ISSN 1818-1015
eISSN 2313-5417

2021 Volume 28 No 3

MODELING AND ANALYSIS OF INFORMATION SYSTEMS

SCIENTIFIC JOURNAL

Start date of publication — 1999
Published quarterly

FOUNDER
P.G. Demidov Yaroslavl State University

EDITORIAL OFFICE
14 Sovetskaya str., Yaroslavl 150003, Russian Federation

Website: http://mais-journal.ru
E-mail: mais@uniyar.ac.ru
Phone: +7 (4852) 79-77-73

2021 Том 28 № 3

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ

НАУЧНЫЙ ЖУРНАЛ

Издается с 1999 года
Выходит 4 раза в год

УЧРЕДИТЕЛЬ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Ярославский государственный университет им. П. Г. Демидова»

РЕДАКЦИЯ
ул. Советская, 14, Ярославль, 150003, Российская Федерация
Website: http://mais-journal.ru
E-mail: mais@uniyar.ac.ru
Телефон: +7 (4852) 79-77-73

Свидетельство о регистрации СМИ ПИ № ФС 77–66186 от 20.06.2016 выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций. Подписной индекс — 31907 в Объединенном каталоге «Пресса России». Технический редактор, компьютерная вёрстка – М. С. Каряева. Корректор английского текста – М. С. Комар. Подписано в печать 10.09.2021. Дата выхода в свет 30.09.2021. Формат 200×265 мм. Объем 102 с. Тираж 40 экз. Свободная цена. Заказ 054/020. Адрес типографии: ул. Советская, 14, оф. 109, Ярославль, 150003, Россия. Адрес издателя: Ярославский государственный университет им. П. Г. Демидова, ул. Советская, 14, Ярославль, 150003, Россия.
Содержание предназначено для детей старше 12 лет.

Contents

Содержание

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

ALGORITHMS

Algorithm for Finding Feedback in a Problem with Constraints for One Class of Nonlinear Control Systems

M. G. Dmitriev[1], Z. N. Murzabekov[2], G. A. Mirzakhmedova[2]

DOI: 10.18255/1818-1015-2021-3-220-233

[1]Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 44/2, Vavilova str., Moscow 119333, Russia.
[2]Al - Farabi Kazakh National University, 71 al-Farabi Ave., Almaty 050040, Republic of Kazakhstan.

MSC2020: 49J15
Research article
Full text in Russian

Received August 20, 2021
After revision August 31, 2021
Accepted September 1, 2021

For a continuous nonlinear control system on a finite time interval with control constraints, where the right-hand side of the dynamics equations is linear in control and linearizable in the vicinity of the zero equilibrium position, we consider the construction of a feedback according to the Kalman algorithm. For this, the solution of an auxiliary optimal control problem with a quadratic functional is used by analogy with the SDRE approach.
Since this approach is used in the literature to find suboptimal synthesis in optimal control problems with a quadratic functional with formally linear systems, where all coefficient matrices in differential equations and criteria can contain state variables, then on a finite time interval it becomes necessary to solve a complicated matrix differential Riccati equations, with state-dependent coefficient matrices. This circumstance, due to the nonlinearity of the system, in comparison with the Kalman algorithm for linear-quadratic problems, significantly increases the number of calculations for obtaining the coefficients of the gain matrix in the feedback and for obtaining synthesis with a given accuracy. The proposed synthesis construction algorithm is constructed using the extension principle proposed by V. F. Krotov and developed by V. I. Gurman and allows not only to expand the scope of the SDRE approach to nonlinear control problems with control constraints in the form of closed inequalities, but also to propose a more efficient computational algorithm for finding the matrix of feedback gains in control problems on a finite interval. The article establishes the correctness of the application of the extension principle by introducing analogs of the Lagrange multipliers, depending on the state and time, and also derives a formula for the suboptimal value of the quality criterion. The presented theoretical results are illustrated by calculating suboptimal feedbacks in the problems of managing three-sector economic systems.

Keywords: optimal control problem; Lagrange multiplier method; nonlinear system; quadratic functional; feedback; SDRE approach; three-sector economic control object

INFORMATION ABOUT THE AUTHORS

Michail G Dmitriev correspondence author	orcid.org/0000-0002-1184-129X. E-mail: mdmitriev@mail.ru Doctor of Sciences, Professor, Chief Researcher.
Zainelkhriet N Murzabekov	orcid.org/0000-0002-9074-4753. E-mail: murzabekov-zein@mail.ru Doctor of Sciences, Professor, Chief Researcher.
Gulbanu A Mirzakhmedova	orcid.org/0000-0001-7915-945X. E-mail: gulbanu.myrzahmedova@mail.ru Research Fellow, Master of science, Senior Lecturer.

Funding: The research was carried out with partial support of the Russian Science Foundation, grant No. 21-11-00202.

For citation: M. G. Dmitriev, Z. N. Murzabekov, and G. A. Mirzakhmedova, "Algorithm for Finding Feedback in a Problem with Constraints for One Class of Nonlinear Control Systems", *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 220-233, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

ALGORITHMS

Алгоритм нахождения обратной связи в задаче с ограничениями для одного класса нелинейных управляемых систем

М. Г. Дмитриев[1], З. Н. Мурзабеков [2], Г. А. Мирзахмедова [2]
DOI: 10.18255/1818-1015-2021-3-220-233

[1]Федеральный исследовательский центр "Информатика и управление" Российской академии наук, ул. Вавилова, д. 44/2, г. Москва, 119333 Россия.
[2]Казахский национальный университет имени аль-Фараби, проспект аль-Фараби, д. 71, г. Алматы, 050040 Казахстан.

УДК 517.977.55
Научная статья
Полный текст на русском языке

Получена 20 августа 2021 г.
После доработки 31 августа 2021 г.
Принята к публикации 1 сентября 2021 г.

Для непрерывной нелинейной управляемой системы на конечном интервале времени с ограничениями на управление, где правая часть уравнений динамики линейна по управлению и линеаризуема в окрестности нулевого положения равновесия рассматривается построение обратной связи по схеме алгоритма Калмана. Для этого используется решение вспомогательной задачи оптимального управления с квадратичным функционалом по аналогии с подходом SDRE.

Так как этот подход в литературе применяется для нахождения субоптимального синтеза в задачах оптимального управления с квадратичным функционалом с формально линейными системами, где все матрицы коэффициентов в дифференциальных уравнениях и в критерии могут содержать переменные состояния, то на конечном интервале времени здесь появляется необходимость решения усложненного матричного дифференциального уравнения Риккати, с матрицами коэффициентов зависящими от состояния. Это обстоятельство вследствие нелинейности системы, по сравнению с алгоритмом Калмана для линейно-квадратичных задач, значительно увеличивает количество вычислений для получения коэффициентов матрицы коэффициентов усиления в обратной связи и для получения синтеза с заданной точностью. Предложенный в работе алгоритм построения синтеза строится с помощью принципа расширения, предложенного В. Ф. Кротовым и развитого В. И. Гурманом, и позволяет не только расширить сферу использования подхода SDRE на нелинейные задачи управления с ограничениями на управление в виде замкнутых неравенств, но и предложить более эффективный вычислительный алгоритм нахождения матрицы коэффициентов усиления обратной связи в задачах управления на конечном интервале. В работе устанавливается корректность применения принципа расширения с помощью введения аналогов множителей Лагранжа, зависящих от состояния и времени, а также выводится формула субоптимального значения критерия качества. Приведенные теоретические результаты иллюстрируются на расчетах субоптимальных обратных связей в задачах управления трехсекторными экономическими системами.

Ключевые слова: задача оптимального управления; метод множителей Лагранжа; нелинейная система; квадратичный функционал; обратная связь; подход SDRE; трехсекторный экономический объект управления

ИНФОРМАЦИЯ ОБ АВТОРАХ

Михаил Геннадьевич Дмитриев
автор для корреспонденции
orcid.org/0000-0002-1184-129X. E-mail: mdmitriev@mail.ru
ГНС, доктор физ.-мат. наук, профессор.

Зайнелхриет Нугманович Мурзабеков
orcid.org/0000-0002-9074-4753. E-mail: murzabekov-zein@mail.ru
ГНС, доктор технических наук, профессор.

Гулбану Абсаматовна Мирзахмедова
orcid.org/0000-0001-7915-945X. E-mail: gulbanu.myrzahmedova@mail.ru
НС, магистр, старший преподаватель.

Финансирование: Исследование выполнено при частичной поддержке гранта РНФ № 21-11-00202.

Для цитирования: M. G. Dmitriev, Z. N. Murzabekov, and G. A. Mirzakhmedova, "Algorithm for Finding Feedback in a Problem with Constraints for One Class of Nonlinear Control Systems", *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 220-233, 2021.

Введение

Задачи и вопросы техники построения синтезирующих законов управления продолжают быть актуальными, в связи с необходимостью построения законов обратной связи в различных приложениях. Это происходит на фоне усложняющихся математических моделей из-за необходимости учета нелинейностей, возмущений, повышений размерности векторов состояния и управления и т. д. В связи с этим происходит постоянный поиск новых подходов к построению синтезирующих законов управления в нелинейных задачах и развитие имеющихся способов построения законов обратной связи в динамических системах. При этом наличие ограничений, конечно, осложняет поиск таких алгоритмов управления. Для приближенного решения задач синтеза в нелинейном случае с 90-х годов прошлого века в литературе активно развивается так называемый подход SDRE (см., например, [1—6]) для задач оптимального управления в классической постановке без ограничений на управление. Для приближенного решения нелинейных задач оптимального управления как на бесконечном интервале, так и на конечном, здесь правые части обыкновенных дифференциальных уравнений динамики сначала приводятся к формально линейному виду по состоянию и управлению, где коэффициенты всех матриц могут быть зависящими от состояния. Обратная связь, при этом, строится с помощью решения соответствующих линейно-квадратичных задач оптимального управления, где коэффициенты весовых матриц в критерии оптимальности также могут зависеть от переменных состояния. Затем матрица коэффициентов усиления регулятора находится с помощью решения матричных уравнений типа Риккати, как алгебраических для задач стабилизации на полуоси, так и дифференциальных, для задач управления на конечном интервале времени соответственно. Как показали многочисленные эксперименты, такой эвристический подход вследствие неоднозначности представления нелинейной системы в виде системы линейной структуры и погрешностей, возникающих при численном решении матричных уравнений Риккати, коэффициенты также зависят от состояния, порождает множество возможных субоптимальных решений. Но учитывая сложность построения и важность для приложений управления в виде законов обратной связи в нелинейных системах, подход SDRE получил широкое распространение в литературе при приближенном решении нелинейных задач оптимального управления без ограничений на управление.

Здесь, для одного класса нелинейных управляемых систем на конечном интервале времени, показывается возможность обоснования подхода SDRE при построении обратной связи в задачах с ограничениями на управление с помощью принципа расширения, предложенного В. Ф. Кротовым [7] и развитым В. И. Гурманом [8]. При этом, предлагается более эффективный алгоритм с точки зрения объема вычислений, не требующий многократного интегрирования матричных дифференциальных уравнений Риккати с коэффициентами, зависящими от состояния. Отметим, что впервые применение принципа расширения Кротова в рамках подхода SDRE, иллюстрировалось в [9] для задачи построения стабилизирующего регулятора без ограничений на управление на бесконечном интервале времени.

В конце настоящей работы приводятся результаты вычислительного эксперимента, иллюстрирующие предлагаемый алгоритм построения обратной связи на примере решения задачи оптимального управления для трехсекторной нелинейной модели экономического объекта, с ограничениями на управление в виде замкнутых неравенств.

1. Теоретические результаты

Пусть нелинейная управляемая система имеет вид

$$\dot{y}(t) = Ay(t) + B(y)u(t) + h(y), \quad y(t_0) = y_0, \quad t \in [t_0, T], \tag{1}$$

где $y(t) \in R^n$, $u(t) \in R^m$ вектор-функции состояния и управления, $n \times n$ матрица A – постоянная, коэффициенты $n \times m$ матрицы $B(y)$ и компоненты вектора $h(y)$ ограничены и непрерывно диффе-

ренцируемы по $y(t) \in R^n$, а управление $u \in R^m$ – кусочно-непрерывная вектор-функция; t_0, T – заданные начальный и конечный момент времени, управление $u(t)$ в каждый момент времени, удовлетворяет ограничениям

$$u(t) \in U(t) = \{u : \gamma_1(t) \le u(t) \le \gamma_2(t),\ t \in [t_0, T]; \quad \gamma_1,\ \gamma_2 \in C^m[t_0, T]\}. \tag{2}$$

Вид (1) связан с рассмотрением нелинейных задач управления, когда исходная система уравнений динамики может иметь ненулевое положение равновесия x^s при некотором постоянном управлении v^s, а неоднородность $h(y)$ связана с преобразованием $y = x - x^s$ и выделением линейной части по y, т.е. $h(0) = 0$.

Будем искать управление, на основе подхода SDRE, используя критерий оптимальности

$$J(u) = \frac{1}{2} y'(T) F y(t) + \frac{1}{2} \int_{t_0}^{T} (y' Q(y) y + u' R u)\, dt, \tag{3}$$

где матрица Q предполагается положительно полуопределенной при всех допустимых y, а матрицы F, R постоянные и положительно определенные.

Теперь задача состоит в том, чтобы найти управление $u(t)$ в виде обратной связи путем подбора матрицы $Q(y)$ так, чтобы достигался минимум критерия (3) вдоль траекторий системы (1) при выполнении ограничений (2). Такой подход связан, с одной стороны, со стремлением, чтобы вдоль такого управления траектории замкнутой системы, как минимум, оставались ограниченными, а, с другой стороны, являлись бы рациональными в некотором смысле.

Заметим, что весовая матрица $Q(y)$ в критерии будет представляться в виде

$$Q(y) = (K + W^{-1}) B(y) R^{-1} B'(y) (K + W^{-1}) - (K + W^{-1}) B_s R^{-1} B_s' (K + W^{-1}) + Q_1, \tag{4}$$

где $B_s = B(0)$, Q_1 – некоторая положительно определенная постоянная матрица и будет подбираться так, чтобы матрица $Q(y)$ при всех y была бы положительно полуопределенной, матрица K являлась бы решением некоторого матричного алгебраического уравнения Риккати с постоянными коэффициентами, а невырожденная матрица W будет определяться ниже.

Итак, будем искать оптимальное управление в форме обратной связи $u(y, t)$ в задаче (1)–(3). Для решения используем принцип расширения [7, 8], который состоит в том, что исходная задача оптимального управления с ограничениями сводится к задаче без ограничений и при этом решение новой задачи является одновременно и решением первоначальной задачи [9—13]. Для этого задачу (1)–(3) заменяем задачей без ограничений с использованием множителей Лагранжа $\lambda_1(t)$, $\lambda_2(t)$, $\lambda_3(t)$, $\mu(y, t)$. Неотрицательные функции $\lambda_1(t)$, $\lambda_2(t)$ отвечают ограничениям на управление, множитель Лагранжа $\mu(y, t)$ вводится для учета дифференциальных связей в (1) и ищется в виде $\mu(y, t) = Ky + q(y, t)$, а функция $\lambda_3(t)$ отвечает вводимой связи типа равенства $y(t) - W(t)q(y, t) = 0$.

