

15. Muttillio V., Pomante L., Santic M., Valente, G. SystemC-based co-simulation / analysis for system-level hardware/software co-design, *Computers and Electrical Engineering*, 2023, Vol. 110, 108803. DOI: 10.1016/j.compeleceng.2023.108803.
16. dsPIC33FJXXXGPX06/X08/X10 high-performance, 16-bit digital signal controllers: Data sheet DS70286C. Microchip Technology Inc., 2009, 320 p. Available at: <https://ww1.microchip.com/downloads/en/DeviceDoc/70286C.pdf> (accessed 20 May 2024).
17. dsPIC33F/PIC24H family reference manual. Section 22. Direct memory access (DMA) DS70182A. Microchip Technology Inc., 2006, 60 p. Available at: <http://ww1.microchip.com/downloads/en/devicedoc/70183d.pdf> (accessed 20 May 2024).
18. dsPIC33F/PIC24H family reference manual. Section 16. Analog-to-digital converter (ADC) DS70183D. Microchip Technology Inc., 2006-2012, 88 p. Available at: <https://ww1.microchip.com/downloads/en/DeviceDoc/70182A.pdf> (accessed 20 May 2024).
19. dsPIC33F family reference manual. Section 13. Output compare DS70209A. Microchip Technology Inc., 2007, 24 p. Available at: <http://ww1.microchip.com/downloads/en/DeviceDoc/70209A.pdf> (accessed 20 May 2024).
20. MPLAB starter kit for dsPIC® digital signal controllers: Users Guide DS51700A. Microchip Technology Inc., 2008, 42 p. Available at: <https://ww1.microchip.com/downloads/en/DeviceDoc/51700B.pdf> (accessed 20 May 2024).

Статью рекомендовал к опубликованию д.т.н., профессор Ю.А. Кравченко.

Хлуденев Александр Владимирович – Оренбургский государственный университет; e-mail: avhludenev@yandex.ru; г. Оренбург, Россия, тел.: +73532372874; кафедра промышленной электроники и информационно-измерительной техники; к.т.н.; доцент.

Сильвашко Сергей Анатольевич – e-mail: sisean@yandex.ru; тел.: +73532372874; кафедра промышленной электроники и информационно-измерительной техники; к.т.н.; доцент.

Khudenev Alexander Vladimirovich – Orenburg State University; e-mail: avhludenev@yandex.ru; Orenburg, Russia; phone: +73532372874; the department of Industrial Electronics and Information Measuring Engineering; cand. of eng. sc.; associate professor.

Silvashko Sergey Anatolievich – e-mail: sisean@yandex.ru; phone: +73532372874; the department of Industrial Electronics and Information Measuring Engineering; cand. of eng. sc.; associate professor.

УДК 004.912

DOI 10.18522/2311-3103-2024-4-110-122

Ю.М. Вишняков, Р.Ю. Вишняков

ФОРМАЛИЗАЦИЯ РАСПОЗНАВАНИЯ И ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИХ ОБЪЕКТОВ В ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВЫХ ПОТОКАХ

Участившиеся случаи совершаемых в киберпространстве преступлений, в особенности, в социальных сетях и различного рода мессенджерах требуют создания адекватных и эффективных мер противодействия. Рост киберпреступлений настолько большой, что они уже могут нанести невосполнимый урон государству и обществу. Однако выявление подобного рода преступлений и преступных деяний, наталкивается на большие трудности, так как преступники присутствуют в социальных сетях виртуально и лингвистически, используют всячески их возможности и особенности для сокрытия следов своих преступлений. И, тем не менее, такими инструментами противодействия могли бы быть различного рода распознаватели и идентификаторы, способные автоматически обрабатывать естественный язык, выделять в нем специфические смысловые черты преступных деяний, распознавать и идентифицировать их. Поскольку по многим параметрам и обстоятельствам применительно к данным ситуациям использования нейросетевого подхода представляется нецелесообразным, в предлагаемой работе разрабатывается собственный формальный метод проектирования распознавателя для идентификации в текстовых потоках семантических объектов по их лингвистическим следам. Вводятся такие формальные понятия, как формальная модель семантического объекта, функция поведения, сценарий, лингвистический след, функция распознавания. Рассуждения строятся на теоретико-множественных положениях вычислительной теории семантической интерпретации и используют вычислительное представ-

ление смысла текстовых фрагментов для их сравнения на семантическую близость. Предлагаемый подход носит общий и универсальный характер, он позволяет формальным образом синтезировать распознаватель семантических объектов по их лингвистическим описаниям и поведению. В работе все рассуждения и построения иллюстрируются конкретными примерами.

Семантический объект; семантический след; семантический распознаватель; семантическое сравнение; текстовый фрагмент; семантическая близость; семантическая идентификация.

Yu.M. Vishnyakov, R.Yu. Vishnyakov

FORMALIZATION OF RECOGNITION AND IDENTIFICATION OF SEMANTIC OBJECTS IN NATURAL LANGUAGE TEXT STREAMS

The increasing incidence of crimes committed in cyberspace, particularly on social networks and various messengers, necessitates the development of adequate and effective countermeasures. The rise in cybercrime is so significant that it poses a potential threat of inflicting irreparable harm to the state and society. However, detecting such crimes and criminal activities is challenging because offenders operate virtually and linguistically within social networks, exploiting their features to conceal their traces. Nonetheless, various detection and identification tools capable of automatically processing natural language, highlighting specific semantic features of criminal activities, and recognizing and identifying them could serve as effective countermeasures. Given the impracticality of applying neural network approaches to these situations for several reasons, this study proposes a formal method for designing a recognizer to identify semantic objects in text streams based on their linguistic traces. Formal concepts such as the formal model of a semantic object, behavior function, scenario, linguistic trace, and recognition function are introduced. The reasoning is based on set-theoretical principles of computational theory of semantic interpretation and utilizes computational representations of the meaning of text fragments for their comparison in terms of semantic similarity. The proposed approach is general and universal, allowing for the formal synthesis of a recognizer for semantic objects based on their linguistic descriptions and behavior. All discussions and constructions in the work are illustrated with specific examples.

Semantic object; linguistic trace; semantic recognizer; semantic comparison; text fragment; semantic proximity; semantic identification.

