.

ВВЕДЕНИЕ

В различных предметных областях существует острая необходимость повышения эффективности структурно-семантического анализа доступных научных текстов, их структурирования, выявления скрытых закономерностей за счет применения инструментов интеллектуального анализа больших объемов научных текстов (корпуса) и построения базы знаний предметной области.

В статье рассматриваются вопросы структурно-семантического анализа научных текстов (статьи, диссертации, монографии, учебные пособия); разработки корпусов текстов и организации хранения на основе единой информационной платформы. Платформа предназначена для широкого круга ученых, и позволяет выполнить не только поверхностный анализ (временной, географический), но и глубокий содержательный, в результате которого могут быть выявлены наилучшие результаты в рассматриваемых предметных областях.

– Структурно-семантический анализ текстовых документов.

– Семантический анализ представляет собой механизм автоматического понимания

текстов, включающий формирование перечня семантических отношений между выделяемыми сущностями и построение семантического представления текстов. Семантическое представление текста может быть формализовано в виде графа, отражающего бинарные отношения между смысловыми единицами текста. Глубина семантического анализа текстов может быть различной и зависит от конкретных прикладных задач [1].

– Структурно-семантический анализ (ССА) является результатом фактографического анализа первичных документов, научного анализа изложенных в них положений – фактов и концепций и позволяет ориентироваться в информационном потоке по данному направлению (проблеме) и давать оценку состояния проблемы, выявляя тенденции ее развития [2, 3]. Повышение эффективности ССА возможно за счет автоматизации процесса сбора, обработки, хранения и последующего интеллектуального анализа накапливаемых данных на основе методов машинного обучения.

ССА текстовых документов включает [4] ряд основных операций, направленных на построение перечня ключевых слов и словосочетаний и формирования семантических отношений между ними.

Обобщенная функциональная модель ССА представлена на рис. 1.

ССА включает процесс выделения ключевых слов, с последующим выявлением их смыслового содержания. Ключевые слова представляют собой лексические единицы (слова и словосочетания), передающие основной смысл тематики документа.

Массив ключевых слов должен быть упорядочен по любому признаку, не противоречащему логике изучения темы: от общего к частному, от простого к сложному, поаспектно, по логике развития темы и т.д.

Выделенные и упорядоченные ключевые слова выполняют две функции:

1) определяют логическое построение содержания;

2) являются входами в информационные массивы при поиске литературы.

На рис. 1. введены следующие обозначения:

① – для каждого слова w_i в W строится упорядоченное по убыванию приоритетности для текущего контекста множество смысловых значений $S = \{s_1, s_2, \ldots, s_p\}$;

② – для каждой пары w_i и w_j близких по смыслу слов строится оценка меры их семантической близости d, которая сравнивается с заранее заданным экспертом пороговым значением Θ. При выявлении близких по смыслу слов происходит уточнение их смысловых значений: $d(w_i, w_j) < \Theta$ $s_i \cap s_j$. В результате для каждого слова и словосочетания формируется множество S, характеризующее их уточненный смысл;

③ – эксперт предметной области;

④ – специалист по ССА текстов на естественном языке;

⑤ – нормативно-справочные документы и материалы;

⑥ – лексические единицы, которые передают основной смысл тематики документов. $D = \{d_1, d_2, d_m\}$ – множество документов; для $\forall d \in D \; \exists W$: для каждого документа d_k строится множество W ключевых слов и словосочетаний;

⑦ – информационный массив для поиска литературы;

⑧ – инструменты разметки и аннотирования текстовой документации;

⑨ – список кортежей «документ-ключевые слова и словосочетания документа» (d, W);

⑩ – список кортежей «документ-смысловое описание ключевых слов и словосочетаний» (d, S);

⑪ – критерий упорядочивания.