Итак, функционал (3) по схеме [7, 8] заменяем следующим

$$\begin{aligned} L(y, u) = \frac{1}{2} y'(T) F y(T) + \int_{t_0}^{T} \{ & \frac{1}{2} y' Q(y) y + u' R u + \mu'(y, t)(Ay + B(y)u + h(y) - \\ & - \dot{y}) + \lambda_1'[\gamma_1 - u] + \lambda_2'[u - \gamma_2] + \lambda_3'[y - W(t)q)] \}\, dt, \end{aligned} \tag{5}$$

и введем функции

$$\begin{aligned} M(y, u, t) = \frac{1}{2} y' Q(y) y + \frac{1}{2} u' R u + (Ky + q)'(Ay + B(y)u + \\ + h(y)) + y' \dot{q} + \lambda_1'[\gamma_1 - u] + \lambda_2'[u - \gamma_2] + \lambda_3'[y - W(t)q], \end{aligned} \tag{6}$$

$$G(y_0, y(T)) = \frac{1}{2}y'(T)(F-K)y(T) - y'(T)q(T) + \frac{1}{2}y(t_0)'Ky(t_0) + y(t_0)'q(t_0), \tag{7}$$

$$v(y,t) = \frac{1}{2}y'Ky + y'q(y,t). \tag{8}$$

Перепишем (5) в виде

$$L(y,u) = \frac{1}{2}y'(T)Fy(T) + [v(y_0,t_0) - v(y(T),T)] + \int_{t_0}^{T} M(y,u,t)\,dt. \tag{9}$$

Введем множество всех допустимых управлений, удовлетворяющих условию $u(t) \in U(t)$, $t \in [t_0, T]$ и соответствующих им траекторий $y(t)$ системы (1), определенных на отрезке $t_0 \le t \le T$, т. е. множество всех допустимых пар $\{y(t), u(t)\}$, которое обозначим

$$\Delta(t_0, T, y_0) = \{(y,u) : u(t) \in U(t),\ \dot{y}(t) = Ay + B(y)u(t) + h(y),\ y(t_0) = y_0,\ t_0 \le t \le T\}. \tag{10}$$

Учитывая (7), $L(y,u)$ принимает вид

$$L(y,u) = G(y_0, y(T)) + \int_{t_0}^{T} M(y(t), u(t), t)\,dt, \tag{11}$$

где исходная задача с ограничениями сведена к другой задаче, но уже без ограничений.

Теперь обратимся к непрерывно дифференцируемой функции $v(y,t)$ в (8), и вычислим ее полную производную по времени $\frac{dv}{dt} = (\frac{\partial v}{\partial y})'\dot{y} + \frac{\partial v}{\partial t}$.

Ниже будем использовать обозначения $M_{min} \equiv inf\{M(y,u,t),\ \{y,u\} \in \Delta(t_0,T,y_0)\}$, $G_{min} \equiv inf\{G(y_0,y(T)),\ \{y_0,y(T)\} \in \Delta(t_0,T,y_0)\}$ при почти всех $t \in [t_0, T]$.

Имеет место

Лемма. Пусть выполнены условия:

1) Пара $\{y(t), u(t)\} \in \Delta(t_0, T, y_0)$;
2) Существуют $\lambda_1(t) \ge 0$, $\lambda_2(t) \ge 0$ такие, что на оптимальной паре $(\tilde{y}(t), \tilde{u}(t)) \in \Delta(t_0, T, y_0)$, которая доставляет минимальное значение функционалу (11), выполняются условия

$$\lambda_1'(\gamma_1 - \tilde{u}) = 0,\ \lambda_2'(\tilde{u} - \gamma_2) = 0. \tag{12}$$

Тогда справедливы следующие соотношения

$$L(y,u) = \int_{t_0}^{T} M(y(t), u(t), t)\,dt + G(y_0, y(T)) \le J(u), \quad L(\tilde{y}, \tilde{u}) = J(\tilde{u}), \tag{13}$$

где $(y(t), u(t))$, $(\tilde{y}(t), \tilde{u}(t))$ допустимая и оптимальная пары соответственно.

Доказательство. Для функции $v(y,t)$ из (8) имеем

$$\frac{dv(y(t),t)}{dt} = \frac{\partial v(y(t),t)}{\partial y}\dot{y} + \frac{\partial v(y(t),t)}{\partial t} = (Ky+q)'\dot{y} + y'\frac{dq(y(t),t)}{dt}. \tag{14}$$

Интегрируя (14) и суммируя полученное выражение с функционалом (5), получаем

$$L(y,u) = \int_{t_0}^{T} \{\frac{1}{2}y'Q(y)y + u'Ru + (Ky+q)'(Ay + B(y)u + h(y)) + y'\dot{q} - \\ - \frac{dv}{dt} + \lambda_1'[\gamma_1 - u] + \lambda_2'[u - \gamma_2] + \lambda_3'[y - W(t)q]\}\,dt + \frac{1}{2}y'Fy(T).$$

Отсюда имеем, с учетом (6) и (7), что $L(y,u) = \int_{t_0}^{T} M(y(t),u(t),t)\,dt + G(y_0,y(T))$ и

$$G(y_0,y(T)) = \frac{1}{2}y'(T)(F-K)y(T) - y'(T)q(T) + \frac{1}{2}y'(t_0)Ky(t_0) + y'(t_0)q(t_0),$$

Покажем, что на допустимых решениях выполняется неравенство $L(y(t),u(t)) \le J(u)$. Действительно, пусть $(y(t),u(t)) \in \Delta(t_0,T,y_0)$, тогда в силу (1) из (5), учитывая $\lambda_1'(\gamma_1 - u) \le 0$, $\lambda_2'(u-\gamma_2) \le 0$, $\lambda_3'[y - W(t)q] = 0$, получаем

$$\begin{aligned} L(y,u) = \int_{t_0}^{T} \{\frac{1}{2}y'Q(y)y + \frac{1}{2}u'Ru + (Ky+q)'(Ay + B(y)u + h(y) - \\ - \dot{y}) + \lambda_1'[\gamma_1 - u] + \lambda_2'[u-\gamma_2] + \lambda_3'[y - W(t)q]\}\,dt + \frac{1}{2}y'(T)Fy(T) = \\ = J(u) + \int_{t_0}^{T} \{\lambda_1'[\gamma_1 - u] + \lambda_2'[u-\gamma_2] + \lambda_3'[y - W(t)q]\}\,dt \le J(u). \end{aligned} \tag{15}$$

Из цепочки рассуждений в (15) видно, что вдоль оптимальной пары $(\tilde{y}(t),\tilde{u}(t)) \in \Delta(t_0,T,y_0)$, учитывая (12), выполняется следующее равенство

$$L(\tilde{y}(t),\tilde{u}(t)) = \int_{t_0}^{T} [\lambda_1'(\gamma_1 - u) + \lambda_2'(u-\gamma_2)]\,dt + J(\tilde{u}) = J(\tilde{u})$$

и, таким образом, утверждение леммы имеет место.

Для определения пары $\{\tilde{y}(t),\tilde{u}(t)\}$, минимизирующей функционал (11), необходимо найти управление $u(t)$ и определить множители $\mu(y,t) = Ky + q(y,t)$, $\lambda_1(t)$, $\lambda_2(t)$ так, чтобы при каждом фиксированном $t \in (t_0,T)$ подынтегральная функция $M(y,u,t)$ в (11) достигала наименьшего значения среди $(y,u) \in \Delta(t_0,T,y_0)$.

Учитывая, что при $t = T$ функция $G(y_0,y(T))$ принимает минимальное значение, а пара $\{\tilde{y}(t),\tilde{u}(t)\}$ удовлетворяет (1) при выполнении ограничений (2), тогда из необходимых условий минимума по управлению для функции $M(y,u,t)$ получаем выражение для управления

$$u = -R^{-1}B'(y)(Ky+q) - R^{-1}(\lambda_2 - \lambda_1), \tag{16}$$

а из минимизации терминальной части (7) функционала (11) находим $q|_{t=T} = (F-K)y(T)$.

Теперь определим неизвестные матрицы K, $W(t)$ и вектор-функцию q как решения на отрезке $t \in [t_0,T]$ следующих уравнений

$$KA + A'K - KB_sR^{-1}B_s'K + Q_1 = 0, \tag{17}$$

$$\dot{W} = WA_{1,s}' + A_{1,s}W - B_{1,s}, \quad W(T) = (F-K)^{-1}, \tag{18}$$

$$\dot{q} = -(A - B_sR^{-1}B_s'K)'q + W^{-1}(t)(h(y) - (B_1(y) - B_{1,s})(Ky+q) + B(y)\varphi(y,t)), \tag{19}$$

$$q(y,T) = W^{-1}(T)y(T),$$

где постоянные матрицы $A_{1,s}$, $B_{1,s}$ имеют вид $A_{1,s} = A - B_sR^{-1}B_s'K$, $B_{1,s} = B_sR^{-1}B_s'$, и предполагается, что K положительно определенная матрица, матрица $W(t)$ невырожденная, постоянная матрица F такая, что $F >> K$, а под $\dot{q}$ понимается полная производная по времени.

Покажем, что если существуют решения уравнений (17), (18), тогда замкнутая система вдоль управления (16) принимает вид

$$\dot{y} = A_1(y)y - B_1(y)q + B(y)\varphi(y,t) + h(y), \qquad y(t_0) = y_0, \tag{20}$$

где

$$A_1(y) = A - B(y)R^{-1}B(y)'K, \qquad B_1(y) = B(y)R^{-1}B(y)', \qquad \varphi(y,t) = R^{-1}[\lambda_1(y,t) - \lambda_2(y,t)],$$
$$\lambda_1(y,t) = Rmax\{0;\ \gamma_1 - \omega(y,t)\} \geq 0, \qquad \lambda_2(y,t) = Rmax\{0;\ \omega(y,t) - \gamma_2\} \geq 0, \qquad (21)$$
$$\omega(y,t) = -R^{-1}B'(y)(Ky + q).$$

Действительно, т.к. $q = q(y,t),\ q|_{t=T} = (F-K)y(T)$ и мы ее ввели как функцию $q(y,t) = W^{-1}(t)y(t)$ или $y \equiv Wq$, то вычисляя теперь $\dot{y} \equiv \dot{W}q + W\dot{q}$ и учитывая уравнения (17)-(19), получаем

$$A_1(y)y(t) - B_1(y)q(t) + B(y)\varphi(y,t) + h(y) = (WA'_{1,s} + A_{1,s}W - B_{1,s})q + W\dot{q}.$$

После преобразования левой части, используя уравнения в (17), (18) имеем

$$(A_1(y) - A_{1,s})Wq + A_{1,s}Wq - (B_1(y) - B_{1,s})q - B_{1,s}q + B(y)\varphi(y,t) + h(y) =$$
$$= (WA'_{1,s} + A_{1,s}W - B_{1,s})q + W\dot{q},$$

$$-(B_1(y) - B_{1,s})(Ky + q) + B(y)\varphi(y,t) + h(y) = WA'_{1,s} + W\dot{q},$$

Отсюда находим, что

$$\dot{q} = -(A_{1,s})'q + W^{-1}(t)(h(y) - (B_1(y) - B_{1,s})(Ky + q) + B(y)\varphi(y,t)),$$
$$q|_{t=t_0} = W^{-1}(t_0)y(t_0). \qquad (22)$$

Результаты, установленные для задачи (1)–(3), сформулируем в виде следующего утверждения.

Теорема 1. Пусть выполняются условия леммы и дополнительно предположим, что:

1. Существует постоянная положительно определенная матрица Q_1, что матрица $Q(y)$ при всех y является положительно полуопределенной.
2. Пара постоянных матриц $\{A, B_s\}$ удовлетворяет условию управляемости.
3. Постоянная положительно определенная матрица F такая, что $F - K$ также является положительно определенной матрицей.
4. Функция $q(y,t)$ удовлетворяет соотношению (22).

Тогда

a) существует положительно определенное решение матричного алгебраического уравнения Риккати (17);

b) решение $W(t)$ дифференциального уравнения (18) существует единственно и является невырожденной матрицей при всех $t \in [t_0, T]$;

c) оптимальное управление $u(y,t)$ в задаче (1)–(3) имеет вид

$$u(y,t) = -R^{-1}B'(y)(Ky + q(y,t)) + \varphi(y,t), \qquad (23)$$

где $q(y,t),\ \varphi(y,t)$ определяется в (19), (21).

Доказательство. Учитывая утверждения леммы, вследствие первых двух условий теоремы, подынтегральное выражение в (3) неотрицательное при всех допустимых парах (y,u). Поэтому из необходимых условий оптимальности $\dfrac{\partial M}{\partial u} = 0$ имеем, что экстремальное управление представляется как $u = -R^{-1}[B'(y)(Ky+q) - \lambda_1 + \lambda_2]$, которое с учетом обозначений в (21) принимает вид (23). Далее определяем множители $\lambda_1 \geq 0,\ \lambda_2 \geq 0$ таким образом, чтобы, с одной стороны, выполнялись условия

$$\lambda'_1(\gamma_1 - \tilde{u}) = 0, \quad \lambda'_2(\tilde{u} - \gamma_2) = 0, \qquad (24)$$

а с другой, осуществляем выбор $\lambda_1,\ \lambda_2,\ \varphi$ так, что имеют место представления

$$\lambda_1(y,t) = -Rinf(0,\ \omega(y,t) - \gamma_1(t)), \quad \lambda_2(y,t) = -Rinf(0,\ \gamma_2(t) - \omega(y,t)) \tag{25}$$

и

$$\varphi(y,t) = -inf(0,\ \omega(y,t) - \gamma_1(t)) + inf(0,\ \gamma_2(t) - \omega(y,t)). \tag{26}$$

Теперь определим функцию $M(y,u,t)$ на управлении (23). Подставляя управление (23) в выражение (6) и группируя подобные члены, в результате получим функцию

$$P(y,t) = \frac{1}{2}y'Q(y)y + (Ky+q)'Ay - \frac{1}{2}(Ky+q)'B(y)R^{-1}B'(y)(Ky+q) +$$
$$+ \frac{1}{2}\varphi'R\varphi + (Ky+q)'h(y) + y'\dot{q} = \frac{1}{2}y'[Q_1 + KA + A'K - KB_sR^{-1}B_s'K]y +$$
$$+ y'(A_{1s}'q + \dot{q}) + \frac{1}{2}\varphi'R\varphi + (Ky+q)'h(y) - \frac{1}{2}q'B_{1s}q.$$

Тогда вдоль положительно определенной матрицы K из (17) и функции $q(t,y)$ из (22), для $\{\tilde{y}(t),\tilde{u}(t)\} \in \Delta(t_0,T,y_0)$ имеем

$$L(\tilde{y},\tilde{u}) = \int_{t_0}^{T}\{y'(A_{1s}'q + \dot{q}) + \frac{1}{2}\varphi'R\varphi + (Ky+q)'h(y) - \frac{1}{2}q'B_{1s}q\}\,dt + G(y_0,\tilde{y}(T)), \tag{27}$$

где в точке $t = T$ функция $G(y_0,\tilde{y}(T))$ в (7) также принимает минимальное значение, что определяет минимальное значение (27) в целом.