Введение. В настоящее время обработка естественно-языковой информации NLP (Natural Language Processing) представляет одну из востребованных и активно развивающихся областей информатики, что связано с интенсивными коммуникациями в интернет-пространстве. В то же время NLP является и одной из слабо формализованных областей информатики, что определяется сложностью и многозначностью самого естественного языка.

В этой связи с авторы посчитали возможным дать некоторые ссылки на работы, которые с разных сторон касаются проблем NLP, хотя список подобного рода работ более чем обширен. Так, в работе [1–3] обсуждаются общие вопросы и проблемы прикладной лингвистики, а в работе [4] – язык программирования Python в качестве основного инструмента разработки. В работах [5, 6] демонстрируется использование языка программирования Python в прикладном аспекте – текстовом анализе и машинном обучении. Работы [7–9] посвящены решению конкретных вопросов обработки естественно-языковых данных, в работе [10] рассматривается вероятностная модель естественных языков, другие модели естественных языков рассмотрены в [11–14], модели языков в информационном поиске представлены [15, 16]. Из приведенного можно сделать вывод, что обработка естественно-языковой информации в различных аспектах, является более чем актуальна.

Однако со стороны пользователей постоянно усиливаются требования к точности и эффективности естественно-языковой обработки, и данное обстоятельство активно обсуждается в научном сообществе. Именно по этой причине совсем не случайно нейросетевой подход переживает ренессанс, но, как показывает практика, и он не представляет собой панацею.

Проблема состоит в том, что сегодня традиционная частотная парадигма семантики не в полной мере отвечает возрастающим требованиям пользователей и назрела необходимость в поиске иных подходов к моделированию смыслов. В этой связи уход в естественно-языковой обработке от частотной парадигмы представления семантики представляется весьма актуальным. В контексте данного обстоятельства авторы, например в

[17], развивают вычислительную теорию семантической интерпретации. Одним из ее приложений является поиск и идентификация объектов в естественно-языковых текстовых потоках [18, 19]. Под потоками понимаются чаты мессенджеров, социальные сети и пр., а под семантическим объектом - лингвистически представленный в текстовом потоке некий объект (субъект). Например, в одном случае это может быть музыкальный коллектив, в другом – роман или фильм, в третьем – преступная группа, создающие угрозы обществу.

Объекту присущи цель и индивидуальный образ, отображенные в характерные выражения и поведенческие схемы. Индивидуальность выражена в общей логике общения, в синтаксической структуре и логике сообщений, их семантике и пр. Именно такой лингвистический образ объекта понимается под семантическим объектом, а основная задача предлагаемого исследования заключается в том, чтобы выявить и идентифицировать присутствие объекта на основе его лингвистического образа и оставляемым им в текстовом потоке лингвистическим следам.

Востребованность подобных исследований явствует из публикации обозревателя «Новой газеты» Г. Мурсалиевой (НГ № 51 от 16.05.2016г.), которая рассматривалась в нашей работе [3]. В публикации шла речь об обосновавшейся в соцсети ВКонтакте преступной организации «Синий кит», склоняющих подростков к суициду. Владение приемами психологического воздействия в купе с возможностями социальных сетей для сокрытия следов преступлений позволяло преступникам долгое время действовать безнаказанно, создавая угрозы обществу, о чем свидетельствует большое число жертв преступлений. Очевидно, что подобного рода преступные деяния требует создания адекватных мер противодействия.

В связи с вышеизложенным целью настоящей работы является представление исследований по разработке методов формализованного описания и идентификации семантического объекта в текстовых потоках с оценкой степени достоверности. Предлагаемое исследование является развитием результатов работ [18, 19] и их продолжает.

Метод решения. В начале определим ряд базовых формальных понятий, которые положены в основу исследования.

Распознаватель семантических объектов (*Semantic Object Recognizer – RecSemObj*) представляет собой некоторое гипотетическое устройство (модель), который на основе информации о некотором семантическом объекте выявляет и идентифицирует присутствие данного семантического объекта в текстовом потоке и оценивает степень достоверности результата распознавания. *RecSemObj* можно представить тройкой вида:

$$RecSemObj = (T, SemObj, F), \quad (1)$$

где T – текстовый поток, $SemObj$ (*Semantic Object – SemObj*) – семантический объект, F – функция распознавания.

Текстовый поток – это множество упорядоченных (во времени) и различных по смыслу естественно-языковых конструкций (цепочек) – токенов (слов, предложений и их фрагментов):

$$T = \{t_1, t_2, \dots, t_m\}. \quad (2)$$

Отдельный токен t_i , представляет собой элементарную неделимую единицу текстового потока.

Семантический объект (*Semantic object – SemObj*) – это тройка вида:

$$SemObj = (Q, P, Z), \quad (3)$$

где Q – множество лингвистических характеристик; P – функция поведения, Z – цель.

Лингвистическая характеристика. Лингвистическая характеристика q_i представляет собой слова, целостные по смыслу текстовые фрагменты или предложения, которые использует *SemObj* для достижения своих целей. Совокупность лингвистических характеристик семантического объекта (возможно упорядоченных и сведенных в классы) представляется в виде:

$$Q = \{q_1, q_2, \dots, q_n\}. \quad (4)$$

Можно было бы сказать, что множество Q это своеобразный лексикон *SemObj*.

По аналогии с лексическим анализом лингвистической характеристике можно придать лексемоподобную нотацию:

$$\langle ID, \text{Текстовое значение, Класс} \rangle, \quad (5)$$

где ID-идентификатор, Класс – принадлежность определенному классу (классы определяются исходя из конкретных целей). Подобное представление представляется удобным с точки зрения в реализации.

Из лингвистических характеристик можно составить различные последовательности (цепочки), причем потенциальное множество всех таких цепочек описывает полная итерация Q^* множества Q следующим образом:

$$\begin{cases} Q^* = \underbrace{Q^0 \cup Q^1 \cup Q^2 \cup \dots \cup Q^m}_{m \rightarrow \infty}; \\ Q^i = Q^{i-1} \times Q = Q \times Q^{i-1} = \underbrace{Q \times Q \times \dots \times Q}_i; \\ Q^0 = \{\lambda\} - \text{аксиома.} \end{cases} \quad (6)$$

Здесь: $\times$ – операция прямого произведения множеств, λ – пустая цепочка, она не содержит ни одной характеристики. Q^0 – одноэлементное множество, состоящее из пустой цепочки λ . Других цепочек, не включенных в Q^* , нет.

