A

УДК 681.518.9; 621.384.3 **DOI** 10.34757/2413-7383.2023.30.3.002

С. С. Анцыферов, К. Н. Фазилова, К. Е. Русанов
МИРЭА – Российский технологический университет, г. Москва, Россия
119454, Россия, г. Москва, пр. Вернадского, 78

ПРИНЦИПЫ СТРУКТУРНОГО ПОСТРОЕНИЯ СИСТЕМ «ДОКУМЕНТАЛЬНЫЙ ИНФОРМАЦИОННЫЙ ПОТОК»

S. S. Antsyferov, K. N. Fazilova, K. E. Rusanov
MIREA – Russian Technological University, c. Moscow, Russia
Russia, 119454, c. Moscow, Vernadsky ave., 78

«DOCUMENTARY INFORMATION FLOW» SYSTEMS STRUCTURAL CONSTRUCTION PRINCIPLES

В статье определены принципы построения и функционирования системы документальных информационных потоков. Представлены основные этапы развития научного направления, которые находят свое отражение в документальном информационном потоке. Указано на взаимосвязь принципов построения системы «документальный информационный поток» с характером этапов развития научного направления и что принципы структурного построения системы «документальный информационный поток» могут быть использованы при создании экспертных систем.
Ключевые слова: документальные информационные потоки, информационная база, интеллектуальные системы управления, входные данные, блоки семантических шаблонов, информация, контроль.

The article defines the principles of construction and functioning of the system of documentary information flows. The main stages of the development of the scientific direction, which are reflected in the documentary information flow, are presented. It is pointed out that the principles of the construction of the "documentary information flow" system are interrelated with the nature of the stages of the development of the scientific direction and that the principles of the structural construction of the "documentary information flow" system can be used in the creation of expert systems.
Keywords: documentary information flows, information base, intelligent control systems, input data, semantic template blocks, information, control.

Введение

На современном этапе научно-технического развития индустриального – развитого общества большое значение приобретает задача формирования информационной базы по каждому научно-тематическому направлению. Важность данной задачи определяется тем, что сформированная информационная база может служить основой для создания интеллектуальных систем управления и обработки информации, а также для прогнозирования возможных направлений и тенденций научного развития [1-21].

Любое научное направление проходит определенные этапы своего развития. Как правило, выделяются 4 основных этапа: зарождение, формирование, эволюционное развитие, деградация. Каждому направлению соответствует свой документальный информационный поток (такие публикации как статьи, патенты, отчеты, конференции, учебные пособия, диссертации), позволяющий судить как о текущем состоянии тематического направления, так и прогнозировать его последующее состояние. В свою очередь, документальный информационный поток (ДИП) может рассматриваться как некоторая самоорганизующаяся система, то есть как динамическая совокупность взаимосвязанных информационных документов, содержащих закрепленную научную информацию, предназначенную для передачи во времени и пространстве.

Цель работы – определение принципов построения и функционирования системы ДИП.

Принципы построения системы ДИП

Принципы построения данной системы во многом определяются характером этапов развития научного направления, которые находят свое отражение в документальном информационном потоке (рис. 1).

Так характерной особенностью этапа зарождения научного направления является небольшое число публикаций и высокая степень расселения информации.

Этап формирования характеризуется резким увеличением числа публикаций, содержащих информацию по развивающемуся направлению.

Эволюционный этап отличается тем, что число публикаций приближается к максимальному значению (максимальная продуктивность).

Этап деградации — это период насыщения, когда область основных идей исчерпана, что приводит к уменьшению используемости опубликованных работ и, соответственно, к уменьшению числа публикаций по данному направлению (рис. 1).