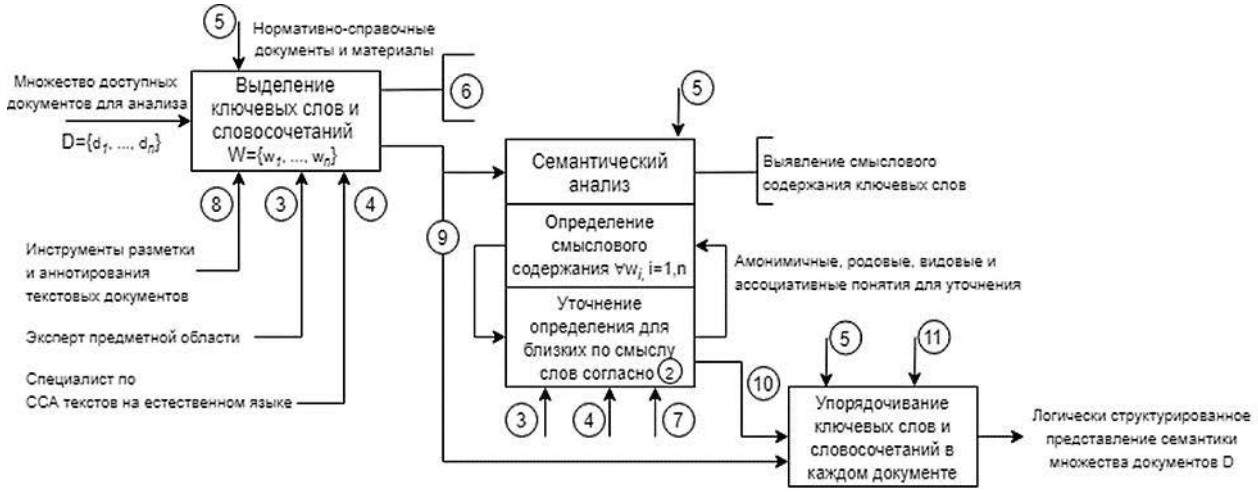

Рис. 1. Функциональная модель структурно-семантического анализа текстовых данных

Fig. 1. Functional model of structural-semantic analysis of text data

СОЗДАНИЕ КОРПУСА СПЕЦИАЛИЗИРОВАННЫХ ТЕКСТОВ

Корпус – это коллекция взаимосвязанных документов (текстов) на естественном языке. Иерархия структурной организации корпуса представлена в табл. 1.

Таблица 1

Иерархическая структурная организация корпуса

Hierarchical structural organization of the corpus

Уровень семантического анализа корпуса	Описание
Категории документов или отдельные документы	Набор документов разного размера и тематической направленности
Абзацы документа	Смысловые единицы речи, выражающие одну идею
Предложения	Синтаксические единицы
Слова и знаки препинания	Лексические единицы
Символы	Единицы, имеющие смысл, при объединении в слова

Наиболее удачные модели естественного языка часто являются узкоспециализированными для конкретного применения. Корпус, специализированный для конкретной области, анализируется и моделируется точнее, чем корпус с документами из разных областей.

ОРГАНИЗАЦИЯ ПОИСКА И ИЗВЛЕЧЕНИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ С ТЕМАТИЧЕСКИХ РЕСУРСОВ ДЛЯ ПОСТРОЕНИЯ КОРПУСА ТЕКСТОВ

Организация поиска и извлечения текстовых документов с тематических ресурсов в задаче построения корпуса текстов основана на [5–8] и представлена на рис. 2.

В схеме организации поиска выделены следующие компоненты:

① – структурированное информационное хранилище;

② – неструктурированные данные в WWW в виде электронных информационных ресурсов;

③ – парсеры и грабберы (специализированные программные модули);

④ – применение API для получения метаинформации;

⑤ – запрос на получение метаинформации о документах;

⑥ – метаинформация о документах;

⑦ – множество специфических парсеров для выделенных электронных информационных ресурсов;

⑧ – агрегатор результатов сбора данных;

⑨ – автоматизированный поисковой модуль (web crawlers + spiders).