Конкретное поведение *SemObj* направлено на достижение цели Z и задается некоторым сценарием. Сам сценарий составляется из последовательности текстовых заготовок, которыми являются лингвистические характеристики *SemObj*:

$$\alpha = q_1^\alpha q_2^\alpha q_3^\alpha \dots q_r^\alpha. \quad (7)$$

Сценарий поведения Sc (Sc - Scenario) *SemObj*. В общем случае семантический объект *SemObj* может иметь множество сценариев, приводящих к достижению цели Z :

$$\begin{cases} Sc = \{\alpha_1, \alpha_2, \dots, \alpha_n\}; \\ Sc \subset Q^*. \end{cases} \quad (8)$$

Функция поведения P *SemObj*. Определим ее как функциональное соответствие (функцию) на множестве цепочек Q^* и множестве $\{0,1\}$ и представим следующим образом:

$$P: Q^* \rightarrow \{0,1\}, \quad (9)$$

где

$$\begin{cases} P(\beta) = 1, \text{ если } \beta \in Sc; \\ P(\beta) = 0, \text{ если } \beta \notin Sc. \end{cases} \quad (10)$$

Данное функциональное соответствие задает поведение семантического объекта *SemObj*.

Исходя из (9) и (10), семантический объект *SemObj* можно представить в виде некоторого гипотетического устройства, которое последовательно обрабатывает сценарий Sc , строит функцию $P(\beta)$ и достигает цели Z , общая схема которого показана на рис. 1:

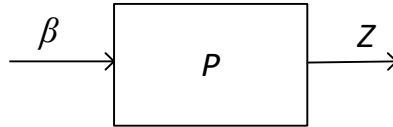


Рис. 1. Общая схема семантического объекта *SemObj*

Для построения формальной поведенческой модели *SemOb* можно воспользоваться различными подходами. Обсудим их кратко.

Нейросетевой подход. В этом случае множество сценариев Sc нужно рассматривать как обучающую выборку. Но опять же в случае бесконечного множества Sc придется создавать репрезентативную выборку, выполнять обучение сети и формировать функцию ошибки, что является далеко не тривиальной процедурой.

Другой – представляет собой формально грамматический подход. Его суть состоит в том, чтобы построить некоторую грамматику $G[Z]$. В данной грамматике $G[Z]$ цель Z является ее начальным символом, множество лингвистических характеристик Q образуют словарь терминальных символов грамматики, отдельный сценарий α_i представляет собой предложение, а множество сценариев образует язык $L(G[Z])$. Тогда процедуру достижения цели можно представить выводом вида $Z \Rightarrow +\alpha_i$.

Еще один подход, который можно рассматривать как частный случай формально-грамматического подхода, основывается на понятиях регулярного множества, регулярно выражения и алгебры регулярных выражений. Здесь множество сценариев Sc соответствуют множеству значений некоторого регулярного выражения E , которое их порождает. И задача проектирования состоит в том, чтобы сконструировать регулярное выражение E и для него построить некоторое гипотетическое устройство (модель $SemObj$). Обсудим данный подход, поскольку, как нам представляется, большинство практических задач, укладываются в него. Кроме того, в языке программирования Python, который стал основным средством разработки программистов, имеются средства описания и обработки подобного рода структур данных [2].

Обобщенное поведение семантического объекта можно представить диаграммой состояний. Как это показано на рис. 2, данная диаграмма имеет одно начальное состояние S , одно заключительное состояние Z и называется системой переходов.

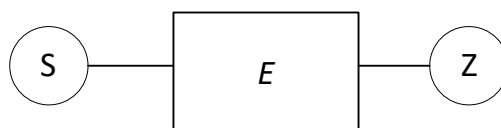


Рис. 2. Диаграмма состояний системы переходов семантического объекта

В системе переходов обобщенная дуга E (представлена прямоугольником) несет следующий смысл – любая цепочка $\alpha_i \in Sc$ переводит $SemObj$ из состояния S в состояние Z .

Диаграмма состояний $SemObj$, в которой дуги переходов представляются только лингвистическими характеристиками и пустой цепочкой λ , называется приведенной. Будем считать, что модель $SemObj$ построена, если она представлена своей приведенной системой переходов. Разбор сценария в приведенной системе переходов $SemObj$ происходит последовательно по его лингвистическим характеристикам.

Для конструирования $SemObj$ воспользуемся алгеброй регулярных выражений. В нашем случае алгебра определяется аксиоматически через операции “|” – ИЛИ, “*” – катенации, “{ }” – итерации следующим образом:

1. $\emptyset$ (пустое множество), λ (пустая цепочка) и имена лингвистических характеристик $q_1, q_2, \dots, q_n$ – регулярные выражения по определению (аксиома, основание);
2. Если e_1 и e_2 – регулярные выражения, то $e_1|e_2$ и $e_1 * e_2$ также регулярные выражения. Обычно в регулярных выражениях знак операции $*$ опускают.

Например, если $e_1 = q_1|q_2$ и $e_2 = q_1q_2$, то $e_1e_2 = (q_1|q_2)q_1q_2 = q_1q_1q_2|q_2q_1q_2$ и для $e_1|e_2 = (q_1|q_2)|q_1q_2 = q_1|q_2|q_1q_2$;

1. Если e регулярное выражение, то и $\{e\}$ также регулярное выражение. В данном случае фигурные скобки $\{ \}$ – представляют операцию итерации;
2. Других регулярных выражений нет.