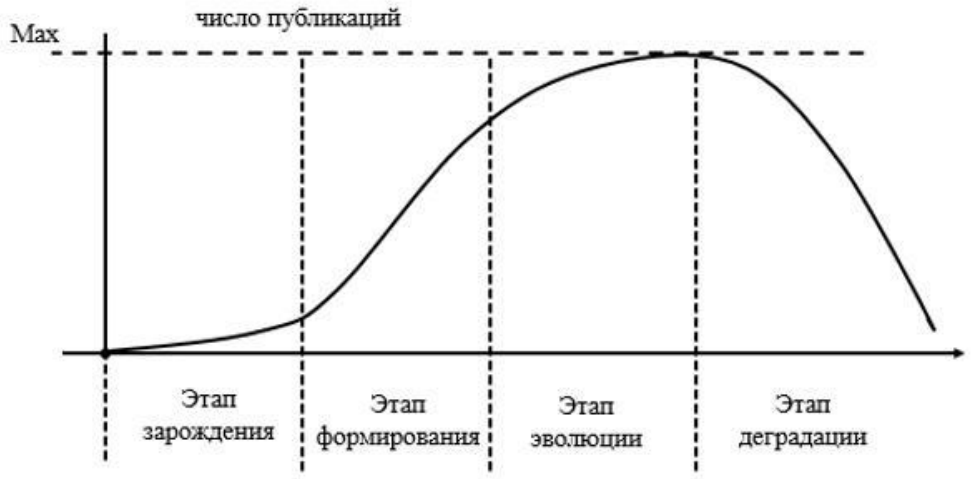

Рисунок 1 – Этапы развития научного направления

Признаки ДИП, которые можно разделить на первичные – смысловое содержание информации, и вторичные – классификационные индексы, термины, авторы, принадлежность к специализированным изданиям и др. Исходя из отмеченных этапов и учитывая указанные признаки, построение системы ДИП определяющее ее структуру на каждом этапе, должно включать блок адаптивного группирования входных данных по вторичным признакам, блоки формирования семантических шаблонов, сравнения и текущих семантических образов для каждой из ранее полученных групп, блок установления аналогии между шаблоном и текущим образом.

Отметим, что функционирование блока адаптивного группирования может происходить в соответствии с одним из известных алгоритмов (ПСОМАД, Мак - Куина, Джонсона и другие). Здесь отдельное внимание можно обратить на теоретико-графовый подход, который эффективен в случаях, когда входные данные характеризуются не числовыми оценками, а отношениями.

Функционирование блоков семантической обработки может происходить с использованием известных семантических стратегий (например, программа GRIN1), связанных с формированием семантических шаблонов, а также алгоритмов установления аналогии между семантическим шаблоном и семантическим образом текущих данных (например, программа ZOBRA).

Результатом действия этих блоков является выделенное число информативных документов d_{ik} из общего числа документов D_{iN} с высокой степенью вероятности, соответствующее тематике данного (i-го) научного направления.

Для отслеживания и прогнозирования динамики развития научного направления (динамики системы ДИП) важно иметь следующую статистическую информацию:

* приращение числа документов (публикаций) в системе Δd (индекс i опущен) за время наблюдения $\Delta t=t_j-t_{j-1}$

$\Delta d=d_j-d_{j-1}, j=1,2\ldots$

* d_j - число публикаций в момент t_j;

* d_o - число публикаций в момент начала наблюдения;

* ΔB - приращение числа использований документов за время Δt;

Данная информация позволяет определить такие показатели динамики функционирования системы ДИП как интенсивность роста числа публикаций

$$\Delta I(t)=\frac{\Delta d}{d_0\Delta t}$$

и интенсивность их использования

$$\Delta J(t)=\frac{\Delta B}{\Delta d\Delta t}.$$

В результате можно воспользоваться нелинейным дифференцированным уравнением, устанавливающим связь между этими показателями и энтропией H, представляющим математическую модель динамики функционирования системы ДИП

$$\dot{H}=\frac{dH}{dt}=\Delta I(t)H(t)-\Delta J(t)H^2(t),$$

где

$$H=\sum_k\frac{d_k}{D_{iN}}\ln\frac{d_k}{D_{iN}},k=\overline{1,N}$$

N – число источников, содержащих d_j публикаций по i-му направлению.

Этапу зарождения научного направления соответствует минимальное значение энтропии H_{min} и $\dot{H}>0(\Delta J\approx 0)$.