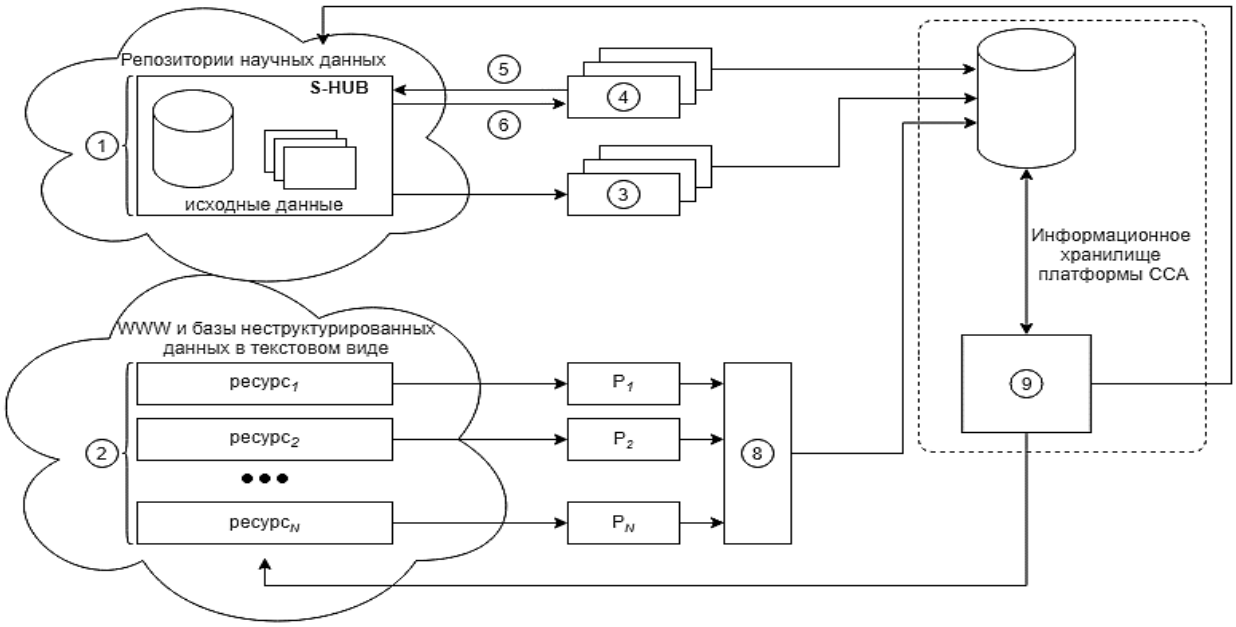

Рис. 2. Схема организации поиска и извлечения текстовых документов с тематических ресурсов

Fig. 2. Scheme of organizing the search and extraction of text documents from thematic resources

ОРГАНИЗАЦИЯ ХРАНЕНИЯ КОРПУСА ТЕКСТОВЫХ ДОКУМЕНТОВ

Приложения данных часто используют архивные хранилища вида «записал один раз, прочитал много раз» (Write-Once Read-Many, WORM) в качестве промежуточного уровня управления данными между механизмами сбора и предварительной обработки данных. Архивные хранилища WORM (иногда их называют «озерами данных») поддерживают потоковый доступ для чтения исходных данных повторяемым и масштабируемым способом, стремясь удовлетворить требование высокой производительности. Кроме того, сохраняя данные в хранилище WORM, можно повторно выполнить предварительную обработку без повторного извлечения данных источников, что позволяет легко проверять новые гипотезы в отношении исходных, необработанных данных.

Для управления данными наиболее часто применяются документно-ориентированные базы данных NoSQL, поддерживающие потоковое чтение документов с минимальными накладными расходами.

СУЩЕСТВУЮЩИЕ ПЛАТФОРМЫ СЕМАНТИЧЕСКОГО АНАЛИЗА

Подобный подход [9] к извлечению информации из семантической сети (IE) реализован в платформе для семантического индексирования, аннотации и поиска. Он сочетает в себе семантическую сеть, основанную на платформе обработки текста (GATE), с представлением знаний и управлением. Ключевым элементом системы является автоматическое создание аннотаций именованных сущностей (NE) со ссылками на классы и экземпляры семантического репозитория.

Разработана и используется упрощенная онтология верхнего уровня, обеспечивающая подробное описание наиболее популярных типов сущностей (более 250 классов). База знаний (БЗ) с фактическим исчерпывающим охватом реальных сущностей общего значения поддерживается, используется и постоянно пополняется.

Распознавание отношений идентичности между объектами используется для унификации их ссылок на базу знаний. В результате последнего, база знаний обогащается признанными отношениями между сущностями. На заключительном этапе процесса IE ранее неизвестные псевдонимы и объекты добавляются в базу знаний с их конкретными типами.

Платформа ССА должна включать поддержку методов интеллектуального анализа с целью автоматизации процесса структурирования накопленных данных, выявления

скрытых закономерностей и построения базы знаний предметной области.

Представленные решения по построению платформы семантического анализа не в полной мере реализуют механизмы сбора и подготовки массива данных с различных источников в сети интернет. Также открытым остается вопрос о перечне технологий интеллектуального анализа, необходимых для построения естественно-языковых моделей, например, на основе нейросетевых решений.