Диаграммы состояний систем переходов элементарных регулярных выражений представлены на рис. 3:

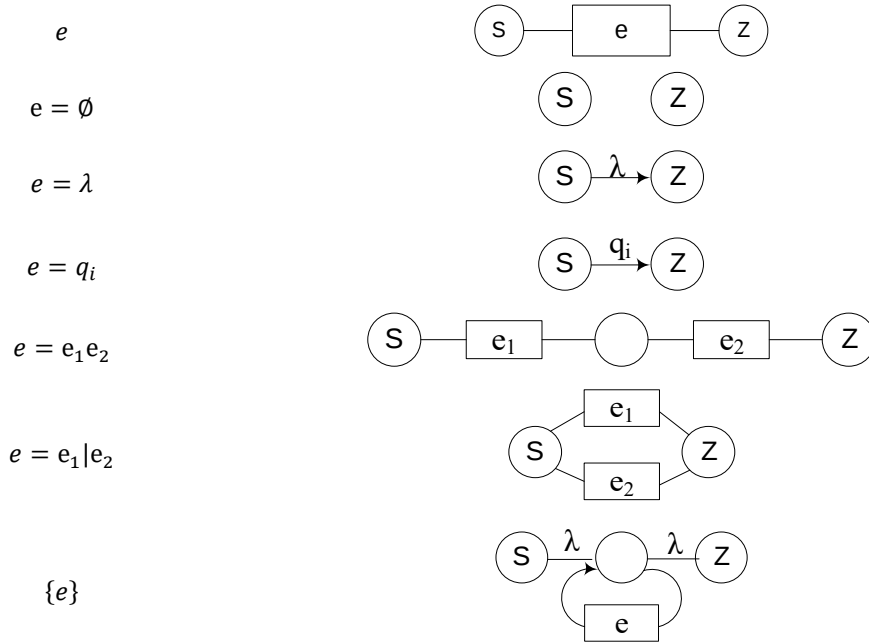


Рис. 3. Диаграммы состояний систем переходов элементарных регулярных выражений

Конструирование *SemObj* состоит в преобразовании системы переходов *SemObj*, показанной на рис. 2, к приведенной системе переходов. Преобразование выполняется декомпозицией получаемой на каждом шаге системы переходов по правилам рис. 3 до тех пор, пока не получится искомая приведенная система переходов.

Пример конструирования приведенной системы переходов для регулярного выражения $E = (q_1|q_2)q_3$ показан ниже на рис. 4:

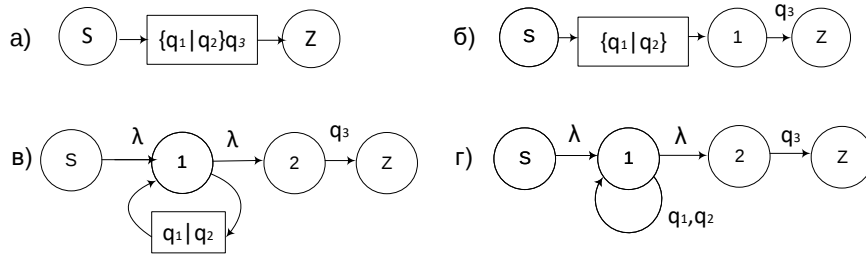


Рис. 4. Пример конструирования диаграммы состояний приведенной системы переходов

Заметим, что регулярное выражение e представляет собой формулу алгебры регулярных выражений. Значение e обозначается как $|e|$ и представляет множество цепочек, порождаемых регулярным выражением e . В случае лингвистической характеристики q_i значение имеет вид:

$$|q_i| = \{\text{"текст"}\}. \quad (11)$$

Например, пусть $|q_1| = \{\text{"расскажи о себе"}\}$, $|q_2| = \{\text{"ты сильный"}\}$. Значения для составных регулярных выражения имеют вид $|(q_1|q_2)| = \{\text{расскажи о себе, ты сильный}\}$ и $|q_1q_2| = \{\text{расскажи о себе ты сильный}\}$.

Значение регулярного выражения $\{e\}$ представляется в виде $|\{e\}| = \{\lambda, e, ee, eee, \dots, \underbrace{ee \dots e}_{n \rightarrow \infty}\}$. Например, если $e = q_2$, то $|\{q_2\}|$ представляет множество

$\{\lambda, \text{ты сильный}, \text{ты сильный ты сильный}, \text{ты сильный ты сильный ты сильный} \dots\}$.

Пусть для *SemObj* рис. 4 заданы следующие конкретные сценарии, которые он обрабатывает:

$$\begin{cases} \alpha_1 = q_3 \\ \alpha_2 = q_1 q_2 q_3 \\ \alpha_3 = q_1 q_1 q_2 q_2 q_3 \end{cases} \quad (12)$$

Отображение приведенной системы переходов на сценарий будем называть линейной разверткой *SemObj* и для (12) такие линейные развертки имеют вид:

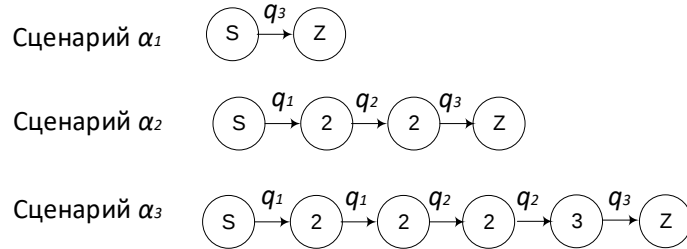


Рис. 5. Диаграммы состояний линейных разверток семантического объекта *SemObj*

Обратимся к вычислительной теории семантической интерпретации [8] и обсудим сравнение тестовых цепочек на семантическую близость. Пусть задана некоторая цепочка слов q :

$$q = a_1 a_2 \dots a_n. \quad (13)$$

Она представляет целостный по смыслу текстовый фрагмент (предложение или его часть). Целостность означает, что каждое слово a_i входит в словосочетания, а сами словосочетания образуют отношение непосредственного подчинения на множестве слов цепочки.

Пусть $S(a_i)$ множество смысловых значений слова a_i , тогда множество смысловых значений цепочки $S(q)$ представляется функционалом вида:

$$\begin{cases} S(q) = \Phi(S(a_1), S(a_2), \dots, S(a_n)); \\ S(q) \subset S(a_i), \text{ если } a_i \text{ главное слово } q. \end{cases} \quad (14)$$

Пусть заданы два текстовых фрагмента q и t и их смысловые значения $S(q)$ и $S(t)$ соответственно. Введем функцию, которая определяет их смысловую близость и назовем ее критерием семантической близости *Cprox* [8]:

$$\begin{cases} Cprox(S(q), S(t)) \in D; \\ D = [0..1]. \end{cases} \quad (15)$$

Если $Cprox=0$, то семантическая близость между q и t отсутствует, при $Cprox=1$ имеет место полная семантическая близость.