Этапу формирования соответствует рост энтропии, при этом $\dot{H} < 0$ из-за высоких значений ΔJ.

На эволюционном этапе энтропия достигает H_{max} и $\dot{H} > 0$, так как число публикаций приближается к максимальному значению.

На этапе деградации система входит в насыщение, поэтому $\dot{H} \le 0$, так как $\Delta I \approx 0$ и $\Delta J \approx 0$.

Алгоритм, реализующий предложенные принципы, представлен на рис 2.

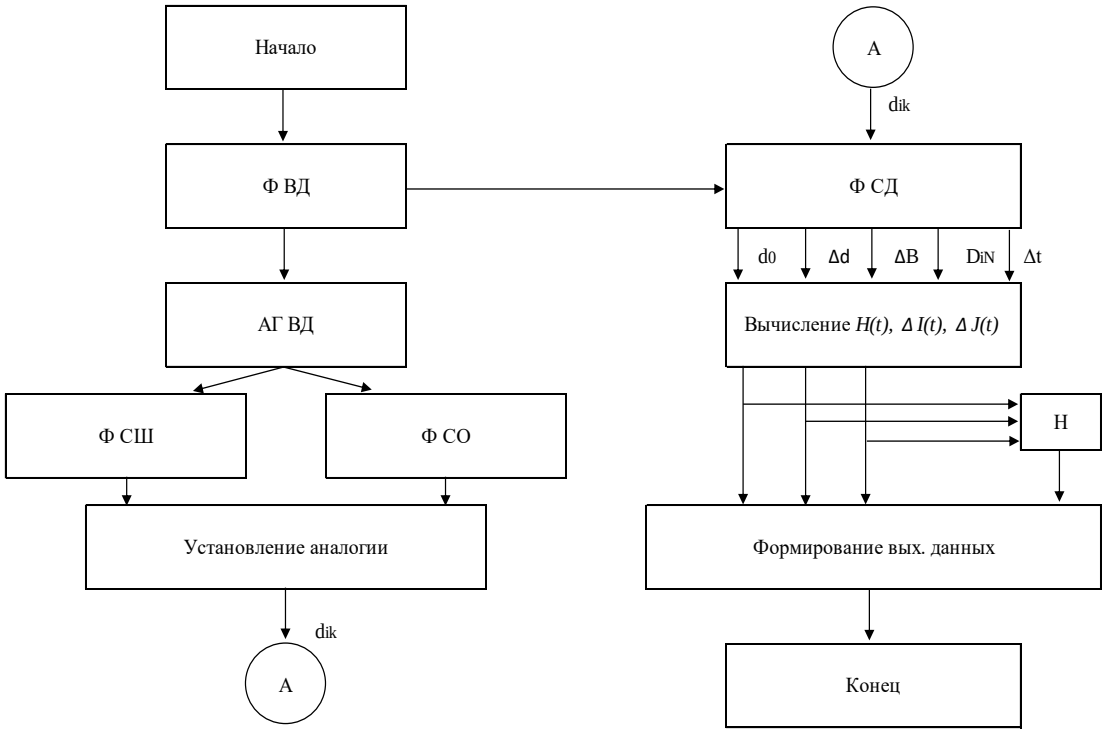

Рисунок 2 – Алгоритм функционирования системы ДИП, где Ф – формирование; ВД – входные данные; АГ – адаптивное группирование; СШ – семантический шаблон; СО – семантический образ; СД – статистические данные.

Система контроля

Структура системы контроля должна обеспечивать выполнение таких операций как извлечение знаний из входных данных, семантическая и статистическая обработка входных данных, диалоговое общение (рис. 3).