ПРИМЕР АНАЛИЗА НАУЧНЫХ ТЕКСТОВ ПО ПОДБОРКЕ «ВИХРЕВЫЕ ЭЛЕКТРОМАГНИТНЫЕ ПОЛЯ В ИНФОКОММУНИКАЦИОННЫХ СИСТЕМАХ»

Существуют доступные библиотеки научных публикаций (репозитории) (Google Scholar, eLIBRARY.RU, CiteSeer, arXiv.org, Scirus, Scopus и т.д.), позволяющие получать с помощью программного интерфейса API сами публикации и метаинформацию о публикации.

Ниже представлен анализ публикаций по подборке «Вихревые электромагнитные поля в инфокоммуникационных системах» инструментами анализа публикаций eLIBRARY.RU. Общая информация: всего публикаций в библиотеке по данной теме – 273, из них статей в журналах – 9, статей в журналах, входящих в Web of Science или Scopus – 0, входящих в ядро РИНЦ – 2, входящих в RSCI – 2. Взвешенный импакт-фактор журналов, в которых были опубликованы статьи – 0,393. Общее число авторов – 702, среднее число публикаций в расчете на одного автора – 0,39. Суммарное число цитирований – 297, среднее число цитирований в расчете на одну статью – 1,09. Число статей, процитированных хотя бы один раз – 56, число самоцитирований – 0. Индекс Хирша – 6.

Далее построены 6 видов статистических отчетов: распределение публикаций из подборки по годам, по авторам, по организациям, по журналам, по тематике и по ключевым словам.

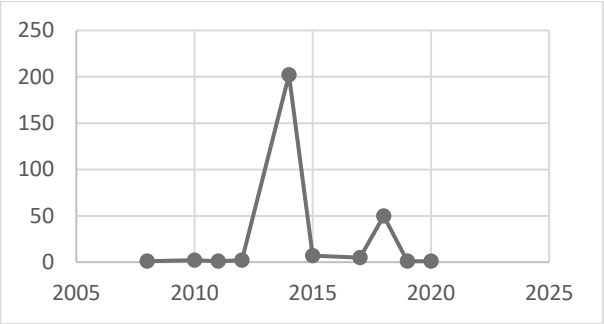

Рис. 3. Распределение статей по годам в области знаний «Вихревые электромагнитные поля в инфокоммуникационных системах»

Fig. 3. Distribution of articles by years in the field of knowledge "Vortex electromagnetic fields in infocommunication systems"

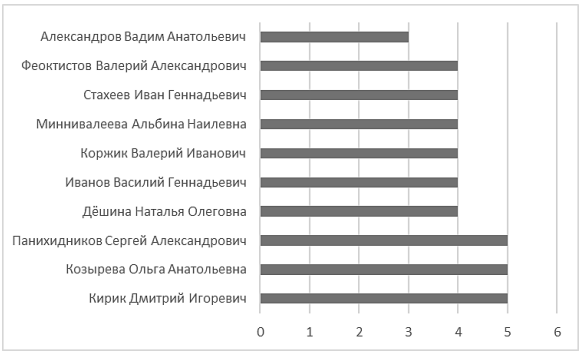

Рис. 4. Распределение статей по авторам в области знаний «Вихревые электромагнитные поля в инфокоммуникационных системах»

Fig. 4. Distribution of articles by authors in the field of knowledge "Vortex electromagnetic fields in infocommunication systems"

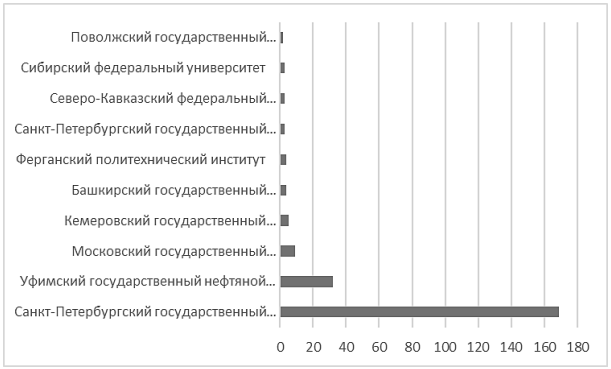

Рис. 5. Распределение статей по организациям в области знаний «Вихревые электромагнитные поля в инфокоммуникационных системах»

Fig. 5. Distribution of articles by organizations in the field of knowledge "Vortex electromagnetic fields in info-communication systems"