В простейшем случае критерий семантической близости можно представить долей совпадающих элементов смысла p в семантической схеме цепочек q и t к общему числу элементов смысла m цепочки q :

$$Cprox(S(q), S(t)) = \frac{p}{m}. \quad (16)$$

Отметим, что традиционный частотный критерий релевантности является частным случаем (16), когда m представляет число слов в цепочке q , а p – число слов из q , совпадающих со словами в цепочке t . Из (16) следует справедливость соотношения:

$$Cprox(S(q), S(t)) \neq Cprox(S(t), S(q)). \quad (17)$$

Из выше обсужденного для нас важен факт возможности измерения семантической близости двух текстовых отрезков.

Как ранее отмечалось, в текстовом потоке имеются токены, семантика которых в той или иной степени близка лингвистическим характеристикам *SemObj*. Такие токены будем называть лингвистическими следами характеристик, а их совокупность лингвистическим следом семантического объекта *SemObj*. Формализуем данные понятия.

Лингвистический след характеристики для некоторого *SemObj* (3) представляется тройкой вида:

$$\begin{cases} (q_i, t_j, w_k): \\ q_i \in Q, t_j \in T, w_k = Cprox(q_i, t_j), 0 < w_k \leq 1, \end{cases} \quad (18)$$

здесь величина w_i является мерой семантической близости, которую можно рассматривать как меру присутствия характеристики q_i в текстовом потоке.

Пусть тройка (q, w, t) задает некоторый семантический след в текстовом потоке, тогда из (18) по определению справедливо соотношение:

$$(q, w) = t. \quad (19)$$

Функция распознавания F . Пусть *SemObj* реализует в тестовом потоке T один из сценариев $\alpha \in Sc$ вида $\alpha = q_1 q_2 \dots q_n$, тогда существует такая же последовательность семантических следов (q_i, t_j, w_k) и последовательность токенов $\alpha' = t_{r1} t_{r2} \dots t_{rn}$, представляющая отображение (19) сценария α на текстовый поток T . Тогда в общем случае функция распознавания имеет вид:

$$\begin{cases} F(\alpha, \alpha') \in D: \\ D = [0..1]. \end{cases} \quad (20)$$

Функции распознавания F можно придать смысл сравнения на семантическое подобие (*Csemsim* - comparison for semantic similarity) сценария α и последовательности токенов α' следующим образом:

$$F = Csemsim(\alpha, \alpha'). \quad (21)$$

$Csemsim(\alpha, \alpha')$ можно представить в виде:

$$Csemsim(\alpha, \alpha') = Csemsim(Cprox(q_1, t_{r1}), Cprox(q_2, t_{r2}), \dots, Cprox(q_n, t_{rn})) \quad (22)$$

или с учетом (18) получим:

$$Csemsim(\alpha, \alpha') = Csemsim(w_{k1}, w_{k2}, \dots, w_{kn}). \quad (23)$$

В зависимости от практических целей для F можно подбирать различные виды функциональной аппроксимации, обеспечивая тонкую настройку *RecSemObj*. В нашем случае для понимания вполне достаточно понимать F как средневзвешенную величину:

$$F = Csemsim(\alpha, \alpha') = \frac{(w_{k1} + w_{k2} + \dots + w_{kn})}{n} \quad (24)$$

Рассмотрим функцию распознавания для сценария α_3 (12) *SemObj*, показанного на рис. 4:

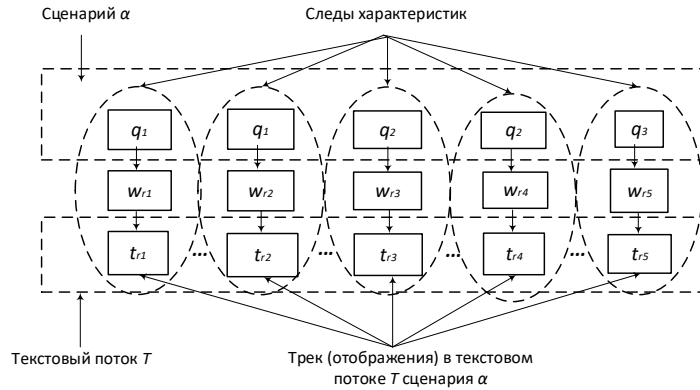


Рис. 6. Схема семантических следов *SemObj* в текстовом потоке

На рис. 6 представлен некоторый текстовый поток T , в который встроен сценарий α_3 *SemObj*. Сценарий представлен последовательностью характеристик $\alpha_3 = q_1 q_1 q_2 q_2 q_3$, которые будучи отображены в текстовый поток определяют в нем трек токенов $\alpha'_3 = t_{r1} t_{r2} t_{r3} t_{r4} t_{r5}$. Характеристики сценария, токены трека и степени их семантической близости задают семантические следы *SemObj* в текстовом потоке T . На рис. 6 семантические следы выделены пунктирными овалами.

Выполнив подстановку в $\alpha'_3 = t_{r1} t_{r2} t_{r3} t_{r4} t_{r5}$ в соответствии с (19), получим:

$$\alpha'_3 = (q_1, w_{p1})(q_1, w_{p2})(q_2, w_{p3})(q_2, w_{p4})(q_3, w_{p5}), \quad (25)$$

тогда для трека токенов α'_3 функция распознавания будет иметь вид:

$$F = Csemsim(\alpha, \alpha') = \frac{(w_1 + w_2 + w_3 + w_4 + w_5)}{5}. \quad (26)$$

Проиллюстрируем вышеизложенное на примере реально подобного текстового потока. В нем некто Куратор (К) склоняет Жертву (Ж) совершить «ЭТО».

Набор лингвистических характеристик $Q = \{q_1, q_2, \dots, q_{30}\}$ *SemObj* “Куратор”, сведенных в классы “Контакт”, “Вовлечение”, “Игра”, представлен в табл. 1. Поведение *SemObj* “Куратор” описывается нижеуказанными регулярным выражением (27) и приведенной системой переходов рис. 7:

$$E = q_1 \{q_2 | q_3 | \dots | q_7\} q_8 \{q_9 | q_{10} | \dots | q_{16}\} q_{17} \{q_{16} | q_{17} | \dots | q_{29}\} q_{24} \quad (27)$$

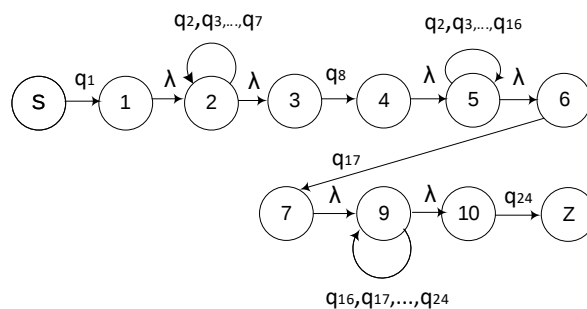


Рис. 7. Диаграмма состояний приведенной системы переходов *SemObj* “Куратор”