База знаний (БЗ), позволяющая решать задачи контроля, должна содержать форматированное в рамках метода и языка представления знаний описание среды, которую должна контролировать система. Знания о среде формируются подсистемой «Извлечение знаний» путем объединения интегрированной входной информации G, и корректирующей информации Δg от эксперта. Подсистема «Диалоговое общение» обрабатывает неформализованное задание Z в интерактивном режиме и использует для своей работы соответствующую БЗ, содержащую правила анализа и синтеза естественно-языковой или графической информации в проблемной области, а также интерпретатор, служащий для преобразования неформализованного задания Z в формализованное T в рамках внутреннего языка системы. Подсистема «Семантическая обработка» обеспечивает выполнение таких операций (рис. 3), как адаптивное груп-

пирование входных данных по вторичным признакам (классификационные индексы, термины, причастность к специализированным изданиям и др.), формирование семантических шаблонов для каждой из ранее полученных групп, формирование текущих семантических образов, установление аналогии между шаблоном и текущим образом. Результатом выполнения этих операций является выделенное число информативных единиц d_{ik}, из общего их числа D_{iN} с высокой степенью вероятности, соответствующее тематике данного (i-го) направления. Подсистема «Статистическая обработка» обеспечивает получение статистической информации по выделенному числу информативных единиц d_{ik}, обеспечивающую эксперту возможность отслеживания и прогнозирования динамики развития научного направления по показателям динамики информационного потока:

- если значение энтропии минимально H_{min} и $\dot{H} > 0 (\Delta J \approx 0)$, то это соответствует этапу зарождения научного направления;
- если происходит рост энтропии и при этом $\dot{H} < 0$ (высокие значения ΔJ), то это соответствует этапу формирования:
- если энтропия максимальна H_{max} и $\dot{H} > 0$, то это соответствует эволюционному этапу;
- если $\dot{H} \le 0$ ($\Delta I \approx 0$ и $\Delta J \approx 0$), то это соответствует этапу деградации.

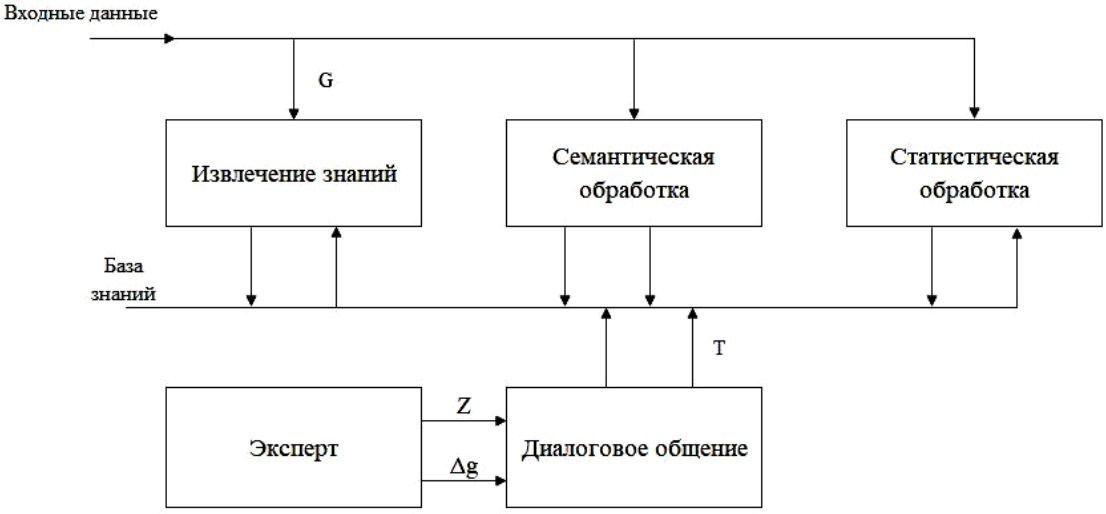

Рисунок 3 – Структурное построение системы контроля

Система контроля прошла апробацию при решении вопросов организации и прогнозирования развития учебного процесса по ряду направлений подготовки.

Выводы

Предложенные принципы структурного построения системы ДИП могут быть использованы при создании экспертных систем, предназначенных для контроля и управления научно-технической информацией.