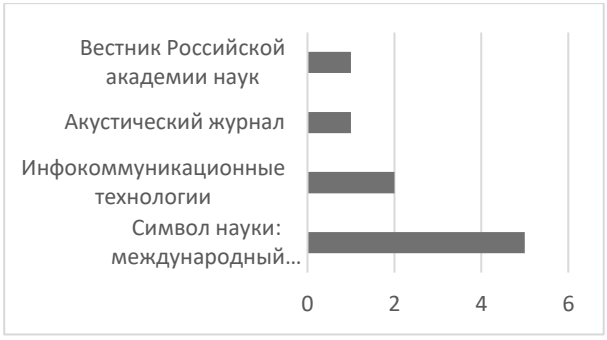

Рис. 6. Распределение статей по журналам в области знаний «Вихревые электромагнитные поля в инфокоммуникационных системах»

Fig. 6. Distribution of articles by journals in the field of knowledge "Vortex electromagnetic fields in infocommunication systems"

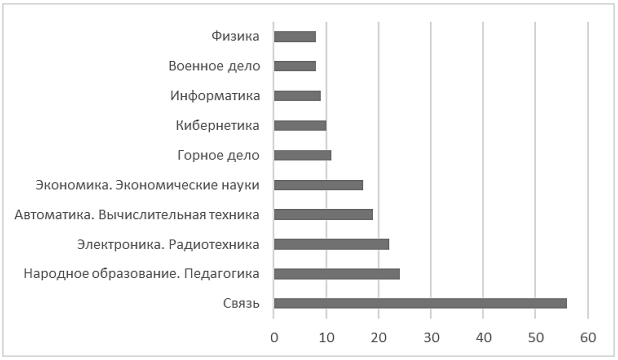

Рис. 7. Распределение статей по журналам в области знаний «Вихревые электромагнитные поля в инфокоммуникационных системах»

Fig. 7. Distribution of articles by journals in the field of knowledge "Vortex electromagnetic fields in infocommunication systems"

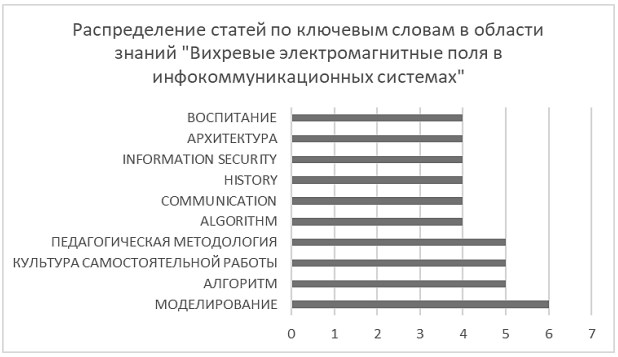

Рис. 8. Распределение статей по журналам в области знаний «Вихревые электромагнитные поля в инфокоммуникационных системах»

Fig. 8. Distribution of articles by journals in the field of knowledge "Vortex electromagnetic fields in infocommunication systems"

Таким образом, мы получили 6 видов отчетов о публикациях по данной тематике, позволяющих сделать определенные выводы о состоянии вопроса. Анализ публикаций на основе данных платформы elibrary.ru является предварительным этапом изучения научных текстов по выбранной тематике исследований.

ЗАКЛЮЧЕНИЕ

Процесс «ручного» анализа массивов текстовых документов очень трудоемок и основан далеко не на всех потенциально доступных в сети данных. Следовательно, необходима автоматизация процесса сбора, подготовки и последующего анализа за счет создания единой платформы в виде надстройки над существующими WORM хранилищами в сочетании с механизмами автоматизации структурно-семантического анализа на основе методов и инструментов интеллектуального анализа и машинного обучения.

СПИСОК ЛИТЕРАТУРЫ

1. **Липницкий С. Ф.** Семантический анализ текста на основе ситуативно-синтагматической сети // Информатика. 2019. № 2 (6). С. 102–110. **[** S. F. Lipnitskiyi, “Semantic text analysis based on situational-syntagmatic network”, (in Russian), in *Informatika,* no. 2 (6), pp. 102-110, 2019. **]**

2. **Отделение** ГПНТБ СО РАН. Структурно-семантический анализ темы обзора. [Электронный ресурс]. URL: http://www.spsl.nsc.ru/ (дата обращения 10.04.2022). **[** Branch of the State Public Scientific and Technical Library of the Siberian Branch of the Russian Academy of Sciences. Structural and semantic analysis of the review topic (2022, Apr. 10). [Online]. Available: http://www.spsl.nsc.ru/ **]**

3. **Chowdhary K. R.** Natural language processing // Fundamentals of artificial intelligence. 2020. Pp. 603-649.