Таблица 1

Лингвистические характеристики *SemObj* “Куратор”

ID	Значение характеристики	Класс	ID	Значение характеристики	Класс
q ₁	Привет, как зовут тебя?	КОНТАКТ	q ₁₆	проблемы уйдут	ИГРА
q ₂	расскажи о себе		q ₁₇	твое задание	
q ₃	Я живу в городе Н.		q ₁₈	сядь на стул и замри	
q ₄	попробую угадать твой адрес		q ₁₉	молодец, ты справился	
q ₅	пять лет назад переехал с родителями сюда		q ₂₀	ножом надави на руку	
q ₆	Мой папа врач, а мама учительница		q ₂₁	терпи, даже если будет кровь	
q ₇	раньше часто сорился с родителями		q ₂₂	нужно продержаться пять минут	
q ₈	когда-то я освоил одну игру	ВОВЛЕЧЕНИЕ	q ₂₃	ты сильный	
q ₉	игра помогло мне		q ₂₄	сделай ЭТО	
q ₁₀	стал уверенным и смелым		q ₂₅	ничего не бойся	
q ₁₁	веду игру в соцсетях и помогаю людям		q ₂₆	избавишься от проблем и страданий	
q ₁₂	даю задания, ты выполняешь		q ₂₇	ты слабак и неудачник	
q ₁₃	у тебя всё получится		q ₂₈	ты никому не нужен	
q ₁₄	Нужно делать все, что я говорю		q ₂₉	наказание будет страшным	
q ₁₅	закалим твой характер				

В табл. 2 представлен текстовый поток и для *SemObj* “Куратор” один из возможных сценариев α , который будучи отображен в текстовый поток порождает трек токенов α' .

Таблица 2

Текстовый поток с треком токенов сценария α'

№ пп	(q_i, w)	Трек сценария α
1.	$(q_1, 1, 0)$	К. Привет, я Костя, а как зовут тебя?
2.		Ж. Я Миша.
3.	$(q_2, 1, 0)$	К. Миша, расскажи о себе.
4.	$(q_3, 1, 0)$	К. Я живу в городе Н. А ты?
5.		Ж. Ого! А я тоже живу в городе Н.
6.	$(q_4, 1, 0)$	К. Попробую угадать твой адрес. Улица Усова, дом 35?
7.		Ж. Угадал! Здорово!
8.	$(q_5, 0, 5)$	К. Миша, я переехал сюда с родителями из города Н.
9.	$(q_6, 1, 0)$	К. Мой папа врач, а мама учительница.
10.	$(q_7, 1, 0)$	К. Я раньше часто сорился с родителями, а ты, Миша?
11.		Ж. Постоянные ссоры, меня не понимают, заклевали!
12.	$(q_8, 1, 0)$	К. Миша, я когда-то освоил одну игру.
13.	$(q_9, 1, 0)$	К. Игра помогла мне.
14.		К. Я решил свои проблемы, наладились отношения.
15.	$(q_{10}, 1, 0)$	К. Стал уверенным и смелым.
16.	$(q_{11}, 1, 0)$	К. Теперь сам веду игру в соцсетях и помогаю людям.
17.		Ж. Костя, интересно! А что это за игра?
18.	$(q_{12}, 1, 0)$	К. Я даю задания, а ты выполняешь.
19.	$(q_{13}, 1, 0)$	К. Уверен, у тебя всё получится.
20.	$(q_{14}, 1, 0)$	К. Нужно делать все, что я говорю.
21.	$(q_{15}, 1, 0)$	К. Мы закалим твой характер.
22.	$(q_{16}, 1, 0)$	К. Проблемы сами уйдут.
23.		К. Начнем?
24.		Ж. Да, хорошо, попробуем
25.	$(q_{17}, 1, 0)$	К. Это твое первое задание.
26.	$(q_{18}, 1, 0)$	К. Сядь на стул и замри на 15 минут.
27.		Ж. У меня получилось! Ура!
28.	$(q_{19}, 1, 0)$	К. Молодец, ты справился!
29.	$(q_{17}, 0, 5)$	К. Завтра новое задание
30.		Ж. Буду ждать, интересно!
31.	$(q_{17}, 1, 0)$	К. Привет, Миша. Теперь твое новое задание.
32.		К. Укрепи характер.
33.		Ж. Костя, я весь внимание
34.	$(q_{20}, 1, 0)$	К. Острием ножа сильно надави на руку.
35.	$(q_{21}, 1, 0)$	К. Терпи, даже если будет кровь.
36.	$(q_{22}, 0, 5)$	К. продержись хотя бы 10 минут.
37.		Ж. Я не смогу это сделать.
38.	$(q_{23}, 1, 0)$	К. Сможешь, ты сильный
39.		Ж. Хорошо, я попробую.

40.		Ж. Я сделал, получилось!
41.	(q ₂₃ ,1,0)	К. Молодец, ты сильный!
42.	(q ₁₇ ,1,0)	К. Теперь твое самое сложное задание
43.	(q ₂₄ ,1,0)	К. Сделай ЭТО!
44.	(q ₂₅ ,1,0)	К. Ничего не бойся!
45.		Ж. Я не могу, я боюсь. Если меня не станет? А как же родители?
46.	(q ₂₆ ,0,7)	К. Разом избавишься от всех проблем!
47.	(q ₂₇ ,1,0)	К. Для друзей ты слабак и неудачник
48.	(q ₂₈ ,1,0)	К. Ты никому не нужен!
49.	(q ₂₄ ,1,0)	К. Ну! давай же! сделай ЭТО!
50.		Ж. Я очень боюсь!
51.	(q ₂₈ ,1,0)	К. Давай же, ты никому не нужен!
52.	(q ₂₄ ,1,0)	К. Немедленно сделай ЭТО!!
53.		Ж. Родители любят меня!
54.	(q ₂₈ ,1,0)	К. Ты никому не нужен!
55.	(q ₂₄ ,1,0)	К. Сделай ЭТО!
56.	(q ₂₆ ,0,7)	К. Избавься от страданий!
57.	(q ₂₄ ,1,0)	К. Сделай ЭТО
58.		Ж. Я боюсь, пожалей меня!
59.	(q ₂₉ ,1,0)	К. Не зли меня, наказание будет страшным
60.	(q ₂₄ ,1,0)	К. Сделай ЭТО!