Список литературы

1. Бурлаева Е.И., Зори С.А. Сравнение некоторых методов машинного обучения для анализа текстовых документов. *Проблемы искусственного интеллекта*. 2019. № 1 (12). С. 42-51.
2. Маталов Д.П., Плискин Е.Л. Веб-сервис на основе SDK для распознавания документов. *Информационные технологии и вычислительные системы*. 2019. № 2. С. 32-43.
3. Пикалёв Я.С. Обзор архитектур систем интеллектуальной обработки естественно-языковых текстов. *Проблемы искусственного интеллекта*. 2020. № 4 (19). С. 45-68.

4. Timofeev A. The future of our society and technical thinking systems. *Problems of Artificial Intelligence*. 2020. № 1 (16). P. 16-22.
5. Андриевская Н.К. Онтологический подход в системах обработки данных научных и научно-образовательных организаций. *Проблемы искусственного интеллекта*. 2020. № 1 (16). С. 23-36.
6. Андриевская Н.К. Гибридная интеллектуальная мера оценки семантической близости. *Проблемы искусственного интеллекта*. 2021. № 1 (20). С. 4-17.
7. Садовская Л.Л., Гуськов А.Е., Косяков Д.В., Мухамедиев Р.И. Обработка текстов на естественном языке: обзор публикаций. *Искусственный интеллект и принятие решений*. 2021. № 3. С. 66-86.
8. Хорев П.Б., Сергеев А.В. Анализ свойств и критерии обнаружения скрытых данных в документах MICROSOFT WORD. *Информационные технологии*. 2021. Т. 27, № 9. С. 470-477.
9. Соловьев А.В. Проблема определения электронного документа долговременного хранения. *Информационные технологии и вычислительные системы*. 2022. № 1. С. 47-54.
10. Анцыферов С.С., Фазилова К.Н., Ханова М.К. Стандартизация показателей свойств документальных информационных потоков. *Национальная научно-техническая конференция с международным участием. Перспективные материалы и технологии (ПМТ-2022).* Сборник докладов конференции Института перспективных технологий и индустриального программирования РТУ МИРЭА. Москва, 2022. С. 118-122.
11. Пикалёв Я.С. Разработка системы нормализации текстовых корпусов. *Проблемы искусственного интеллекта*. 2022. № 2 (25). С. 64-78.
12. Соловьев А.В. Математическая модель электронного документа долговременного хранения. *Информационные технологии и вычислительные системы*. 2022. № 2. С. 30-36.
13. Баканова Н.Б. Многокритериальная оценка публикационной результативности научных подразделений организации. *Искусственный интеллект и принятие решений*. 2022. № 3. С. 88-95.
14. Анцыферов С.С., Фазилова К.Н., Ханова М.К Методика оценки показателей свойств документальных информационных потоков. *Управление документацией в цифровой среде*. Сборник трудов IV национальной научно-практической конференции. Москва, 2022. С. 11-15.
15. Славин О.А. Применение дескрипторов объектов для привязки структурных элементов зашумленных образов деловых документов. *Информационные технологии и вычислительные системы*. 2022. № 4. С. 13-24.
16. Ткаченко А.Л., Денисова Л.А. Автоматическая классификация текстовых документов в системе электронного документооборота вуза. *Информационные технологии и вычислительные системы.* 2023. № 1. С. 3-19.
17. Смирнов И.В. Разноуровневая обработка естественного языка для интеллектуального поиска и анализа текстов. *Искусственный интеллект и принятие решений*. 2023. № 1. С. 90-99.
18. Соболь В.М. Об инфологии документального обмена. *Информационные технологии и вычислительные системы*. 2023. № 2. С. 3-17.
19. Анцыферов С.С., Фазилова К.Н., Ханова М.К. Интеллектуальная система контроля динамики информационных потоков. *Наукоемкие технологии*. 2023. Т. 24, № 5. С. 64-68.
20. Арлазаров В.Л., Славин О.А. Вопросы распознавания и верификации текстовых документов. *Информационные технологии и вычислительные системы*. 2023. № 3. С. 55-61.
21. Анцыферов С.С., Фазилова К.Н., Ханова М.К. Методика контроля динамики информационных потоков. *Качество и жизнь*. 2023. № 3 (39). С. 81-83.