Итак, имеем

$$M(\tilde{y}(t),\tilde{u}(t),t) = q'(h(y) - (B_1(y) - B_{1s})(Ky+q) + B(y)\varphi(y,t)) + \frac{1}{2}\varphi'R\varphi + (Ky+q)'h(y) - \frac{1}{2}q'B_{1s}q,$$

$$G(y_0,\tilde{y}(T)) = \frac{1}{2}y'(T)(F-K)y(T) + \frac{1}{2}y'(t_0)Ky(t_0) + y'(t_0)q(t_0). \tag{28}$$

Теперь нетрудно показать, что управление (23) является оптимальным, т. е. вдоль него критерий (3) принимает минимальное значение. Пусть существуют функции $\lambda_1(y,t)$, $\lambda_2(y,t)$ удовлетворяющие условию теоремы и произвольная допустимая пара $(y(t),u(t)) \in \Delta(t_0,T,y_0)$, тогда согласно (24) будем иметь

$$J(u) - J(\tilde{u}) \geq L(y(t),u(t)) - L(\tilde{y}(t),\tilde{u}(t)) =$$
$$= \int_{t_0}^{T}[M(y(t),u(t),t) - M(\tilde{y}(t),\tilde{u}(t),t)]\,dt + G(y_0,y(T)) - G(y_0,\tilde{y}(T)) =$$
$$= \int_{t_0}^{T}[M(y(t),u(t),t) - M_{min}]\,dt + G(y_0,y(T)) - G_{min} \geq 0,$$

отсюда следует $J(\tilde{u}) = L(\tilde{y}(t),\tilde{u}(t)) = inf_{\langle x,u\rangle\in\Delta}L(y(t),u(t))$ и, таким образом, теорема доказана.

Замечание 1. Прием в подходе SDRE с подбором Q_1, чтобы матрица подобная $Q(y)$ была положительно определенной при всех значениях вектора состояния, демонстрировался в [14].

Имеет место

Следствие 1. В условиях теоремы 1 вдоль оптимального управления (23) минимальное значение критерия в задаче (1)–(3) имеет вид

$$J(\tilde{u}) = L_{min}(y,u) = \int_{t_0}^{T}[\frac{1}{2}\varphi'R\varphi + y'(K + W^{-1})h(y)]\,dt + \frac{1}{2}y_0'(K + W_0^{-1})y_0. \tag{29}$$

Доказательство. В условиях теоремы 1 справедливо выражение (27). Учитывая (23) и (24) вычислим производную от вспомогательной квадратичной формы $\frac{1}{2}q'Wq$. Имеем

$$\frac{d}{dt}(\frac{1}{2}q'Wq) = q'W[-(A_{1s})'q + W^{-1}(t,T)(h(y) - (B_1(y) - B_{1s})(Ky+q) + B(y)\varphi(y,t))] + \\ + \frac{1}{2}q'[WA'_{1s} + A_{1s}W - B_{1s}]q, \tag{30}$$

и преобразуем минимальное значение функционала (27) следующим образом. Сначала подставим (30) в выражение (27). Группируя члены, получаем

$$L_{min}(y,u) = \int_{t_0}^{T}[\frac{1}{2}\varphi'R\varphi + y'(K + W^{-1})h(y)]\,dt + \frac{1}{2}y'_0(K + W_0^{-1})y_0. \tag{31}$$

Теперь преобразуем минимальное значение функционала (3). Для этого в (3) подставляем оптимальное управление (23) и выражение для $Q(y)$ из (4).

$$J(\tilde{u}) = \int_{t_0}^{T}[\frac{1}{2}y'Q(y)y + \frac{1}{2}\tilde{u}'R\tilde{u}]\,dt + \frac{1}{2}y'(T)Fy(T) = \\ = \int_{t_0}^{T}[\frac{1}{2}y'Q_1y - \frac{1}{2}(Ky+q)'B_{1s}(Ky+q) + (Ky+q)'B_1(y)(Ky+q)]\,dt + \\ + \int_{t_0}^{T}[-(Ky+q)'B(y)\varphi + \frac{1}{2}\varphi'R\varphi]\,dt + \frac{1}{2}y'(T)Fy(T).$$

Затем, используя дифференциальное уравнение в (20) и алгебраическое матричное уравнение Риккати (17), последнее выражение преобразуется к виду

$$J(\tilde{u}) = \int_{t_0}^{T}[\frac{1}{2}y'Q_1y - \frac{1}{2}(Ky+q)'B_{1s}(Ky+q) + (Ky+q)(-\dot{y} + Ay + h(y)) + \frac{1}{2}\varphi'R\varphi]\,dt + \\ + \frac{1}{2}y'(T)Fy(T) = \int_{t_0}^{T}[y'A'_{1s}q - \frac{1}{2}q'B_{1s}q + (Ky+q)(-\dot{y} + Ay + h(y)) + \frac{1}{2}\varphi'R\varphi]\,dt + \frac{1}{2}y'(T)Fy(T). \tag{32}$$

Теперь, вычисляя производную от квадратичной формы $\frac{1}{2}y'(K+W^{-1})y$ с учетом дифференциальных связей в (18), имеем

$$\frac{d}{dt}(\frac{1}{2}y'(K + W^{-1})y) = y'(K + W^{-1})\dot{y} - \frac{1}{2}y'(W^{-1}\dot{W}W^{-1})y = \\ = y'(K + W^{-1})\dot{y} - \frac{1}{2}q'[WA'_{1s} + A_{1s}W - B_{1s}]q, \tag{33}$$

Далее, преобразуем минимальное значение функционала (32) используя (33) и получаем

$$J(\tilde{u}) = \int_{t_0}^{T}[(Ky+q)'h(y) + \frac{1}{2}\varphi'R\varphi]\,dt + \frac{1}{2}y'(T)Fy(T) - \frac{1}{2}y'(T)(K + W^{-1}(T))y(T) + \\ + \frac{1}{2}y'(t_0)(K + W^{-1}(T))y(t_0).$$

Теперь, учитывая представление $q(y,t) = W^{-1}(t)y(t)$ и условие, что $q|_{t=T} = (F-K)y(T)$ получаем, что формула (29)

$$J(\tilde{u}) = \frac{1}{2}y'_0(K + W^{-1}(t_0))y_0 + \int_{t_0}^{T}[\frac{1}{2}\varphi'R\varphi + y'(K + W^{-1}(t))h(y)]\,dt$$

имеет место.

Замечание 2. Отметим, что формула (29) включает в себя известное представление начального условия для функции Беллмана в задаче оптимальной стабилизации для стационарной линейно квадратичной задачи на полуоси без ограничений на управление, и в ней еще, естественно, отражается возрастание минимального значения из-за наличия ограничений на управление, а также учитывается влияние неоднородности $h(y)$.

Опишем алгоритм решения задачи оптимального управления (1)–(3).

1. Находится положительно определенная матрица Q_1 такая, что $Q(y)$ также будет положительно определенной матрицей при всех y.
2. Решаются системы алгебраических и дифференциальных уравнений (17) и (18) для определения матриц K и $W(t)$ на отрезке $[t_0, T]$.
3. Задаются условия $y(t_0) = y_0$, и вычисляется $q(t_0) = W^{-1}(t_0)y(t_0)$.
4. Интегрируются на отрезке $[t_0, T]$ системы дифференциальных уравнений (20), (22) при начальных условиях $y(t_0) = y_0, \quad q(t_0) = W^{-1}(t_0)y(t_0)$.

2. Численные эксперименты

Рассматривается задача оптимального управления для экономической модели объекта управления, состоящего из трех секторов: $i = 0$ (материальный сектор), $i = 1$ (фондосоздающий сектор), $i = 2$ (потребительский сектор), математическая модель включает [15, 16]:

a) три дифференциальных уравнения, описывающих динамику фондовооруженностей:

$$\dot{k}_i = -\lambda_i k_i + \frac{s_i}{\theta_i} x_1, \quad k_i(0) = k_i^0, \lambda_i > 0, \quad (i = 0,\ 1,\ 2); \tag{34}$$

b) три функции удельного выпуска типа Кобба-Дугласа:

$$x_i = \theta_i A_i k_i^{\alpha_i}, \quad A_i > 0, \quad 0 < \alpha_i < 1 \quad (i = 0,\ 1,\ 2), \tag{35}$$

c) три балансовых соотношения:

$$s_0 + s_1 + s_2 = 1, \quad s_0 \geq 0, \quad s_1 \geq 0, \quad s_2 \geq 0, \tag{36}$$

$$\theta_0 + \theta_1 + \theta_2 = 1, \quad \theta_0 \geq 0, \quad \theta_1 \geq 0, \quad \theta_2 \geq 0, \tag{37}$$

$$(1 - \beta_0)x_0 = \beta_1 x_1 + \beta_2 x_2, \quad \beta_0 \geq 0, \quad \beta_1 \geq 0, \quad \beta_2 \geq 0. \tag{38}$$

Здесь состояние экономической системы (фондовооруженность) описывается вектором (k_0, k_1, k_2), а $(s_0, s_1, s_2, \theta_0, \theta_1, \theta_2)$ – вектор управлений, (s_0, s_1, s_2) – доли секторов в распределении инвестиционных ресурсов, $(\theta_0, \theta_1, \theta_2)$ - доли секторов в распределении трудовых ресурсов; x_i – удельный выпуск продукции в соответствующем секторе; β_i – прямые материальные затраты в i-ом секторе; $i = 0, 1, 2$. Начальное состояние системы равно (k_0^0, k_1^0, k_2^0), где $k_i^0 = k_i(0)$. Для решения задачи перевода нелинейной системы из начального состояния в желаемое состояние на отрезке времени $[0, T]$. В качестве желаемого конечного состояния (k_0^s, k_1^s, k_2^s) используется состояние равновесия системы, которое определено в работе [12] в следующем виде:

$$k_0^s = \frac{s_0 \theta_1 A_1 (k_1^s)^{\alpha_1}}{\lambda_0 \theta_0}, \quad k_1^s = \left(\frac{s_1 A_1}{\lambda_1}\right)^{\frac{1}{1-\alpha_1}}, \quad k_2^s = \frac{s_2 \theta_1 A_1 (k_1^s)^{\alpha_1}}{\lambda_2 \theta_2}. \tag{39}$$

Значения фондовооруженностей $k_i^s, (i = 0, 1, 2)$ в стационарном состоянии (39) зависят от управлений $(s_0, s_1, s_2, \theta_0, \theta_1, \theta_2)$, для которых в работе [12] получены стационарные значения $(s_0^s, s_1^s, s_2^s, \theta_0^s, \theta_1^s, \theta_2^s)$. Приведем теперь математическую модель объекта управления (34) к виду (1) и запишем в виде системы

$$\dot{y}(t) = Ay(t) + B(y)u(t) + D(y)v^s \quad y(t_0) = y_0, \quad t \in [t_0, T]. \tag{40}$$

используя следующие обозначения $y_1 = k_1 - k_1^s, \quad y_2 = k_2 - k_2^s, \quad y_3 = k_0 - k_0^s, \quad u_1 = s_1 - v_1^s,$

$$u_2 = \frac{s_2\theta_1}{\theta_2} - v_2^s, \quad u_3 = \frac{s_0\theta_1}{\theta_0} - v_3^s, \quad v_1^s = s_1^s, \quad s_2^s\theta_1^s/\theta_2^s = v_2^s, \quad s_0^s\theta_1^s/\theta_0^s = v_3^s,$$

$$f_1(y_1) = (y_1 + k_1^s)^{\alpha_1}, \quad f_2(y_2) = (y_2 + k_2^s)^{\alpha_2}, \quad f_3(y_3) = (y_3 + k_0^s)^{\alpha_0},$$

$$A = \begin{pmatrix} -\lambda_1 + b\alpha_1(k_1^s)^{\alpha_1-1}v_1^s & 0 & 0 \\ b\alpha_1(k_1^s)^{\alpha_1-1}v_2^s & -\lambda_2 & 0 \\ b\alpha_1(k_1^s)^{\alpha_1-1}v_3^s & 0 & -\lambda_0 \end{pmatrix}, \quad B = \begin{pmatrix} b(y_1 + k_1^s)^{\alpha_1} & 0 & 0 \\ 0 & b(y_1 + k_1^s)^{\alpha_1} & 0 \\ 0 & 0 & b(y_1 + k_1^s)^{\alpha_1} \end{pmatrix},$$

$$B(0) = B_s, \quad D(y) = \begin{pmatrix} d(y_1) & 0 & 0 \\ 0 & d(y_1) & 0 \\ 0 & 0 & d(y_1) \end{pmatrix}, \quad D(y_1) = b(y_1 + k_1^s)^{\alpha_1} - (k_1^s)^{\alpha_1} - \alpha_1(k_1^s)^{\alpha_1-1}y_1.$$

Здесь постоянные значения k^s и v^s определяются в (40) и для них имеет место алгебраическая связь

$$Ak^s + BD(k^s)v^s = 0.$$

Для применения алгоритма решения задачи в примере дополнительно введем функции

$$f_1(y_1) = (y_1 + k_1^s)^{\alpha_1}, \quad f_2(y_2) = (y_2 + k_2^s)^{\alpha_2}, \quad f_3(y_3) = (y_3 + k_3^s)^{\alpha_3},$$

$$\xi = \frac{\beta_1 A_1 f_1(y_1) + \beta_2 A_2 f_2(y_2)(1 - u_1 - v_1^s)/(u_2 + v_2^s)}{(1 - \beta_0)A_0 f_3(y_3)(1 - u_1 - v_1^s)/(u_3 + v_3^s) + \beta_2 A_2 f_2(y_2)(1 - u_1 - v_1^s)/(u_2 + v_2^s)}, \tag{41}$$

которые обеспечивают выполнение условия (38); при этом инвестиционные ресурсы

$$s_1 = u_1 + v_1^s, \quad s_2 = (1 - \xi)(1 - u_1 - v_1^s), \quad s_0 = \xi(1 - u_1 - v_1^s), \tag{42}$$

обеспечивают выполнение условия (36); а трудовые ресурсы

$$\theta_1 = \frac{1}{1 + (s_0)/(u_3 + v_3^s) + (s_2)/(u_2 + v_2^s)}, \quad \theta_2 = \frac{(1 - \xi)(1 - s_1)\theta_1}{u_2 + v_2^s}, \quad \theta_0 = \frac{\xi(1 - s_1)\theta_1}{u_3 + v_3^s} \tag{43}$$

обеспечивают выполнение условия (37).