В каждом токене жирным шрифтом выделены слова порождающей его лингвистической характеристики и указана степень w семантической близости с ней. Вычисленное значение функции распознавания $C_{\text{sensim}}(\alpha, \alpha')$ для нашего примера равно 0,95, что с высокой долей уверенности указывает на присутствие в потоке семантического объекта *SemObj* “Куратор”, реализующего свой сценарий поведения для достижения цели.

Эксперименты. Проведение экспериментов с представленной моделью семантического объекта осуществлялось на специально разработанном программном комплексе [20]. Эксперименты проводились для разных поведенческих сценариев семантического объекта и в различных текстовых потоках, они показали хорошую работоспособность сконструированного распознавателя семантических объектов и соответствие процесса распознавания заявляемым теоретическим характеристикам.

Заключение. В работе предложен формальный метод проектирования распознавателя для идентификации в естественно-языковых текстовых потоках семантических объектов по их лингвистическим следам. Разработаны формальная модель семантического объекта, функция поведения, сценарий, лингвистический след, функция распознавания. Предложена формальная методика синтеза распознавателя, основанная на алгебре регулярных выражений и автоматной модели – системе переходов. Для сравнения текстовых фрагментов на семантическую близость использовано вычислительное представление смысла. По мнению авторов предложенные решения можно рекомендовать в первую очередь для обнаружения и предотвращения преступлений в социальных сетях.

Исследование выполнено при финансовой поддержке Кубанского научного фонда в рамках научно-инновационного проекта «НИП-20.1.4».

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Николаев И.С. Компьютерная и прикладная лингвистика / Николаев И.С., Митренина О.В., Ландо Т.М. (ред.). – М.: Ленанд, 2016. – 316 с.
2. Шемякин Ю.И. Начала компьютерной лингвистики: учеб. пособие. – М.: Изд-во МГОУ, А/О “Росвузнаука”, 1992. – 120 с.

3. *Тестелец Я.Г.* Введение в общий синтаксис: учеб. пособие. – М.: Изд-во РГГУ, 2001. – 830 с.
4. *Прохоренок Н.А., Дронов В.А.* Python 3. Самое необходимое. – СПб.: БХВ-Петербург, 2019. – 608 с.
5. *Бенгфорт Бенджамин, Билбро Ребекка, Охеда Тони.* Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – СПб.: Питер, 2019. – 368 с.
6. *Луис Педро Коэльо, Вилли Ричарт.* Построение Системы машинного обучения на языке Python: пер. с англ. Слинкин А.А. – М.: ДМК Пресс, 2016. – 302 с.
7. *Koncel-Kedziorski R., Hajishirzi H. and Sabharwal A. et al.* Parsing algebraic word problems into equations // Transactions of the Association for Computational Linguistics. – 2015. – 3. – P. 585-597.
8. *Devlin J., Chang M.W., Lee K. and Toutanova K.* Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805, 2018.
9. *Hu K., Wu H. and Qi K. et al.* A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model // Scientometrics. – Springer, 2017. – P. 1-38.
10. *Налимов В.В.* Вероятностная модель языка. О соотношении естественных и искусственных языков. – М.: Наука, 1979. – 303 с.
11. *Гладкий А.В.* Синтаксические структуры естественного языка в автоматизированных системах общения. – М.: Наука. Главная редакция физико-математической науки, 1985. – 144 с. – (Серия «Проблемы искусственного интеллекта»).
12. *Найханова Л.В., Евдокимова И.С.* Методы и алгоритмы трансляции естественно-языковых запросов к базе данных в SQL-запросы: монография. – Улан-Удэ: Изд-во ВСГТУ, 2004. – 148 с.
13. *Севбо И.П.* Графическое представление синтаксических структур и стилистическая диагностика. – Киев: Наукова думка, 1983. – 192 с.
14. *Рубашкин В.Ш.* Представление и анализ смысла в интеллектуальных информационных системах. – М.: Наука. Главная редакция физико-математической науки, 1989. – 192 с. – (Серия «Проблемы искусственного интеллекта»).
15. *Aggarwal C.C., Al-Garawi F., Yu P.S.* Intelligent crawling on the world wide web with arbitrary predicates // In Proc. of the WWW10, May 2001. – P. 96-105.
16. *Agichtein E., Lawrence S., Gravano L.* Learning search engine specific query transformations for question answering // In Proc. of the WWW10. – 2001. – P. 169-178.
17. *Вишняков Ю.М., Вишняков Р.Ю.* Вычислительная семантическая интерпретация текстов научно-технического стиля // Современные наукоемкие технологии. – 2016. – № 12–2. – С.236-242.
18. *Vishnyakov R.Y., Vishnyakov Y.M.* Identification of semantic objects // Journal of Physics: Conference Series, Bristol May 2021. – Vol. 1902. – DOI: 10.1088/1742-6596/1902/1/012104.
19. *Вишняков Ю.М. Вишняков Р.Ю.* О применении вычислительной теории семантической интерпретации к выявлению киберпреступлений // Актуальные проблемы прикладной математики, информатики и механики: Сб. трудов международной научно-технической конференции, Воронеж, 13–15 декабря 2021 г. – Воронеж, 2022. – С. 1526-1530.
20. *Степанова Е.В., Вишняков Ю.М.* Программное обеспечение для вычисления семантической близости текстов // Прикладная математика: современные проблемы математики, информатики и моделирования: Сб. материалов III-й Всероссийской научно-практической конференции молодых ученых. – Краснодар, 2021. – С. 218-223.