References

1. Burlaeva E.I., Zori S.A. Comparison of some machine learning methods for analyzing text documents. *Problems of artificial intelligence*. 2019. No. 1 (12). pp. 42-51.
2. Matalov D.P., Pliskin E.L. Web service based on SDK for document recognition. *Information technologies and computing systems*. 2019. No. 2. P. 32-43.
3. Pikalev Y.S. Review of architectures of systems for intelligent processing of natural language texts. *Problems of artificial intelligence*. 2020. No. 4 (19). pp. 45-68.
4. Timofeev A. The future of our society and technical thinking systems. *Problems of Artificial Intelligence*. 2020. No. 1 (16). P. 16-22.
5. Andrievskaya N.K. Ontological approach in data processing systems of scientific and scientific-educational organizations. *Problems of artificial intelligence*. 2020. No. 1 (16). pp. 23-36.
6. Andrievskaya N.K. Hybrid intellectual measure for assessing semantic proximity. *Problems of artificial intelligence*. 2021. No. 1 (20). pp. 4-17.
7. Sadovskaya L.L., Guskov A.E., Kosyakov D.V., Mukhamediev R.I. Text processing in natural language: review of publications. *Artificial intelligence and decision making*. 2021. No. 3. P. 66-86.

8. Khorev P.B., Sergeev A.V. Analysis of properties and criteria for detecting hidden data in MICROSOFT WORD documents. *Information Technology*. 2021. T. 27, no. 9. pp. 470-477.
9. Soloviev A.V. The problem of determining an electronic document for long-term storage. *Information technologies and computing systems*. 2022. No. 1. P. 47-54.
10. Antsyferov S.S., Fazilova K.N., Khanova M.K. Standardization of indicators of the properties of documentary information flows. *National scientific and technical conference with international participation. Advanced materials and technologies (PMT-2022).* Collection of reports from the conference of the Institute of Advanced Technologies and Industrial Programming of RTU MIREA. Moscow, 2022. pp. 118-122.
11. Pikalev Y.S. Development of a system for normalizing text corpora. *Problems of artificial intelligence*. 2022. No. 2 (25). pp. 64-78.
12. Soloviev A.V. Mathematical model of an electronic document for long-term storage. *Information technologies and computing systems*. 2022. No. 2. P. 30-36.
13. Bakanova N.B. Multicriteria assessment of the publication performance of scientific divisions of an organization. *Artificial intelligence and decision making*. 2022. No. 3. P. 88-95.
14. Antsyferov S.S., Fazilova K.N., Khanova M.K. Methodology for assessing indicators of the properties of documentary information flows. *Document management in a digital environment*. Collection of proceedings of the IV National Scientific and Practical Conference. Moscow, 2022. pp. 11-15.
15. Slavin O.A. Using object descriptors to link structural elements of noisy images of business documents. *Information technologies and computing systems.* 2022. No. 4. pp. 13-24.
16. Tkachenko A.L., Denisova L.A. Automatic classification of text documents in the university electronic document management system. *Information technologies and computing systems*. 2023. No. 1. P. 3-19.
17. Smirnov I.V. Multi-level natural language processing for intelligent search and text analysis. *Artificial intelligence and decision making*. 2023. No. 1. P. 90-99.
18. Sobol V.M. On the informationology of documentary exchange. *Information technologies and computing systems*. 2023. No. 2. P. 3-17.
19. Antsyferov S.S., Fazilova K.N., Khanova M.K. Intelligent system for monitoring the dynamics of information flows. *High technology*. 2023. T. 24, No. 5. P. 64-68.
20. Arlazarov V.L., Slavin O.A. Issues of recognition and verification of text documents. *Information technologies and computing systems*. 2023. No. 3. P. 55-61.
21. Antsyferov S.S., Fazilova K.N., Khanova M.K. Methodology for monitoring the dynamics of information flows. *Quality and life*. 2023. No. 3 (39). pp. 81-83.