Были проведены численные расчеты на компьютере при следующих значениях параметров:

$$\alpha_0 = 0,46; \quad \alpha_1 = 0,68; \quad \alpha_2 = 0,49; \quad \beta_0 = 0,39; \quad \beta_1 = 0,29; \quad \beta_2 = 0,52;$$

$$\lambda_i = 0,05,\ i = 0,1,2; \quad A_0 = 6,19; \quad A_1 = 1,35; \quad A_2 = 2,71;$$

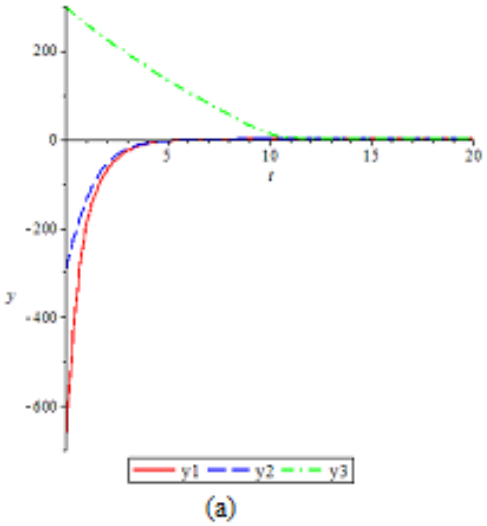
$$s_0^s = 0,2763; \quad s_1^s = 0,4476; \quad s_2^s = 0,2761; \quad \theta_0^s = 0,3944; \quad \theta_1^s = 0,2562; \quad \theta_2^s = 0,3494;$$

$$k_0^s = 966,4430; \quad k_1^s = 2410,1455; \quad k_2^s = 1090,1238;$$

Начальные условия задаются следующие $y(t_0) = (-700,\ -300,\ 300)'$, а матрицы $R,\ Q_1,\ K$ имеют вид

$$R = \begin{pmatrix} 350 & 0 & 0 \\ 0 & 70 & 0 \\ 0 & 0 & 55 \end{pmatrix}, \ Q_1 = \begin{pmatrix} 49 \cdot 10^{-4} & 0 & 0 \\ 0 & 9 \cdot 10^{-4} & 0 \\ 0 & 0 & 9 \cdot 10^{-4} \end{pmatrix}, \ K = \begin{pmatrix} 0,4598 \cdot 10^{-2} & 0,6894 \cdot 10^{-5} & 0,5129 \cdot 10^{-5} \\ 0,6894 \cdot 10^{-5} & 0,8888 \cdot 10^{-3} & -3,5598 \cdot 10^{-9} \\ 0,5129 \cdot 10^{-5} & -3,5598 \cdot 10^{-9} & 0,7926 \cdot 10^{-3} \end{pmatrix}.$$

Результаты расчетов состояния системы представлены на рисунке 1(а). Из рисунка 1(б) видно, что оптимальные управления не выходят за пределы области U, определяемой ограничениями $-0.3477 \le u_1 \le 0.4523, \quad -0.1024 \le u_2 \le 0.6976, \quad -0.07945 \le u_3 \le 0.7205.$



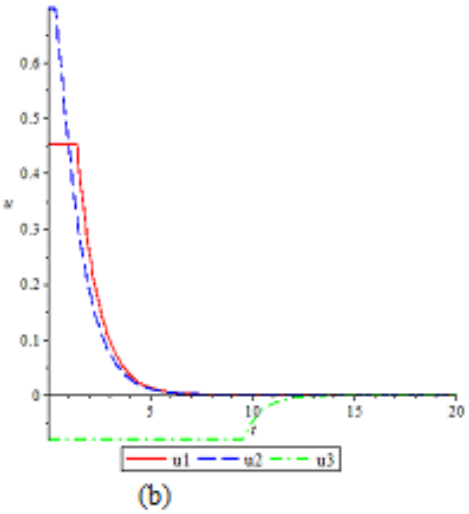
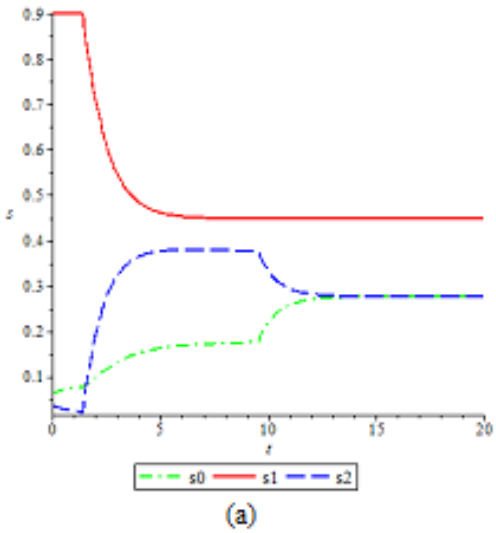


Fig. 1. Graphs of trajectories $y(t)$ (a) and the optimal control $u(t)$ (b)

Рис. 1. Графики траекторий $y(t)$ (a) и оптимального управления $u(t)$ (b)



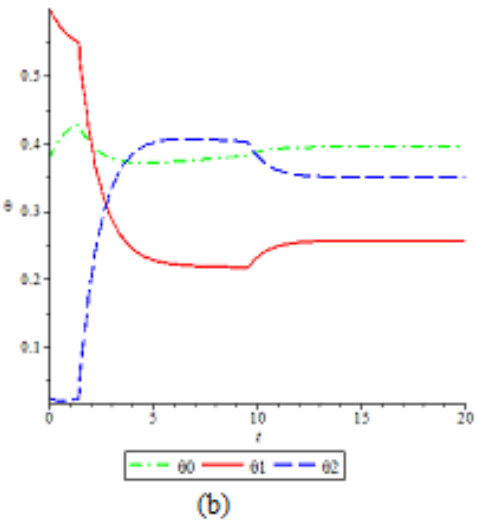


Fig. 2. Graphs of the distribution of investment (a) and labor resources (b)

Рис. 2. Графики распределения инвестиций (a) и трудовых ресурсов (b)

Здесь все компоненты управления $u_1(t)$, $u_2(t)$ и $u_3(t)$ лежат на границе области U на отрезке $[0, t_1]$, $[0, t_2]$ и $[0, t_3]$ соответственно, затем при $t \in (t_1, T]$, $t \in (t_2, T]$ и $t \in (t_3, T]$ заходят внутрь области U. Переключение управлений происходит в момент времени $t_1 = 1,439$ для компонента $u_1(t)$, а для $u_2(t)$ при $t_2 = 0,4$, для $u_3(t)$ при $t_3 = 9,785$. $y_1(T) = -4.8692 \cdot 10^{-6}$; $y_2(T) = -1.7315 \cdot 10^{-6}$; $y_3(T) = 0.229 \cdot 10^{-3}$, и оптимальные значения управлений в конечный момент времени при $T = 20$: $u_1(T) = 2.2873 \cdot 10^{-8}$; $u_2(T) = 6.5404 \cdot 10^{-9}$; $u_3(T) = -3.4357 \cdot 10^{-7}$. На рисунке 2 показаны изменения ресурсов, которые удовлетворяют балансовым соотношениям (36), (37). Значения инвестиций ($s_0(t)$, $s_1(t)$, $s_2(t)$) и трудовых ресурсов ($\theta_0(t)$, $\theta_1(t)$, $\theta_2(t)$) в конечный момент времени при $T = 20$ стремятся к стационарному состоянию.

Заключение

Для частного класса нелинейных управляемых систем на конечном интервале показывается возможность применения подхода SDRE при построении обратной связи на задаче с ограничениями на управление, применяя принцип расширения, предложенного В. Ф. Кротовым и развитый В. И. Гурманом, и при этом предлагается эффективный алгоритм, не требующий многократного интегрирования дифференциального уравнения Риккати с коэффициентами, зависящими от состояния.

Также приведены численные эксперименты, иллюстрирующие предлагаемый алгоритм построения оптимального синтеза при наличии ограничений на управление в виде замкнутых неравенств на примере трехсекторной нелинейной экономической системы.

References

[1] C. Mracek and J. Cloutier, "Control designs for the nonlinear benchmark problem via the state-dependent Riccati equation method", *International Journal of Robust and Nonlinear Control*, vol. 8, no. 4–5, pp. 401–433, 1998.

[2] J. R. Cloutier and D. T. Stansbery, "The Capabilities and Art of State-Dependent Riccati Equation-Based Design", in *Proceedings of the American Control Conference*, vol. 1, IEEE, Piscataway, May, 2002, pp. 86–91.

[3] V. Afanas'ev and P. Orlov, "Suboptimal control of a nonlinear object linearized by feedback", *Bulleten RAS. Control theory and systems*, no. 3, pp. 13–22, 2011, (In Russian).

[4] T. Cimen, "State-dependent Riccati Equation (SDRE) control: A Survey", *IFAC Proceedings Volumes*, vol. 41, no. 2, pp. 3761–3775, 2008.

[5] A. Heydari and S. N. Balakrishnan, "Path Planning Using a Novel Finite Horizon Suboptimal Controller", *Journal if Guidance, Control and Dynamics*, vol. 36, no. 4, pp. 1210–1214, 2013.

[6] A. Heydari and S. N. Balakrishnan, "Approximate closed-form solutions to finite-horizon optimal control of nonlinear systems", in *American Control Conference (ACC), IEEE*, 2012, pp. 2657–2662.

[7] V. F. Krotov and V. Gurman, *Methods and Problems of Optimal Control*. Nauka, Moscow, 1973.

[8] V. Gurman, *The principle of expansion in control tasks*. Moscow "Science" Main Editing of Physical-Mathematical Literature, 1985, (In Russian).

[9] M. Dmitriev, Z. Murzabekov, D. Makarov, and G. Mirzakhmedova, "SDRE based stabilization of the affine control system with the stationary linear part", in *23rd International Conference on System Theory, Control and Computing*, 2019, pp. 739–743.

[10] S. Aipanov and Z. Murzabekov, "Analytical solution of a linear quadratic optimal control problem with control value constraints", *Journal of Computer and Systems Sciences International*, vol. 1, no. 53, pp. 84–91, 2014.

[11] Z. Murzabekov, "The synthesis of the proportional-differential regulators for the systems with fixed ends of trajectories under two-sided constraints on control values", *Asian Journal of Control*, vol. 2, no. 18, pp. 494–501, 2016.

[12] Z. Murzabekov, M. Milosz, and K. Tussupova, "The optimal control problem with fixed-end trajectories for a three-sector economic model of a cluster", in *10th International scientific conferences on research and applications in the field of intelligent information and database systems, ACIIDS, Dong Hoi City*, 2018, pp. 382–391.

[13] Z. Murzabekov, M. Milosz, and K. Tussupova, "Modeling and optimization of the production cluster", in *Proceedings of 36th International Conference on Information Systems and Architecture and Technology - ISAT-2015 / Part II, Advances in Intelligent Systems and Computing. - Karpacz*, 2016, pp. 99–108.

[14] M. Dmitriev and D. A. Makarov, "A weak non-linear regulator in a weakly non-linear control system with efficiency", *Proceedings of the ISA RAS*, vol. 4, no. 64, pp. 53–58, 2014, (In Russian).

[15] V. A. Kolemaev, "Optimal balanced space of the open three-sector economy", *Applied econometrics*, vol. 3, no. 11, pp. 15–42, 2008, (In Russian).

[16] S. M. Aseev, K. Besov, and A. Kryazhimsky, "Optimal control problems on infinite time management in economics", *Advances in mathematical sciences*, vol. 404, no. 2, pp. 3–64, 2012, (In Russian).

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

ALGORITHMS

A Simple Algorithm for Finding a Non-negative Basic Solution of a System of Linear Algebraic Equations

G. D. Stepanov[1]

DOI: 10.18255/1818-1015-2021-3-234-237

[1]Voronezh State Pedagogical University, 86 Lenin str, Voronezh 394043, Russia.

MSC2020: 90C05
Research article
Full text in Russian

Received July 11, 2021
After revision August 24, 2021
Accepted August 25, 2021

This article describes an algorithm for obtaining a non-negative basic solution of a system of linear algebraic equations. This problem, which undoubtedly has an independent interest,
in particular, is the most time-consuming part of the famous simplex method for solving linear programming problems. Unlike the artificial basis Orden's method used in the classical simplex method, the proposed algorithm does not attract artificial variables and economically consumes computational resources.
The algorithm consists of two stages, each of which is based on Gaussian exceptions. The first stage coincides with the main part of the Gaussian complete exclusion method, in which the matrix of the system is reduced to the form with an identity submatrix. The second stage is an iterative cycle, at each of the iterations of which, according to some rules, a resolving element is selected, and then a Gaussian elimination step is performed, preserving the matrix structure obtained at the first stage. The cycle ends either when the absence of non-negative solutions is established, or when one of them is found.
Two rules for choosing a resolving element are given. The more primitive of them allows for ambiguity of choice and does not exclude looping (but in very rare cases). Use of the second rule ensures that there is no looping.

Keywords: linear equation systems; nonnegative solution; linear programming; the rule of choosing a resolving element

INFORMATION ABOUT THE AUTHORS

Gleb D. Stepanov
correspondence author
orcid.org/0000-0003-3237-849X. E-mail: stpnv42@mail.ru
PhD, associate professor.

For citation: G. D. Stepanov, "A Simple Algorithm for Finding a Non-negative Basic Solution of a System of Linear Algebraic Equations", *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 234-237, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

ALGORITHMS

Простой алгоритм отыскания неотрицательного базисного решения системы линейных алгебраических уравнений

Г. Д. Степанов[1]

DOI: 10.18255/1818-1015-2021-3-234-237

[1]Воронежский государственный педагогический университет, ул. Ленина, д. 86, г. Воронеж, 394043 Россия.

УДК УДК 519.852
Научная статья
Полный текст на русском языке

Получена 11 июля 2021 г.
После доработки 24 августа 2021 г.
Принята к публикации 25 августа 2021 г.

В данной статье описывается алгоритм получения неотрицательного базисного решения системы линейных алгебраических уравнений. Эта задача, в частности, является наиболее трудоемким этапом знаменитого симплекс-метода решения задач линейного программирования, хотя бесспорно представляет и самостоятельный интерес. В отличии от метода искусственного базиса Ордена, применяемого в классическом симплекс-методе, предлагаемый алгоритм не использует искусственных переменных и экономно расходует вычислительные ресурсы.

Алгоритм состоит из двух этапов, основу каждого из которых составляют Гауссовы исключения. Первый этап совпадает с основной частью метода полных исключений Гаусса, в котором матрица системы приводится к виду с единичной подматрицей. Второй этап представляет из себя итерационный цикл, на каждой из итераций которого по некоторым правилам выбирается разрешающий элемент, а затем выполняется шаг исключения Гаусса, сохраняющий структуру матрицы, полученную на первом этапе. Цикл завершается, либо когда будет установлено отсутствие неотрицательных решений, либо когда будет найдено одно из них.

Приводятся два правила выбора разрешающего элемента. Более примитивное из них допускает неоднозначность выбора и не исключает зацикливания (но в очень редких случаях). Использование второго правила гарантирует отсутствие зацикливания.

Ключевые слова: система линейных алгебраических уравнений; неотрицательное решение; линейное программирование; правило выбора разрешающего элемента

ИНФОРМАЦИЯ ОБ АВТОРАХ

Глеб Дмитриевич Степанов
автор для корреспонденции

orcid.org/0000-0003-3237-849X. E-mail: stpnv42@mail.ru
канд.физ.-мат.наук, доцент.

Для цитирования: G. D. Stepanov, “A Simple Algorithm for Finding a Non-negative Basic Solution of a System of Linear Algebraic Equations”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 234-237, 2021.

Сразу отметим, что рассматриваемый алгоритм получен в ходе усовершенствования симплекс-метода, который в течение многих десятков лет (см. [1—8]) является основным методом решения важной прикладной задачи – задачи линейного программирования (ЗЛП). Сама по себе очень естественная задача отыскания неотрицательного решения системы

$$\begin{gathered} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = a_{10}, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = a_{20}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = a_{m0}, \end{gathered} \tag{1}$$

где $m < n$, представляет особый интерес именно в связи с тем, что является одним из этапов симплекс-метода.

Задача линейного программирования заключается в поиске максимума или минимума линейной функции на множестве неотрицательных решений системы (1). С алгебраической точки зрения симплекс-метод можно рассматривать как обобщение метода полных исключений Гаусса. Так же как метод Гаусса симплекс-метод удобно формулировать в терминах преобразований расширенной матрицы системы (1), которые не меняют множества решений этой системы. Оба этих метода реализуются в виде последовательности шагов исключения, на каждом из которых в матрице системы выбирается ненулевой разрешающий элемент a_{pq}, а затем элементы расширенной матрицы преобразуются по формулам

$$\begin{aligned} a^{\text{н}}_{pj} &= a_{pj}/a_{pq} && (j \in \{0, 1, \ldots, n\}), \\ a^{\text{н}}_{ij} &= a_{ij} - a_{pj}a_{iq}/a_{pq} && (i \in \{1, \ldots, m\} \setminus p,\ \ j \in \{0, 1, \ldots, n\}), \end{aligned} \tag{2}$$

где через $a^{\text{н}}_{ij}$ обозначены элементы после шага исключения.