REFERENCES

1. *Nikolaev I.S.* Komp'yuternaya i prikladnaya lingvistika [Computer and applied linguistics], Nikolaev I.S., Mitrenina O.V., Lando T.M. (ed.). Moscow: Lenand, 2016, 316 p.
2. *Shemyakin Yu.I.* Nachala komp'yuternoy lingvistiki: ucheb. posobie [The beginnings of computational linguistics: a textbook]. Moscow: Izd-vo MGOU, A/O "Rosvuznauka", 1992, 120 p.
3. *Testeleys Ya.G.* Vvedenie v obshchiy sintaksis: ucheb. posobie [Introduction to the general syntax: a textbook]. Moscow: Izd-vo RGGU, 2001, 830 p.
4. *Prokhorenok N.A., Dronov V.A.* Python 3. Samoe neobkhodimoe [Python 3. The most necessary]. Saint Petersburg: BKhV-Peterburg, 2019, 608 p.
5. *Bengfort Bendzhamin, Bilbro Rebekka, Okheda Toni.* Prikladnoy analiz tekstovyykh dannyykh na Python. Mashinnoe obuchenie i sozdanie prilozheniy obrabotki estestvennogo yazyka [Applied analysis of text data in Python. Machine learning and the creation of natural language processing applications]. Saint Petersburg: Piter, 2019, 368 p.
6. *Luis Pedro Koel'o, Villi Richart.* Postroenie Sistemy mashinnogo obucheniya na yazyke Python [Building a Machine learning System in Python]: transl. from engl. Slinkin A.A. Moscow: DМК Press, 2016, 302 p.

7. Koncel-Kedziorski R., Hajishirzi H. and Sabharwal A. et al. Parsing algebraic word problems into equations, *Transactions of the Association for Computational Linguistics*, 2015, 3, pp. 585-597.
8. Devlin J., Chang M.W., Lee K. and Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*, 2018.
9. Hu K., Wu H. and Qi K. et al. A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model, *Scientometrics*. Springer, 2017, pp. 1-38.
10. Nalimov V.V. Veroyatnostnaya model' yazyka. O sootnoshenii estestvennykh i iskusstvennykh yazykov [A probabilistic model of language. On the relationship between natural and artificial languages]. Moscow: Nauka, 1979, 303 p.
11. Gladkiy A.V. Sintaksicheskie struktury estestvennogo yazyka v avtomatizirovannykh sistemakh obshcheniya [Syntactic structures of natural language in automated communication systems]. Moscow: Nauka. Glavnaya redaktsiya fiziko-matematicheskoy nauki, 1985, 144 p. (Seriya «Problemy iskusstvennogo intellekta» [Series "Problems of artificial intelligence"]).
12. Naykhanova L.V., Evdokimova I.S. Metody i algoritmy translyatsii estestvenno-yazykovykh zaprosov k baze dannykh v SQL-zaprosy: monografiya [Methods and algorithms for translating natural language database queries into SQL queries: monograph]. Ulan-Ude: Izd-vo VSGTU, 2004, 148 p.
13. Sevbo I.P. Graficheskoe predstavlenie sintaksicheskikh struktur i stilisticheskaya diagnostika [Graphical representation of syntactic structures and stylistic diagnostics]. Kiev: Naukova dumka, 1983, 192 p.
14. Rubashkin V.Sh. Predstavlenie i analiz smysla v intellektual'nykh informatsionnykh sistemakh [Representation and analysis of meaning in intelligent information systems]. Moscow: Nauka. Glavnaya redaktsiya fiziko-matematicheskoy nauki, 1989, 192 p. (Seriya «Problemy iskusstvennogo intellekta» [Series "Problems of artificial intelligence"]).
15. Aggarwal C.C., Al-Garawi F., Yu P.S. Intelligent crawling on the world wide web with arbitrary predicates, *In Proc. of the WWW10, May 2001*, pp. 96-105.
16. Agichtein E., Lawrence S., Gravano L. Learning search engine specific query transformations for question answering, *In Proc. of the WWW10, 2001*, pp. 169-178.
17. Vishnyakov Yu.M., Vishnyakov R.Yu. Vychislitel'naya semanticheskaya interpretatsiya tekstov nauchno-tehnicheskogo stilya [Computational semantic interpretation of scientific and technical style texts], *Sovremennye naukoemkie tekhnologii* [Modern high-tech technologies], 2016, No. 12-2, pp. 236-242.
18. Vishnyakov R.Y., Vishnyakov Y.M. Identification of semantic objects, *Journal of Physics: Conference Series, Bristol May 2021*. – Vol. 1902. DOI: 10.1088/1742-6596/1902/1/012104.
19. Vishnyakov Yu.M., Vishnyakov R.Yu. O primeneni vychislitel'noy teorii semanticheskoy interpretatsii k vyavleniyu kiberprestupleniy [On the application of the computational theory of semantic interpretation to the identification of cybercrimes], *Aktual'nye problemy prikladnoy matematiki, informatiki i mekhaniki: Sb. trudov mezhdunarodnoy nauchno-tehnicheskoy konferentsii, Voronezh, 13-15 dekabrya 2021 g.* [Actual problems of applied mathematics, computer science and mechanics: Proceedings of the International Scientific and Technical Conference, Voronezh, December 13-15, 2021]. Voronezh, 2022, pp. 1526-1530.
20. Stepanova E.V., Vishnyakov Yu.M. Programmnoe obespechenie dlya vychisleniya semanticheskoy blizosti tekstov [Software for calculating the semantic proximity of texts], *Prikladnaya matematika: sovremennye problemy matematiki, informatiki i modelirovaniya: Sb. materialov III-y Vserossiyskoy nauchno-prakticheskoy konferentsii molodykh uchenykh* [Applied mathematics: modern problems of mathematics, computer science and modeling: A collection of materials of the III All-Russian Scientific and practical Conference of Young Scientists]. Krasnodar, 2021, pp. 218-223.

Статью рекомендовал к опубликованию д.т.н., профессор В.В. Курейчик.

Вишняков Юрий Муссович – Кубанский государственный университет; e-mail: jury.vishnyakov@gmail.com; г. Краснодар, Россия; тел.: +79996314840; кафедра вычислительных технологий; д.т.н.; профессор.

Вишняков Ренат Юрьевич – e-mail: renat.vishnyakov@mail.ru; тел.: +79996300250; кафедра вычислительной математики и информатики; к.т.н.; доцент.

Vishnyakov Yuriy Mussovich – Kuban State University; e-mail: jury.vishnyakov@gmail.com; Krasnodar, Russia; phone: +79996314840; the department of Computing Technologies; dr. of eng. sc.; professor

Vishnyakov Renat Yurevich – e-mail: renat.vishnyakov@mail.ru; phone: +79996300250; the department of Computational Mathematics and Computer Science; cand. of eng. sc.; associate professor.