RESUME

S. S. Antsyferov, K. N. Fazilova, K.E. Rusanov

«Documentary Information Flow» Systems Structural Construction Principles

At the present stage of scientific and technological development of an industrial - developed society, the task of forming an information base for each scientific and thematic area is of great importance. The importance of this task is determined by the fact that the formed information base can serve as a basis for creating intelligent control systems and information processing, as well as for forecasting possible directions and trends of scientific development. Any scientific direction goes through certain stages of its development. As a rule, there are 4 main stages: origin, formation, evolutionary development, degradation. Each direction has its own documentary information flow (publications such as articles, patents, reports, conferences, textbooks, dissertations), which allows you to judge both the current state of the thematic area and predict its subsequent state. In turn, the documentary information flow (DIF) can be considered as a kind of self-organizing system, that is, as a dynamic set of interrelated information documents containing fixed scientific information intended for transmission in time and space.

The functioning of semantic processing blocks can occur using well-known semantic strategies (for example, the GRIN1 program) associated with the formation of semantic templates, as well as algorithms for establishing an analogy between the semantic template and the semantic image of the current data (for example, the ZOBRA program).

As a result, we can use a nonlinear differential equation that establishes a relationship between these indicators and entropy, representing a mathematical model of the dynamics of the functioning of the DIF system.

The proposed principles of the structural design of the DIF system can be used to create expert systems designed to control and manage scientific and technical information.

РЕЗЮМЕ

С. С. Анцыферов, К. Н. Фазилова, К. Е. Русанов
Принципы структурного построения систем «документальный информационный поток»

На современном этапе научно-технического развития индустриального - развитого общества большое значение приобретает задача формирования информационной базы по каждому научно-тематическому направлению. Важность данной задачи определяется тем, что сформированная информационная база может служить основой для создания интеллектуальных систем управления и обработки информации, а также для прогнозирования возможных направлений и тенденций научного развития. Любое научное направление проходит определенные этапы своего развития. Как правило, выделяются 4 основных этапа: зарождение, формирование, эволюционное развитие, деградация. Каждому направлению соответствует свой документальный информационный поток (такие публикации как статьи, патенты, отчеты, конференции, учебные пособия, диссертации), позволяющий судить как о текущем состоянии тематического направления, так и прогнозировать его последующее состояние. В свою очередь, документальный информационный поток (ДИП) может рассматриваться как некоторая самоорганизующаяся система, то есть как динамическая совокупность взаимосвязанных информационных документов, содержащих закрепленную научную информацию, предназначенную для передачи во времени и пространстве.

Функционирование блоков семантической обработки может происходить с использованием известных семантических стратегий (например, программа GRIN1), связанных с формированием семантических шаблонов, а также алгоритмов установления аналогии между семантическим шаблоном и семантическим образом текущих данных (например, программа ZOBRA).

В результате можно воспользоваться нелинейным дифференцированным уравнением, устанавливающим связь между этими показателями и энтропией, представляющим математическую модель динамики функционирования системы ДИП.

Предложенные принципы структурного построения системы ДИП могут быть использованы при создании экспертных систем, предназначенных для контроля и управления научно-технической информацией.

Анцыферов Сергей Сергеевич – доктор технических наук, профессор, Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет», г. Москва. *Область научных интересов:* системы искусственного интеллекта, эл. почта antsyferov@mirea.ru, адрес: 119454, г. Москва, проспект Вернадского, дом 78, телефон +7499 600-80-80, доб. 23043

Фазилова Ксения Наильевна – кандидат технических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет», г. Москва. *Область научных интересов*: системы искусственного интеллекта, эл. почта fazilova@mirea.ru, адрес: 119454, г. Москва, проспект Вернадского, дом 78, телефон +7499 600-80-80, доб. 25092

Русанов Константин Евгеньевич – кандидат технических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет», г. Москва. *Область научных интересов*: системы искусственного интеллекта, эл. почта rusanov@mirea.ru, адрес: 119454, г. Москва, проспект Вернадского, дом 78, телефон +7499 600-80-80, доб. 23043

Статья поступила в редакцию 15.05.2023.