Симплекс-метод состоит из трех этапов, причем первый этап, по существу, совпадает с основной частью метода полного исключения Гаусса. На этом этапе расширенная матрица системы преобразуется к виду

$$\begin{pmatrix} a_{10} & a_{11} & \ldots & a_{1n-m} & 1 & 0 & \ldots & 0 \\ a_{20} & a_{21} & \ldots & a_{2n-m} & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m0} & a_{m1} & \ldots & a_{mn-m} & 0 & 0 & \ldots & 1 \end{pmatrix}, \tag{3}$$

или отличающемуся от (3) перестановкой столбцов с номерами большими нуля. Такие матрицы здесь будем называть G-матрицами. Перестановок столбцов, приводящих G-матрицу к виду (3), может быть несколько. При выделенной каким-либо образом одной из них через $s_1, s_2, \ldots, s_n$ обозначим соответствующую перестановку исходных номеров столбцов. Номера $r_1 = s_{n-m+1}, r_2 = s_{n-m+2}, \ldots, r_m = s_n$ и переменные с этими номерами называются базисными. Вектор $(x_1, x_2, \ldots, x_n)$, в котором все внебазисные переменные равны нулю, а базисные переменные $x_{r_i} = -a_{i0}$ $(i = 1, 2, \ldots m)$, является решением системы уравнений. Такие решения называются базисными.

Второй этап симплекс-метода заключается в преобразовании G-матрицы в аналогичную матрицу, с неотрицательными элементами нулевого столбца, т.е. в отыскании неотрицательного базисного решения. Этот этап в симплекс-методе является наиболее трудоемким.

На третьем этапе, называющемся основной процедурой, G-матрица с неотрицательным нулевым столбцом преобразуется в такую же матрицу так, чтобы соответствующее базисное решение удовлетворяло критериям оптимальности целевой функции.

Этап 2 симплекс-метода оказывается более трудоемким, чем этап 3, по той причине, что он традиционно сводится методом искусственного базиса к применению основной процедуры. Метод связан с добавлением в систему новых переменных и формированием искусственной целевой функции. Явным недостатком этого приема является увеличение размера задачи, что крайне нежелательно при больших m и n.

На самом деле алгоритм этапа 2 проще составить по аналогии с алгоритмом этапа 3, но с более простым правилом выбора разрешающего элемента a_{pq} и более простыми условиями окончания. Первый вариант такого правила выбора выглядит следующим образом.

Правило 1. *Для G-матрицы p надо выбирать так, что* $a_{p0} < 0$. *При уже выбранном p, надо выбирать q так, что* $a_{pq} < 0$.

Сам алгоритм представляет из себя итерационный цикл, на каждой из итераций которого в G-матрице по правилу 1 (или его уточнению) выбирается разрешающий элемент a_{pq}, а затем матрица преобразуется по формулам (2). Цикл завершается либо при отсутствии требуемого p, что означает неотрицательность нулевого столбца и соответствующего базисного решения, либо при отсутствии требуемого q, что свидетельствует об отсутствии у системы уравнений неотрицательных решений.

Любой из двух вариантов окончания цикла, означает решение задачи, стоящей перед этапом 2. Но при использовании правила 1 не исключено еще и зацикливание. Правило 1 определяет выбор разрешающего элемента неоднозначно и при использовании его приходится уточнять. Возможны более удачные уточнения и менее удачные. При компьютерных экспериментах с естественными вариантами уточнения правила задача решалась очень быстро, т.е. либо находилось неотрицательное решение системы уравнений, либо выявлялось его отсутствие. Зацикливание было выявлено лишь при очень вычурных уточнениях. По сути, можно сказать, что некоторые уточнения правила 1, приводят к довольно эффективным эвристическим алгоритмам, для которых пока не найдены примеры, приводящие к зацикливанию. Более того, удалось получить уточнение правила 1, которое гарантирует отсутствие зацикливания.

Правило 2. *Для G-матрицы в качестве номера разрешающей строки из всех p, при которых* $a_{p0} < 0$, *надо брать номер с наименьшим* r_p. *При уже выбранном p надо брать наименьшее q, при котором* $a_{pq} < 0$.

Теорема 1. *Для G-матриц итерационный цикл, тело которого состоит из выбора разрешающего элемента по правилу 2 и последующего шага исключения, конечен.*

Несмотря на простоту формулировки правила 2, доказательство теоремы весьма непросто и связано с обоснованием конечности этапов сразу четырех алгоритмов решения ЗЛП. Два из этих алгоритмов предназначены для упрощения симплекс-метода, а два – являются двойственными к ним.

О трудоемкости полученного алгоритма можно лишь сказать, что при вычислительных экспериментах количество итераций, затрачиваемое на решение задач, было сравнимо с n, но, наверное, возможны и "плохие" примеры, как пример из [9] для основной процедуры симплекс-метода.

References

[1] S. I. Gass, *Linear Programming: Methods and Applications*. McGraw-Hill: New York, 1958.

[2] D. B. Yudin and E. G. Holstein, *Linear Programming*. Moscow: Nauka, 1963.

[3] G. Dantzig, “Linear programming, its applications and generalizations”, *Princeton University Press*, 1963.

[4] T. Hu, *Integer programming and network flows*. Addison-Wesley, 1969.

[5] S. Ashmanov, *Linear programming*. Moscow: Nauka, 1981.

[6] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Prentice-Hall, 1982.

[7] A. Schrijver, *Theory of linear and integer programming*. Moscow: Mir, 1991.

[8] F. P. Vasiliev and A. Y. Ivanitsky, *Linear Programming*. MCCME, 2020.

[9] V. Klee and G. J. Minty, “How good is the simplex algorithm?”, *Inequalities*, vol. 3, no. 3, pp. 159–175, 1972.

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF COMPUTING

The Investigation of Nonlinear Polynomial Control Systems

S. N. Chukanov[1], I. S. Chukanov[2]

DOI: 10.18255/1818-1015-2021-3-238-249

[1]Sobolev Institute of Mathematics, SB RAS, Omsk branch, 13 Pevtsova str., Omsk 644043, Russia.
[2]Ural Federal University, 19 Mira st., Yekaterinburg 620002, Russia.

MSC2020: 97R40, 68T50
Research article
Full text in English

Received July 18, 2021
After revision August 28, 2021
Accepted September 1, 2021

The paper considers methods for estimating stability using Lyapunov functions, which are used for nonlinear polynomial control systems. The apparatus of the Gröbner basis method is used to assess the stability of a dynamical system. A description of the Gröbner basis method is given. To apply the method, the canonical relations of the nonlinear system are approximated by polynomials of the components of the state and control vectors. To calculate the Gröbner basis, the Buchberger algorithm is used, which is implemented in symbolic computation programs for solving systems of nonlinear polynomial equations. The use of the Gröbner basis for finding solutions of a nonlinear system of polynomial equations is considered, similar to the application of the Gauss method for solving a system of linear equations. The equilibrium states of a nonlinear polynomial system are determined as solutions of a nonlinear system of polynomial equations. An example of determining the equilibrium states of a nonlinear polynomial system using the Gröbner basis method is given. An example of finding the critical points of a nonlinear polynomial system using the Gröbner basis method and the Wolfram Mathematica application software is given. The Wolfram Mathematica program uses the function of determining the reduced Gröbner basis. The application of the Gröbner basis method for estimating the attraction domain of a nonlinear dynamic system with respect to the equilibrium point is considered. To determine the scalar potential, the vector field of the dynamic system is decomposed into gradient and vortex components. For the gradient component, the scalar potential and the Lyapunov function in polynomial form are determined by applying the homotopy operator. The use of Gröbner bases in the gradient method for finding the Lyapunov function of a nonlinear dynamical system is considered. The coordination of input-output signals of the system based on the construction of Gröbner bases is considered.

Keywords: nonlinear systems; polynomial systems; Lyapunov functions; Gröbner bases

INFORMATION ABOUT THE AUTHORS

Sergei Nikolaevich Chukanov
correspondence author
orcid.org/0000-0002-8106-9813. E-mail: ch_sn@mail.ru
Doctor of Sciences in Engineering sciences, Professor.

Ilya Stanislavovich Chukanov
orcid.org/0000-0001-9946-7484. E-mail: chukanov022@gmail.com
Student.

Funding: This work was supported by the Basic Research Program of the Siberian Branch of the Russian Academy of Sciences No. I.5.1., Project No. 0314-2019-0020.

For citation: S. N. Chukanov and I. S. Chukanov, "The Investigation of Nonlinear Polynomial Control Systems", *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 238-249, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF COMPUTING

Исследование нелинейных полиномиальных систем управления

С. Н. Чуканов[1], И. С. Чуканов[2]

DOI: 10.18255/1818-1015-2021-3-238-249

[1]Институт математики им. С. Л. Соболева СО РАН, Омский филиал, ул. Певцова, д. 13, г. Омск, 644043 Россия.
[2]Уральский федеральный университет, ул. Мира, д. 19, г. Екатеринбург, 620002 Россия.

УДК 004.02
Научная статья
Полный текст на английском языке

Получена 18 июля 2021 г.
После доработки 28 августа 2021 г.
Принята к публикации 1 сентября 2021 г.

В работе рассматриваются методы оценивания устойчивости с помощью функций Ляпунова, применяемые для нелинейных полиномиальных систем управления. Для оценивания устойчивости используется аппарат метода базисов Грёбнера. Приводится описание метода базисов Грёбнера. Для применения метода канонические соотношении нелинейной системы аппроксимируются полиномами компонент векторов состоянии и управления. Для вычисления базиса Грёбнера применяется алгоритм Бухбергера, который реализован в программах символьных вычислений для решения систем нелинейных полиномиальных уравнений. Рассматривается использование базиса Грёбнера при нахождения решений нелинейной системы полиномиальных уравнений аналогично применению метода Гаусса для решения системы линейных уравнений. Определяются равновесные состояния нелинейной полиномиальной системы как решения нелинейной системы полиномиальных уравнений. Приводится пример определения равновесных состояний нелинейной полиномиальной системы с использованием метода базисов Грёбнера. Приводится пример нахождения критических точек нелинейной полиномиальной системы с использованием метода базисов Грёбнера и прикладного программного обеспечения Wolfram Mathematica. При использовании прикладного программного обеспечения Wolfram Mathematica применяется функция определения редуцированного базиса Грёбнера. Рассматривается применение метода базиса Грёбнера для оценивания области притяжения нелинейной динамической системы относительно точки равновесия. Для определения скалярного потенциала векторное поле динамической системы декомпозируется на градиентную и вихревую компоненты. По градиентному компоненту скалярный потенциал и функция Ляпунова в полиномиальной форме определяются на основе применения оператора гомотопии. Рассмотрено использование базисов Грёбнера при градиентном методе нахождения функции Ляпунова нелинейной динамической системы. Рассмотрено согласование сигналов ввода-вывода системы на основе построения базисов Грёбнера.

Ключевые слова: нелинейные системы; полиномиальные системы; функции Ляпунова; базисы Грёбнера

ИНФОРМАЦИЯ ОБ АВТОРАХ

Сергей Николаевич Чуканов
автор для корреспонденции
orcid.org/0000-0002-8106-9813. E-mail: ch_sn@mail.ru
д-р техн. наук, профессор.

Илья Станиславович Чуканов
orcid.org/0000-0001-9946-7484. E-mail: chukanov022@gmail.com
студент.

Финансирование: Работа выполнена при поддержке программы фундаментальных научных исследований СО РАН № I.5.1., проект № 0314-2019-0020.

Для цитирования: S. N. Chukanov and I. S. Chukanov, “The Investigation of Nonlinear Polynomial Control Systems”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 238-249, 2021.

Introduction

The most of the dynamical systems in technology and nature are nonlinear dynamical systems. The canonical relations of a nonlinear system can be approximated by polynomials of the components of the state and control vectors. Stability testing using the method of Lyapunov functions is widely applied to nonlinear systems.

There are several methods in the literature to identify candidates for Lyapunov functions [1]:

- decomposition of the sum of squares [2];
- using the Gröbner basis to select parameters [3];
- use of homotopy operators for decomposition of the vector field of states of the system [4, 5];
- the assumption that the derivative of the Lyapunov function is negative definite, and then obtain by integration and check the positive definiteness (gradient method) [6].

Gröbner bases are used to solve problems in the theory of nonlinear systems. Some of the applications of the Gröbner basis can be named: estimation of equilibrium states of a nonlinear system; finding the critical points of a given nonlinear system with the Lyapunov function; coordination of input-output signals of the system.

Gröbner bases facilitate the solution of a system of multidimensional polynomial equations in the same way as the Gaussian elimination algorithm makes it possible to solve a system of linear algebraic equations. In lexical ordering the Gröbner basis has a triangular structure, reminiscent of the triangular structure in the Gaussian elimination method.

The theory of control of dynamic objects can be divided into two subgroups [7]:

(1) systems in which the principle of superposition operates, and linear control methods can be used;

(2) systems in which the superposition principle does not work, and it is necessary to use nonlinear control methods. To improve the quality of the dynamic object control system, it is necessary to take into account the nonlinear features of the system.

1. Gröbner bases

The objects in the theory of Gröbner bases are polynomial ideals and algebraic varieties [8]. Let $p_1, ..., p_s$ be multidimensional polynomials in variables $x_1, ..., x_n$, whose coefficients lie in the field k (we will consider the field of real numbers $\mathbb{R}$). The variables $x_1, ..., x_n$ are considered "place markers" in the polynomials: $p_1, \dots, p_s \in \mathbb{R}[x_1, \dots, x_n]$. Algebraic variety defined by the polynomials $p_1, \dots, p_s$ is the collection of all solutions in $\mathbb{R}^n$ of the system of equations:

$$\begin{aligned} &p_1(x_1, ..., x_n) = 0, \\ &\dots \\ &p_s(x_1, ..., x_n) = 0. \end{aligned} \tag{1}$$

Formally:

$$V(p_1, ..., p_s) := \{(a_1, ..., a_n) \in \mathbb{R}^n : p_i(x_1, ..., x_n) = 0, i = 1, ..., s\}. \tag{2}$$

The polynomial ideal I, which is generated by $p_1, ..., p_s$, is a set of polynomials obtained by combining these polynomials by multiplying and adding with other polynomials:

$$I = \langle p_1, ..., p_s \rangle := \left\{ f = \sum_{i=1}^{s} g_i p_i : g_i \in \mathbb{R}[x_1, ..., x_n] \right\}. \tag{3}$$

The polynomials $p_i, i = 1, \dots, s$ form the basis of the ideal I. A useful interpretation of the polynomial ideal I is in terms of the equations (3). Multiplying g_i by arbitrary polynomials $g_i \in \mathbb{R}[x_1, ..., x_n]$ and adding them, we get the consequence from (1):

$$f = g_1 p_1 + \ldots + g_s p_s = 0,$$

and $f \in I$. Therefore, $I = \langle p_1, \ldots, p_s \rangle$ the ideal contains all "polynomial consequences" of the equations (3).

The Gröbner basis method is based on the concept of monomial ordering (a monomial is a polynomial consisting of one term), since it introduces a corresponding extension of the concept of a leading term and a leading coefficient, familiar for one-dimensional polynomials, to multidimensional polynomials. Let's consider lexicographic or lex order [8]. Let α, β be two n - tuples of integers $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$. n -tuple α follows β (in lex order), which is denoted as $\alpha \succ \beta$, if in the difference of vectors $\alpha - \beta = (\alpha_1 - \beta_1, \ldots, \alpha_n - \beta_n)$ the leftmost nonzero element is positive. For the polynomial $f = x_1^3 x_2 x_3^3 + 2x_1^3 x_3^4$ using lex order $x_1 \succ x_2 \succ x_3$ results in $x_1^3 x_2 x_3^3$ follows $x_1^3 x_3^4$, since the multidegrees of monomials satisfy: $(3, 1, 3) \succ (3, 0, 4)$. In this order, the leading coefficient and the leading term are respectively $LC(f) = 1$ and $LT(f) = x_1^3 x_2 x_3^3$. When using lex of order $x_3 \succ x_2 \succ x_1$ senior term: $LT(f) = 2x_1^3 x_3^4$, since $(4, 0, 3) \succ (3, 1, 3)$.