4. **Cambria E., White B.** Jumping NLP curves: A review of natural language processing research // IEEE Computational intelligence magazine. 2014. Vol. 9, No. 2. Pp. 48-57.

5. **Bengfort B., Bilbro R., Ojeda T.** Applied text analysis with python: Enabling language-aware data products with machine learning. Sebastopol, CA: O'Reilly Media, Inc., 2018. 332 p.

6. **Юсупова Н. И., Шахмаметова Г. Р.** Интеграция инновационных информационных технологий: теория и практика // Вестник УГАТУ. 2010. Т. 14, № 4 (39). С. 112–118. **[** N. I. Yusupova, G. R. Shahmametova, “Integration of innovative information technologies: theory and practice”, (in Russian), in *Vestnik UGATU*, vol. 14, no. 4 (39), pp. 112-118, 2010. **]**

7. **Goldberg Y.** Neural network methods for natural language processing // Synthesis lectures on human language technologies. 2017. Vol. 10, No. 1. Pp. 1-309.

8. **Razno M.** Machine learning text classification model with NLP approach // Computational Linguistics and Intelligent Systems. 2019. Vol. 2. Pp. 71-73.

9. **KIM** - a semantic platform for information extraction and retrieval / B. Popov, et al. // Natural language engineering. 2004. Vol. 10, Iss. 3-4. Pp. 375-392.

ОБ АВТОРАХ

ГАЯНОВА Майя Марсовна, доц. каф. вычислительной математики и инженерной кибернетики. Дипл. математик (БГУ, 1997). Канд. техн. наук по упр. в соц.-техн. системах (УГАТУ, 2006). Иссл. в обл. анализа естественного языка.

ВУЛЬФИН Алексей Михайлович, доц. каф. вычислительной техники и защиты информации. Дипл. инженер-программист (УГНТУ, 2008). Канд. техн. наук по системному анализу (УГАТУ, 2012). Иссл. в обл. интеллектуального анализа данных.

METADATA

Title: Structural and semantic analysis of scientific publications in a selected subject area.

Authors: M. M. Gayanova [1], A. M. Vulfin [2]

Affiliation: Ufa State Aviation Technical University (UGATU), Russia.

Email: [1] gayanova.mm@ugatu.su, [2] vulfin.am@ugatu.su

Language: Russian.

Source: SIIT (scientific journal of Ufa State Aviation Technical University), vol. 4, no. 1 (8), pp. 37-43, 2022. ISSN 2686-7044 (Online), ISSN 2658-5014 (Print).

Abstract: The issues of structural-semantic analysis of scientific texts are considered; development of a system for forming a corpus of texts and organizing storage on the basis of a single information platform. The platform is intended for a wide range of scientists, and is focused on the implementation of not only a superficial analysis (temporal, geographical), but also a deep meaningful analysis, as a result of which the best results in the subject areas under consideration can be revealed. The structural and functional organization of the analysis platform and the hierarchical organization of the corpus of scientific texts are presented. The organization of search and extraction of text documents from thematic resources in the task of constructing a text corpus is considered. The requirements for the platform to support the methods of intellectual analysis in order to automate the process of structuring the accumulated data, identifying hidden patterns and building a knowledge base of the subject area are analyzed. An example of manual analysis of scientific texts on the selection "Vortex electromagnetic fields in infocommunication systems" is considered.

Key words: semantic analysis; processing of semi-structured data; WORM; corpus of texts; semantic analysis platform; text mining; data mining; neural network model; extracting named entities; block diagram.

About authors:

GAYANOVA, Maya Marsovna, Assoc. Prof., Dept. of Computational Mathematics and Engineering Cybernetics. mathematic dipl. (BSU, 1997). Cand. of Tech. Sci. of manag. in social and econ. systems (USATU, 2006).

VULFIN, Alexey Mikhailovich, Assoc. Prof., Dept. of computing equipment and information protection. Software engineer dipl. (USPTU, 2008). Cand. of Tech. Sci., systems analyst, councils recommend measure and information processing (USATU, 2012).