The ideal I has no unique basis, but for any two different bases $\langle p_1, ..., p_s \rangle$ and $\langle g_1, ..., g_m \rangle$ of the ideal I, the varieties $V(p_1, \ldots, p_s)$ and $V(g_1, \ldots, g_m)$ are equal; the variety depends only on the ideal generated by its defining equations. If all polynomials in a given basis of an ideal have a degree lower than the degree of any other polynomial in an ideal, then this basis is the simplest. For an ideal I and a given monomial order, we denote the set of leading terms of elements I as $LT(I)$. The ideal generated by elements from $LT(I)$ is denoted by $\langle LT(I) \rangle$. The Gröbner basis is formally defined as a set of polynomials $g_1, \ldots, g_m$, for which $\langle LT(I) \rangle = \langle LT(g_1), \ldots, LT(g_m) \rangle$. When calculating Gröbner bases, a monomial order is specified. We note two properties of Gröbner bases for a given monomial order:

1. Each ideal $I \subset \mathbb{R}[x_1, ..., x_n]$, different from the trivial $\langle 0 \rangle$, has a Gröbner basis.

2. For the ideal $I \subset \mathbb{R}[x_1, ..., x_n]$, different from the trivial $\langle 0 \rangle$, the Gröbner basis of the ideal I can be calculated using a finite number of algebraic operations.

For a given set of polynomials P, there is an algorithm that computes the Gröbner basis for the (ideal generated by) P in a finite number of steps [9]. Buchberger's algorithm generalizes algorithms: Gaussian elimination for a system of linear algebraic equations and Euclid's algorithm for calculating the greatest common divisor of a set of one-dimensional polynomials. This algorithm was implemented on computers in symbolic computation programs using Gröbner bases for solving systems of polynomial equations [10–12].

2. Finding equilibrium states of a nonlinear dynamical system

The use of the Gröbner basis in finding solutions to a nonlinear system of polynomial equations is similar to the application of the Gauss method for solving a quadratic system of linear equations. Consider an example of reducing a nonlinear system of polynomial equations: $p_1 = x_1 - x_2^2 = 0$, $p_2 = x_2 + x_3^2 = 0$, $p_3 = x_3 - 2x_1^2 = 0$, to a triangular form using the Gröbner basis method for lex order: $x_1 \succ x_2 \succ x_3$. In the WOLFRAM MATHEMATICA package, the function call

$GroebnerBasis[\{p1, p2, p3\}, \{x1, x2, x3\}, \{\}]$

leads to a triangular Gaussian form of polynomial equations:

$$\begin{aligned} x_1 - x_3^4 &= 0, \\ x_2 + x_3^2 &= 0, \\ -x_3 + 2x_3^8 &= 0, \end{aligned}$$

which allows us to get a solution to this system.

Consider a nonlinear system without inputs $\dot{x}(t) = f(x(t))$; $x, f \in \mathbb{R}^n, t \in \mathbb{R}$, where $f(x) = 0$ is a vector of polynomials in x. The equilibrium states for this polynomial system are obtained as solutions of a nonlinear system of polynomial equations: $f(x) = 0$.

Example 1

Equilibrium states of the [8] polynomial system:

$$\begin{aligned}\dot{x}_1 &= x_1 + x_2 - x_3^2,\\ \dot{x}_2 &= x_1^2 + x_2 - x_3,\\ \dot{x}_3 &= -x_1 + x_2^2 + x_3,\end{aligned}$$

can be obtained as solutions of system polynomial equations:

$$\begin{aligned}p_1 &:= x_1 + x_2 - x_3^2 = 0,\\ p_2 &:= x_1^2 + x_2 - x_3 = 0,\\ p_3 &:= -x_1 + x_2^2 + x_3 = 0.\end{aligned}$$

The Gröbner basis for the ideal (p_1, p_2, p_3) using lex order: $x_1 \succ x_2 \succ x_3$, has the form:

$$\begin{aligned}g_1 &:= 4x_1 - 2x_1^2 - 4x_1^3 + x_1^4 + x_1^6,\\ g_2 &:= -x_1^2 + x_1^4 - 2x_2 + 2x_1^2 x_2,\\ g_3 &:= -x_1 + x_1^2 + x_2 + x_2^2,\\ g_4 &:= -x_1^2 - x_2 + x_3.\end{aligned}$$

Algebraic equations $g_i = 0, i = 1, 2, 3, 4$ has the same solutions as $p_j = 0, j = 1, 2, 3$. The polynomial g_4 depends only on x_3; from the algebraic equation $g_4(x_3) = 0$, you can determine x_3. If the numerical value of x_3 substitute in $g_3(x_2, x_3) = 0$, then you can define x_2; from $g_2(x_1, x_2, x_3) = 0$ you can define x_1.

Table 1. English

	x_1	x_2	x_3
$Solution\ 1$:	$x_1 = -1,$	$x_2 = 0.5 - 1.32i,$	$x_3 = 0.5 - 1.32i,$
$Solution\ 2$:	$x_1 = 0,$	$x_2 = 0,$	$x_3 = 0,$
$Solution\ 3$:	$x_1 = 1,$	$x_2 = 0,$	$x_3 = 1,$
$Solution\ 4$:	$x_1 = 1.18,$	$x_2 = -0.69,$	$x_3 = 0.70,$
$Solution\ 5$:	$x_1 = -0.59 - 1.74i,$	$x_2 = -2.35 + 1.03i,$	$x_3 = -5.04 + 3.09i,$
$Solution\ 6$:	$x_1 = -0.59 + 1.74i,$	$x_2 = 1.35 - 1.03i,$	$x_3 = -1.35 - 3.09i,$
$Solution\ 7$:	$x_1 = -1,$	$x_3 = -0.5 + 1.32i,$	$x_3 = 0.5 + 1.32i,$
$Solution\ 8$:	$x_1 = 0,$	$x_2 = -1,$	$x_3 = -1,$
$Solution\ 9$:	$x_1 = 1,$	$x_2 = -1,$	$x_3 = 0,$
$Solution\ 10$:	$x_1 = 1.18,$	$x_2 = -0.31,$	$x_3 = 1.08,$
$Solution\ 11$:	$x_1 = -0.59 - 1.74i,$	$x_2 = 1.35 - 1.03i,$	$x_3 = -1.35 + 1.03i,$
$Solution\ 12$:	$x_1 = -0.59 + 1.74i,$	$x_2 = -2.35 + 1.03i,$	$x_3 = -5.04 - 1.03i.$

In the WOLFRAM MATHEMATICA package: Form an ideal of polynomials:
$p1 = x1 + x2 - x3^2$; $p2 = x1^2 + x2 - x3$; $p3 = -x1 + x2^2 + x3$.

Let us define the Gröbner basis:

$grbas = GroebnerBasis[\{p1, p2, p3\}, \{x3, x2, x1\}, \{\}]$,

$grbas = \{4x1 - 2x1^2 - 4x1^3 + x1^4 + x1^6, -x1^2 + x1^4 - 2x2 + 2x1^2 x2,$
$-x1 + x1^2 + x2 + x2^2, -x1^2 - x2 + x3\}$.

To find the roots of x_1, we define the reduced Gröbner basis:

$grbas = GroebnerBasis[\{p1, p2, p3\}, \{x3, x2, x1\}, \{x3, x2\}]$,

$grbas = \{4x1 - 2x1^2 - 4x1^3 + x1^4 + x1^6\}$.

Let's execute: $Roots[4x1 - 2x1^2 - 4x1^3 + x1^4 + x1^6 == 0, x1]$.

To find the roots of x_2 with known x_1, we define the reduced Gröbner basis:

$grbas = GroebnerBasis\left[\{p1, p2, p3\}, \{x3, x2, x1\}, \{x3\}\right]$,
$grbas = \left\{-x1 + x1^2 + x2 + x2^2\right\}$.

Let's execute: $Roots[-x1 + x1^2 + x2 + x2^2 == 0, x2]$.

To find the roots of x_3 with known x_1, x_2, execute:
$Roots[-x1^2 - x2 + x3 == 0, x3]$. The results are shown in Table 1.

□

3. Application of the Gröbner basis method in the theory of the method of Lyapunov functions

3.1. Estimation of the area of attraction

The set of all initial conditions of a dynamical system, which converge to the same equilibrium state, is called the area of attraction of this equilibrium state [3, 13]. One way to get an estimate of the domain of attraction is to use the Lyapunov functions.

The standard result of Lyapunov's theory is that if $x = 0$ is an equilibrium point for a system with continuous time: $\dot{x} = f(x), x \in D \subset \mathbb{R}^n$, is a domain containing $x = 0$ and $V : D \longrightarrow \mathbb{R}$ is a continuously differentiable Lyapunov function such that $V(0) = 0$ and $V(x) > 0, \dot{V} = V_x f(x) < 0, \forall x \in D - \{0\}$; then the point $x = 0$ is asymptotically stable. For such a Lyapunov function, consider the sets $\Omega = \{x \in \mathbb{R}^n : V_x f(x) < 0\}$ and $B_d = \{x \in \mathbb{R}^n : V(x) \le d\}$. If there is a value $d > 0$ such that $B_d \subset \Omega$, then the set B_d is an estimate of the domain of attraction.

For polynomial systems with a polynomial Lyapunov function V, the Gröbner basis can be used to determine B_d. You can determine the largest B_d by finding a d such that $B_d \subset \Omega$. For polynomial systems with polynomial Lyapunov functions, $V(x) - d$ and $V_x f(x)$ are polynomials and the boundaries of the sets B_d and Ω are varieties $Z(V - d)$ and $Z(V_x f(x))$, respectively. At the points of contact $Z(V - d)$ and $Z(V_x f(x))$, the gradients V and $V_x f(x)$ are parallel [14]. Using this information, we obtain a system of $n + 2$ polynomial equations in $n + 2$ variables $(x_1, \dots, x_n, d, \lambda)$, where λ is the Lagrange multiplier (see Appendix):

$$\begin{aligned} &V - d = 0, \\ &V_x f = 0, \\ &\nabla(V_x f) - \lambda \nabla V = 0. \end{aligned} \tag{4}$$

In the case of the vector of Lagrange multipliers $\lambda = (\lambda_1, \dots, \lambda_m)^T \in \mathbb{R}^m$ we obtain a system of $n + m + 1$ equations from $n + m + 1$ variables $x_1, ..., x_n, d, \lambda_1, \dots, \lambda_m$.

Calculating the Gröbner basis for this system, where the variable d has the lowest rank in the lex order, we obtain a polynomial equation for d. The smallest positive solution of this equation (the value $d_{\min} > 0$), is the best estimate of the area of attraction.

Example 2

Consider a second-order system:

$$\dot{x} = f(x) = \begin{pmatrix} -x_1 \\ -x_2 + 2x_1 x_2^2 \end{pmatrix}$$

and choose the Lyapunov function $V(x) = x^T \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix} x = 4x_1^2 + 4x_1 x_2 + 3x_2^2$, therefore: $V_x = \begin{pmatrix} 8x_1 + 4x_2 \\ 4x_1 + 6x_2 \end{pmatrix}$;
$\dot{V} = V_x f = -8x_1^2 + 12x_1 x_2^3 - 8x_1 x_2 + 8x_1^2 x_2^2 - 6x_2^2$.

The fact that the gradients are parallel $(\nabla(V_x f) - \lambda \cdot \nabla V = 0)$ gives additional equations:

$$\begin{cases} g_1 = 8x_1 + 4x_2 - \lambda(-16x_1 + 12x_2^3 - 8x_2 + 16x_1 x_2^2), \\ g_2 = 4x_1 + 6x_2 - \lambda(36x_1 x_2^2 - 8x_1 + 16x_1^2 x_2 - 12x_2). \end{cases}$$

Let us calculate the Gröbner basis for four polynomials $\{V - d, V_x f, g_1, g_2\}$ in ordering: $d \prec x_1 \prec \lambda \prec x_2$.

This reduces the system to a polynomial: $4d^4 - 147d^3 + 768d^2 + 2048d$, which results in the values of the roots: $\{\ 0 \quad 29.71 \quad -1.92 \quad 8.97\ \}$. The smallest nonzero positive value of d for which there is a solution to the system is $d \approx 8.97$.

In the WOLFRAM MATHEMATICA package: $Vd = 4x1^2 + 4x1x2 + 3x2^2 - d$;

$Vxf = -8x1^2 - 8x1x2 - 6x2^2 + 8x1^2x2^2 + 12x1x2^3$;

$g1 = 8x1 + 4x2 - lam(-16x1 - 8x2 + 16x1x2^2 + 12x2^3)$;

$g2 = 4x1 + 6x2 - lam(-8x1 - 12x2 + 16x1^2x2 + 36x1x2^2)$.

$grbas = GroebnerBasis[\{Vd, Vxf, g1, g2\}, \{d, x1, lam, x2\}, \{x1, lam, x2\}]$,

$grbas = \left\{2048d + 768d^2 - 147d^3 + 4d^4\right\}$,

$Roots[2048d + 768d^2 - 147d^3 + 4d^4 == 0, d] \Longrightarrow d = \{\ 0 \quad 29.71 \quad -1.92 \quad 8.97\ \}$. □

3.2. Decomposition of the dynamical system vector field $\dot{x} = f(x)$

For a dynamic system: $\dot{x} = f(x); x \in \mathbb{R}^n, f(x) \in \mathbb{R}^n, f(0) = 0$, form a vector field $X = f(x)\dfrac{\partial}{\partial x}$. Form the corresponding differential form $\omega = f(x)dx$ in the dual basis $\left\langle dx_i, \dfrac{\partial}{\partial x_j} \right\rangle = \delta_{ij}$. Let us construct a scalar potential from the vector field X by using the homotopy operator centered at the point $x_0 = 0$ for the form $\omega = f(x)dx$:

$$\mathbb{H}(\omega) = \int_0^1 \left(x\frac{\partial}{\partial x} \right) \rfloor (f(\lambda x)dx)\, d\lambda = \int_0^1 x^T f(\lambda x) d\lambda.$$

We will assume that $\varphi(x) = \mathbb{H}\omega(x)$ is a scalar potential.

Example 3

Consider an example of dynamic equations:

$$\begin{aligned} \dot{x}_1 &= -x_1 + x_2^2, \\ \dot{x}_2 &= -x_2 - x_1^2. \end{aligned}$$

Let's construct a dual differential form:

$$\omega = (-x_1 + x_2^2)dx_1 + (-x_2 - x_1^2)dx_2,$$

to which we apply the homotopy operator with $x_0 = 0$:

$$\begin{aligned} \varphi(x) = \mathbb{H}(\omega(x)) &= \int_0^1 x^T f\left(\tilde{\lambda}x\right) d\tilde{\lambda} = \int_0^1 \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} -\tilde{\lambda}x_1 + \tilde{\lambda}^2 x_2^2 \\ -\tilde{\lambda}x_2 - \tilde{\lambda}^2 x_1^2 \end{pmatrix} d\tilde{\lambda} = \\ &= -\frac{1}{2}(x_1^2 + x_2^2) + \frac{1}{3}(x_1 x_2^2 - x_2 x_1^2). \end{aligned}$$

Let us choose the function as the scalar Lyapunov function:

$$\begin{aligned} V(x) &= -6 \cdot \phi(x) = 3\left(x_1^2 + x_2^2\right) + 2\left(x_1 x_2^2 - x_2 x_1^2\right), \\ \dot{V} = V_x f &= \begin{pmatrix} 6x_1 + 2x_2^2 - 4x_1 x_2 & 6x_2 - 2x_1^2 + 4x_1 x_2 \end{pmatrix} \begin{pmatrix} -x_1 + x_2^2 \\ -x_2 - x_1^2 \end{pmatrix} = \\ &= -6x_1^2 - 6x_2^2 - 6x_2^2 x_1 + 6x_1^2 x_2. \end{aligned}$$

Let's find the solution of the system $V\_d = V(x) - d = 0$ in the WOLFRAM MATHEMATICA package:

$$V\_d = 3x_1^2 + 3x_2^2 + 2x_2^2x_1 - 2x_1^2x_2 - d,$$
$$V_xf = -6x_1^2 - 6x_2^2 - 6x_2^2x_1 + 6x_1^2x_2,$$
$$g_1 = 6x_1 + 2x_2^2 - 4x_1x_2 + lam(-12x_1 - 6x_2^2 + 12x_1x_2),$$
$$g_2 = 6x_2 + 4x_2x_1 - 2x_1^2 + lam(-12x_2 - 12x_2x_1 + 6x_1^2),$$
$$grb = GroebnerBasis[\{V\_d, V_xf, g_1, g_2\}, \{d, x_1, lam, x_2\}, \{x_1, lam, x_2\}],$$
$$grb = \left\{54d - 29d^2 + d^3\right\}.$$

The roots of the polynomial $54d - 29d^2 + d^3 are : d_1 = 0, d_2 = 2, d_3 = 27$.

The smallest nonzero positive value d, for which there is a solution to the system: $d_{\min} = 2$.

3.3. Gradient method for finding the Lyapunov function [6]

The method is to first assume that $\dot{V}$ is negative definite and then get V by integration. If V is positive definite, then the stability of the system can be determined.

Consider a nonlinear system:

$$\dot{x} = f(x); x, f \in \mathbb{R}^n, f(0) = 0. \tag{5}$$

The assumed Lyapunov function $V = V(x)$; that is: $\dot{V} = V_x \cdot f(x)$. Let $V_x^T = W$, where the vector W satisfies the condition:

$$\frac{\partial W_i}{\partial x_j} = \frac{\partial W_j}{\partial x_i}(i, j = 1, ..., n). \tag{6}$$

Then the function V can be found as an integral: $V = \int_0^x V_x dx = \int_0^x W^T dx$; since this integral does not depend on the integration path, then:

$$V = \int_0^{x_1} W_1(\tau_1, 0, \dots, 0)d\tau_1 + \cdots + \int_0^{x_n} W_n(x_1, \dots, x_{n-1}, x_n)d\tau_n. \tag{7}$$

It is necessary to choose such a function $\dot{V}$, so that the V obtained from (7) is positive definite; then the equilibrium state $x = 0$ of the system (5) is stable.

Consider the variable gradient method. Using this method, you can find the Lyapunov function, assuming that $W = V_x^T = Bx$, where $B = (b_{ij})$ must be determined; b_{ij} can be a function of x, and $b_{ij}(x) = b_{ji}(x)$. The choice of $B = (b_{ij})$ must ensure the fulfillment of the condition (6) and $x^TB^Tf(x)$ must be positive definite.

If $f(x) = A(x)x$, then $\dot{V}$ takes the form $\dot{V} = x^TH(x)x$, where $H(x) = B^TA(x)$. For the matrix $H(x)$, the following condition is selected:

$$\begin{cases} (h_{ii} < 0) \vee (h_{ii} \le 0), \\ h_{ij} + h_{ji} = 0; i = 1, \dots, n, \end{cases} \tag{8}$$

to ensure that $\dot{V}$ is negative definite. The function $V(x)$ can be defined by the integral (7) and check whether it is positive definite.

Example 4

Consider the system:

$$\dot{x}_1 = -x_1,$$
$$\dot{x}_2 = -x_2 + x_1^3,$$

that is $A(x) = \begin{bmatrix} -1 & 0 \\ x_1^2 & -1 \end{bmatrix}$. Suppose $W = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$.

Then from $H(x) = B^T \cdot A(x) = \begin{bmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ x_1^2 & -1 \end{bmatrix}$ and (8) follows:

$h_{11} = -b_{11} + b_{21}x_1^2 < 0, h_{22} = -b_{22} < 0, h_{12} + h_{21} = -b_{21} - b_{12} + b_{22}x_1^2 = 0.$
$b_{22} > 0, b_{11} > 0, b_{21}x_1^2 < b_{11}, b_{21} = -b_{12} + b_{22}x_1^2.$
Choosing $b_{11} = b_{22} = 1$, $b_{12} = 5$, we get $b_{21} = -5 + x_1^2$ and:

$$W = \begin{bmatrix} 1 & 5 \\ -5 + x_1^2 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + 5x_2 \\ -5x_1 + x_1^3 + x_2 \end{pmatrix}.$$

The function $\dot{V}$:

$$\dot{V} = W^T f(x) = \begin{pmatrix} x_1 + 5x_2 \\ -5x_1 + x_1^3 + x_2 \end{pmatrix}^T \begin{pmatrix} -x_1 \\ -x_2 + x_1^3 \end{pmatrix} = -x_1^2 - 5x_1^4 + x_1^6 - x_2^2.$$

Performing the integration, we obtain:

$$V = \int_0^{x_1} (\tau_1 + 5 * 0) d\tau_1 + \int_0^{x_2} (-5x_1 + x_1^3 + \tau_2) d\tau_2 =$$
$$= \frac{1}{2}x_1^2 + (-5x_1 + x_1^3)x_2 + \frac{1}{2}x_2^2.$$

In the WOLFRAM MATHEMATICA package:
$V\_d = x_1^2 + 2(-5x_1 + x_1^3)x_2 + x_2^2 - d,$
$Vxf = -x_1^2 - 5x_1^4 + x_1^6 - x_2^2,$
$g1 = 2x_1 + 2(-5 + 3x_1^2)x_2 + lam(-2x_1 - 20x_1^3 + 6x_1^5),$
$g2 = 2(-5x_1 + x_1^3) + 2x_2 + lam(-2x_2),$
$grb = GroebnerBasis[\{V\_d, Vxf, g1, g2\}, \{d, x1, lam, x2\}, \{x1, lam, x2\}].$
As a result, we get a polynomial:

$$\{235480000d + 2221000d^2 + 382824847d^3 - 112046495d^4 +$$
$$+5932816d^5 + 117990d^6 + 6561d^7\},$$

whose roots are: $d_1 = 0.0$, $d_2 = 4.99$, $d_3 = 9.28$, $d_{4,5} = -16.04 \pm 32.48i$, $d_{6.7} = -0.09 \pm 0.76i$. The smallest nonzero positive value d, for which there is a solution to the system: $d_{\min} = 4.99$.

4. Conversions of input-output signals of a nonlinear system

Consider a differential ring - a ring on which the differentiation operation is defined. It is assumed that differentiation is carried out with respect to the implicit variable t. A differential ideal is an ideal that is closed under differentiation.

A polynomial system in the state space is a system of differential equations:

$$\dot{x}_1 = f_1(x, u), \dots, \dot{x}_n = f_n(x, u), y = h(x, u),$$

where $h, f_i \in \mathbb{R}[x, u], \forall i.$

Thus, every polynomial system in the form of a state space corresponds to a differential ideal in $\mathbb{R}[x, u, y]$:

$$I = [\varphi_1, \dots, \varphi_n, y - h],$$

where $\varphi_i = \dot{x}_i - f_i(x, u), i = 1, \dots, n.$

The problem of transformation from the state space to the input-output form: let I be a differential ideal; find a generator for the differential ideal $I \cap \mathbb{R}[u, y]$.

Example 5

Suppose that it is necessary to find a differential relationship between u and y from the description in the state space of the system:

$$\dot{x}_1 = -2x_1 + x_2^2; \dot{x}_2 = -x_1x_2 + u; y = x_2.$$

Differentiating the equations of the system with respect to t and replacing $\dot{x}_i$ by f_i, we get:
$g_1 = y - x_2; g_2 = \dot{y} - (u - x_1x_2); g_3 = \ddot{y} - (\dot{u} - (x_2^2 - 2x_1)x_2 - x_1(u - x_1x_2))$.
Replace $y \longrightarrow y_0, \dot{y} \longrightarrow y_1, \ddot{y} \longrightarrow y_2, u \longrightarrow u_0, \dot{u} \longrightarrow u_1$ in g_i, calculate the Gröbner basis G for:
$(y_0 - x_1, y_1 - u_0 + x_1x_2, y_2 - u_1 + (x_2^2 - 2x_1)x_2 + x_1(u_0 - x_1x_2))$ relative to lex order:
$u_0 \prec u_1 \prec y_0 \prec y_1 \prec y_2 \prec x_1 \prec x_2$.
Therefore, the input signals $u, \dot{u}$ and the output signals $y, \dot{y}, \ddot{y}$ are related by:

$$(-2u - \dot{u} + 2\dot{y} + \ddot{y} + y\dot{y})y^3 + (-3u_0^2 + 3u\dot{y} - \dot{y}^2)\dot{y} + u_0^3.$$

In the WOLFRAM MATHEMATICA package:
$g1 = y0 - x1,$
$g2 = y1 - u0 + x1 * x2,$
$g3 = y2 - u1 + (x2^2 - 2 * x1) * x2 + x1 * (u0 - x1 * x2).$
$grbas = GroebnerBasis[\{g1, g2, g3\}, \{u0, u1, y0, y1, y2, x1, x2\}, \{x1, x2\}],$
$grbas = (-2u_0 - u_1 + 2y_1 + y_2 + y_0y_1)y_0^3 + (-3u_0^2 + 3u_0y_1 - y_1^2)y_1 + u_0^3.$
□

Example 6

Let's consider a method for finding the observability of the components of the state vector of a system based on the construction of a reduced Gröbner basis. Let us choose the system from Example 5. Suppose that it is necessary to find the influence of the variation of the component of the state vector x_1 on the output signal y. Differentiating the equations of the system with respect to t and replacing $\dot{x}_i$ by f_i, we obtain the polynomials g_1, g_2, g_3 similar to the polynomials of Example 5 in the absence of $u, \dot{u}$. Replace $y \longrightarrow y_0, \dot{y} \longrightarrow y_1, \ddot{y} \longrightarrow y_2$ in g_i, calculate the Gröbner basis G for: $(y_0 - x_1, y_1 + x_1x_2, y_2(x_2^2 - 2x_1)x_2 - x_1^2x_2)$ relative to lex order: $y_0 \prec y_1 \prec y_2 \prec x_1 \prec x_2$.

In the WOLFRAM MATHEMATICA package:
$g_1 = y_0 - x_2,$
$g_2 = y_1 + x_1x_2,$
$g_3 = y_2 + \left(x_2^2 - 2x_1\right) x_2 - x_1^2x_2.$
Reduced Gröbner basis with the exception of x_2:
$grbas = GroebnerBasis[\{g_1, g_2, g_3\}, \{x_1, x_2, y_0, y_1, y_2\}, \{x_2\}].$
One of the polynomials of the resulting reduced Gröbner basis:
$x_1y_0 + y_1 = 0 \Longrightarrow x_1y + \dot{y} = 0,$
where you can find the expression for the variation δx_1:
$\delta x_1 y + \dot{y} = 0 \Longrightarrow \delta x_1 = y^{-1}\dot{y}, \text{if } y \neq 0.$
□

Conclusion

The paper considers methods for estimating stability using Lyapunov functions, applied to nonlinear systems. The canonical relations of a nonlinear system are approximated by polynomials of the components of the state and control vectors. To assess the stability, Gröbner bases are used. A method for finding the critical points of a given nonlinear system is proposed. The coordination of input-output signals of the system based on the construction of Gröbner bases is considered.

Appendix [14, ch.11]

We introduce vector-valued functions $h = (h_1, ..., h_m)$ and write general nonlinear programming problems as minimizing $f(x)$ for $h(x) = 0, x \in \Omega$. Restrictions $h(x) = 0$ are called functional constraints. A point $x \in \Omega$ satisfying all functional constraints is called admissible. Introduce the subspace $M = \{y : \nabla h(x^*)y = 0\}$ and investigate under what conditions M is a tangent plane at the point x^*.

A point x^*, satisfying the constraints $h(x^*) = 0$, is called a regular constraint point if the gradient vectors $\nabla h_1(x^*), \dots, \nabla h_m(x^*)$ are linearly independent. If the functions h are affine ($h(x) = A \cdot x + b$), then the regularity is equivalent to the condition $\operatorname{rank}(A) = m$ regardless of x. At a regular point x^* of the surface S, defined by the expression $h(x) = 0$, the tangent plane is $M = \{y : \nabla h(x^*)y = 0\}$.

Let x^* be a regular point of constraints $h(x) = 0$ and a point of local extremum of the function, taking into account these constraints. Then for $y \in \mathbb{R}^n$, satisfying $\nabla h(x^*)y = 0$, must hold: $\nabla f(x^*)y = 0$. This means that $\nabla f(x^*)$ is a linear combination of gradients ∇h in x^*; relations lead to the need to introduce the vector of Lagrange multipliers λ.

Let x^* be a local extremum point of the function f subject to the constraints $h(x) = 0$. Then there exists a vector of Lagrange multipliers $\lambda \in \mathbb{R}^m$, such that:

$$\nabla f(x^*) + \lambda^T \nabla h(x^*) = 0.$$

First order necessary conditions $\nabla f(x^*) + \lambda^T \nabla h(x^*) = 0$ together with the constraints $h(x^*) = 0$ give $n + m$ equations with $n + m$ variables x^*, λ. Introduce the Lagrangian: $l(x, \lambda) = f(x) + \lambda^T h(x)$. Then the necessary conditions can be expressed in the form:

$$\nabla_x l(x, \lambda) = 0, \nabla_\lambda l(x, \lambda) = 0.$$

References

[1] N. N. Krasovsky, *Problems of the Theory of Stability of Motion*. Moscow: Mir, 1959.

[2] A. Papachristodoulou and S. Prajna, "On the construction of Lyapunov functions using the sum of squares decomposition", in *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, IEEE, vol. 3, 2002, pp. 3482–3487.

[3] K. Forsman, "Construction of Lyapunov functions using Gröbner bases", in *[1991] Proceedings of the 30th IEEE Conference on Decision and Control*, IEEE, 1991, pp. 798–799.

[4] S. N. Chukanov and D. V. Ulyanov, "Decomposition of the vector field of control system by constructing a homotopy operator", *Probl. Upr.*, vol. 6, pp. 2–6, 2012.

[5] D. G. Edelen, *Applied Exterior Calculus*. John Wiley & Sons, Inc., 1985.

[6] X. Liao, L. Wang, and P. Yu, *Stability of dynamical systems*. Elsevier, 2007.

[7] H. K. Khalil, *Nonlinear systems*. Prentice hall, 2002.

[8] D. Nesic, I. Mareels, T. Glad, and M. Jirstrand, "The Gröbner basis method in control design: An overview", in *IFAC Proceedings Volumes*, vol. 35, 2002, pp. 13–18.

[9] B. Buchberger, "Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems", *Aequationes Muthematicae*, vol. 4, no. 3, pp. 374–383, 1970.

[10] J. L. Awange, B. Paláncz, R. H. Lewis, and L. Völgyesi, *Mathematical Geosciences – Hybrid Symbolic-Numeric Methods*. Springer International Publishers, 2018.

[11] S. Wolfram, *The Mathematica Book*. Wolfram Media, 2003.

[12] M. Demenkov, "A Matlab tool for regions of attraction estimation via numerical algebraic geometry", in *2015 International Conference on Mechanics-Seventh Polyakhov's Reading*, IEEE, 2015, pp. 1–5.

[13] N. Sidorov, D. Sidorov, and Y. Li, “Basins of attraction and stability of nonlinear systems equilibrium points”, *Differential Equations and Dynamical Systems*, pp. 1–10, 2019.

[14] D. G. Luenberger and Y. Ye, *Linear and nonlinear programming*. Springer, 2016.

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF DATA

Comparison of Style Features for the Authorship Verification of Literary Texts

K. V. Lagutina[1]

DOI: 10.18255/1818-1015-2021-3-250-259

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50
Research article
Full text in English

Received May 4, 2021
After revision August 20, 2021
Accepted August 25, 2021

The article compares character-level, word-level, and rhythm features for the authorship verification of literary texts of the 19th-21st centuries. Text corpora contains fragments of novels, each fragment has a size of about 50 000 characters. There are 40 fragments for each author. 20 authors who wrote in English, Russian, French, and 8 Spanish-language authors are considered.

The authors of this paper use existing algorithms for calculation of low-level features, popular in the computer linguistics, and rhythm features, common for the literary texts. Low-level features include n-grams of words, frequencies of letters and punctuation marks, average word and sentence lengths, etc. Rhythm features are based on lexico-grammatical figures: anaphora, epiphora, symploce, aposiopesis, epanalepsis, anadiplosis, diacope, epizeuxis, chiasmus, polysyndeton, repetitive exclamatory and interrogative sentences. These features include the frequency of occurrence of particular rhythm figures per 100 sentences, the number of unique words in the aspects of rhythm, the percentage of nouns, adjectives, adverbs and verbs in the aspects of rhythm. Authorship verification is considered as a binary classification problem: whether the text belongs to a particular author or not. AdaBoost and a neural network with an LSTM layer are considered as classification algorithms. The experiments demonstrate the effectiveness of rhythm features in verification of particular authors, and superiority of feature types combinations over single feature types on average. The best value for precision, recall, and F-measure for the AdaBoost classifier exceeds 90% when all three types of features are combined.

Keywords: stylometry; natural language processing; style features; rhythm features; authorship verification

INFORMATION ABOUT THE AUTHORS

Ksenia Vladimirovna Lagutina
correspondence author

orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
postgraduate student.

Funding: The reported study was funded by RFBR, project number 20-37-90045.

For citation: K. V. Lagutina, “Comparison of Style Features for the Authorship Verification of Literary Texts”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 250-259, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Сравнение стилистических характеристик для верификации авторов художественных текстов

К. В. Лагутина[1]

DOI: 10.18255/1818-1015-2021-3-250-259

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на английском языке

Получена 4 мая 2021 г.
После доработки 20 августа 2021 г.
Принята к публикации 25 августа 2021 г.

В статье сравниваются характеристики уровней символов, слов и ритма для верификации авторства художественных текстов 19-21-го веков. Корпуса текстов содержат фрагменты романов, каждый фрагмент имеет размер около 50 000 знаков. Для каждого автора приводится 40 фрагментов. Рассматриваются по 20 авторов, писавших на английском, русском, французском языках, и 8 испаноязычных авторов.

Авторы статьи используют существующие алгоритмы для вычисления популярных в современной компьютерной лингвистике низкоуровневых характеристик и распространённых в художественной литературе ритмических характеристик. Низкоуровневые характеристики включают в себя n-граммы слов, частоты встречаемости букв и знаков пунктуации, среднюю длину слова и предложения и т. д. Ритмические характеристики основаны на лексико-грамматических средствах: анафоре, эпифоре, симплоке, апозиопезе, эпаналепсисе, анадиплозисе, диакопе, эпизевксисе, хиазме, многосоюзие, повторяющихся восклицательных и вопросительных предложениях. Данные характеристики включают в себя частоты появления отдельных ритмических средств на 100 предложений, количество уникальных слов в аспектах ритма, доли существительных, прилагательных, наречий и глаголов в аспектах ритма. Верификация авторов рассматривается как задача бинарной классификации: принадлежит текст конкретному автору или нет. В качестве алгоритмов классификации рассматриваются AdaBoost и нейросеть со слоем LSTM. Эксперименты демонстрируют эффективность ритмических характеристик при верификации конкретных авторов и превосходство комбинаций типов характеристик над отдельными типами характеристик в среднем. Лучшее значение точности, полноты и F-меры для классификатора AdaBoost превышает 90%, когда комбинируются все три типа характеристик.

Ключевые слова: стилометрия; обработка естественного языка; стилистические характеристики; ритмические характеристики; верификация авторов

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ксения Владимировна Лагутина
автор для корреспонденции

orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
аспирант.

Финансирование: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-37-90045.

Для цитирования: K. V. Lagutina, “Comparison of Style Features for the Authorship Verification of Literary Texts”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 250-259, 2021.

Introduction

The authorship verification is the task of determination whether the text belongs to a given author or not. It is based on the assumption that the author has the individual set of style markers that can distinguish the author from others, but occurs in each of his/her texts [1].

In the state-of-the-art of the authorship verification and close text classification tasks there is no set of style features that would be versatile for different texts. Some feature types like character-level, word-level, and syntactic features appear in many investigations, but are often combined with more complex linguistic features [2, 3]. Researchers admit that the influence of different types of features on the quality of text classification remains underexplored [4].

Rhythm features are the subtype of the linguistic features that most often describe the style of literary texts [5]. They can be applied for authorship verification [6], but are rarely compared with other feature types [2].

The goal of this paper is comparing how different feature types affect the quality of the authorship verification of literary texts. We analyse rhythm features and popular low-level features based on statistics of text elements. The comparison is performed on the corpora of English, Russian, French, and Spanish literary texts.

1. State-of-the-art

The task of authorship verification is usually performed for the texts from the Internet: news articles, emails, reviews, etc. [2, 7].

In many cases the researchers modeled texts using only standard low-level features and experimented with classification. Halvani et al. [8] used stylometric features based on n-grams. The verification was realized by the determination of the proximity of the numerical feature vectors of the texts. Experiments were conducted in five European languages: Dutch, English, Greek, Spanish, and German. The F-measure varied from 67.37 % for Greek up to 83.33 % for Spanish. The method also showed good results at the PAN-2020 competition [9].

Potha and Stamatatos [7] introduced an intrinsic profile-based verification method that apply latent semantic indexing for topic modeling and low-level features: word and character n-grams. Then the algorithm calculated the text model that represents all texts of the same author as a common vector. Then it identified the authors by searching for test texts the closest vector from the authors' train ones. The researchers compared in experiments corpora of prose, newspaper articles, reviews in four languages: Dutch, English, Greek, and Spanish. The method achieved more than 80 % of the AUC.

Boenninghoff et al. [10] proposed a new neural network topology to identify whether two documents with unknown authors were written by the same author. This approach showed the best results of the precision, recall, and F-measure 84 % for short multi-genre social media posts.

Adamovic et al. [11] searched a wide range of word and character-based language-independent text stylistic features. Then they applied the SVM-RFE feature selection method to remove redundant and irrelevant characteristics. Authorship verification of articles in four languages: English, Greek, Spanish, and German showed a high result over 90 % of the accuracy.

To improve the quality of authorship verification and take into account domain peculiarities and the authors' idiolect, the researchers frequently applied linguistic features.

Al-Khatib and Al-qaoud [12] verified native and non-native speakers of online opinion articles. The feature set included statistical and linguistic features: number of unique words, complexity, Gunning-Fog readability index, character space, letter space, average syllables per word, sentence count, average sentence length, and the Flesch-Kincaid Readability. The accuracy varied for text corpora from 47 % to 77 %.

Lagutina et al. [6] investigated application of rhythm features to the authorship verification of the artistic prose. They found the features based on repetitions of words and sentences (anaphora, epiphora,

aposiopesis, etc.) and verified authors of English, Russian, French, and Spanish prose. The F-measure achieved from 60 % to 95 % for different authors and about 80 % in average.

The literary texts are usually analysed not in the authorship verification but in the close task of the authorship attribution. For example, Stanisz et al. [13] created adjacency networks with words frequently appearing in texts, and their co-occurrences as vertices and edges' weights. Then the authors computed various graph characteristics: clustering coefficients of vertices, an average shortest path length, an assortativity coefficient, and modularity. The experiments showed the accuracy of 85–90 % for English and Polish books.

The analysis of the state-of-the-art papers shows the lack of comparison of different feature types with linguistic ones, especially for artistic texts. The authors usually rely on standard statistical features based on words and characters and try to extend them by relatively small number of syntactic, topical, or other linguistic features. Deep linguistic features remains under-researched, most probably, because of their complexity in search. Although such features are directly identify the author's style [5] and can be the most interpretable ones.

2. Style features

We compare three types of features: character-level, word-level, and rhythm-level ones. The first two feature types are the popular effective features from the state-of-the art. The rhythm features describe the specific style marks of the authors that frequently appear in literary texts.

Before feature calculation we search in plain texts the following elements:

- Top-40 unigrams and top-40 bigrams of words among the text corpora. They will be used for computing frequencies of occurrences for n-grams.
- Lexico-grammatical rhythm figures. For each text we found the lists of the following figures: anaphora, epiphora, symploce, anadiplosis, diacope, epizeuxis, epanalepsis, chiasmus, polysyndeton, repeating exclamatory sentences, repeating interrogative sentences, and aposiopesis. Their definitions and search algorithms are taken from the works of Lagutina et al. [6, 14]. The quality of figures search achieves 80–95 % of precision.

We compute the following style features:

- Character-level features:
  - Average sentence length in characters including punctuation marks and spaces.
  - Frequencies of occurrences of each letter among all letters. The uppercase letters are previously reduced to lowercase ones.
  - Frequencies of occurrences of each punctuation mark (.!?:, etc.) among all punctuation.
- Word-level features:
  - Average sentence length in words.
  - Average word length in characters.
  - Frequencies of occurrences of unigrams and bigrams among top-40 n-grams.
- Rhythm features:
  - The density of the figure — the number of occurrences of the rhythm figure (anaphora, epiphora, etc.) divided by the number of sentences.
  - The fraction of unique words—words that appear only once in rhythm figures.
  - The fraction of words of a particular part of speech (noun, verb, adverb, and adjective) in rhythm figures.

All features are calculated separately for each text. Character and word-level features represent the base statistics of the text style. Rhythm features represent the density and linguistic structure of the text rhythm. So the text is modeled as the vector of statistical and linguistic features.

3. Authorship verification

3.1. Design of authorship verification

After feature extraction we get the matrix where rows are texts of particular authors, columns are feature types. We verify each author separately using the whole matrix for the author's language. His/her texts are labeled as belonging or not belonging to him/her. Then the binary classification is performed.

Two classifiers are compared: AdaBoost and Bidirectional LSTM. They have already show their quality in solution of state-of-the-art text classification tasks [15].

The AdaBoost classifier combines the results of 50 Decision Tree classifiers. The Bidirectional LSTM neural network contains the Bidirectional LSTM layer with 64 units and a dense output layer with the sigmoid activation function. The loss function is categorical cross-entropy, the optimization algorithm is Adam, the number of epochs is 100.

In order to estimate the stability of classifiers, we apply the five-fold cross-validation technique: 80 % of texts are the training samples, 20 % are the test ones. The estimation is performed with three standard measures: precision, recall, and F-measure [16], and also their standard deviations.

The code for the feature selection and authorship verification is published at https://github.com/text-processing/prose-rhythm-detector. It is written in Python programming language and uses Stanza 1.1.1 NLP library for text representation and determination of parts of speech. For the verification it uses Scikit-Learn 0.23.2 and Keras 2.4.3.

3.2. Text corpora

We compare literary texts of four languages: English, Russian, French, and Spanish. The corpora were created manually collecting famous works of famous authors written in their native language.

In order to make texts equal in size, we extracted 1–4 fragments with the size about 50 000 characters including spaces from each prose text. In such a way each author is presented by 40 text fragments. English, Russian, and French corpora contain texts of 20 famous authors of 19th–21st centuries, 800 texts per corpora. The Spanish corpus has texts of 8 authors of 19-th–20th centuries, 320 texts in total.

4. Experiments

During experiments we compare features of three types: 36–43 character-level features (the letters differs for corpora in different languages), 82 word-level features, and 17 rhythm features.

Comparing two classifiers, we discover that AdaBoost outperforms the neural network by 10–15 % of precision, recall, and F-measure. Most probably, it happens due to the fact that the training sample has the insufficient size for better performance of the LSTM network. So the tables in this section contains classification quality for the AdaBoost algorithm.

Table 1 describes authorship verification quality for all feature types and their combinations. Ch means character-level features, W — word-level ones, Rh — rhythm ones, + marks the combination of two feature types, All — the combination of three feature types. Precision, recall, and F-measure are calculated as the averages for all authors. Bold marks the lines with best quality and best F-measures.

From Table 1 we can see that rhythm features provide the good classification quality. It is lower by 3–11 % of F-measure in the most cases, but has quite high values of 78–87 %. Besides, the number of rhythm features is several times less than character- and word-level ones, so the relatively small number of specific style parameters allow to achieve significant authorship verification quality.

Any combination of feature types improve quality by 2–14 %, but the combination of all types is slightly higher than of the two types.

Authors of Russian, French, and Spanish texts in most cases are verified better than English. In English and French texts the best feature type is character-level, In Russian and Spanish texts it is word-level.

Table 1. Mean measure values of the authorship verification

Language	Feature type	Precision	Recall	F-measure
English	Ch	87.8	80.7	84.1
English	W	85.8	78.2	81.8
English	Rh	82.0	74.2	77.9
English	Ch + W	92.2	84.0	**88.0**
English	Ch + Rh	90.8	80.9	85.6
English	W + Rh	88.8	81.7	85.1
English	All	**94.7**	**85.4**	**89.8**
Russian	Ch	91.2	81.4	86.0
Russian	W	92.0	81.9	86.7
Russian	Rh	84.7	76.7	80.5
Russian	Ch + W	96.9	86.7	**91.5**
Russian	Ch + Rh	94.3	85.4	89.6
Russian	W + Rh	92.2	82.6	87.1
Russian	All	**96.9**	**87.4**	**91.9**
French	Ch	93.7	86.5	90.0
French	W	91.8	80.1	85.6
French	Rh	83.5	75.9	79.5
French	Ch + W	95.4	89.2	**92.2**
French	Ch + Rh	96.2	86.6	**91.2**
French	W + Rh	93.3	83.0	87.9
French	All	**97.5**	**90.0**	**93.6**
Spanish	Ch	89.9	85.0	87.4
Spanish	W	92.3	87.9	90.1
Spanish	Rh	88.5	86.3	87.4
Spanish	Ch + W	92.5	87.8	90.1
Spanish	Ch + Rh	94.5	88.8	**91.6**
Spanish	W + Rh	93.7	88.6	91.1
Spanish	All	**94.1**	**90.0**	**92.0**

Table 2. Verification of English authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
W. Scott	Ch	95.4	8.4	88.0	6.9	88.5	5.7
W. Scott	W	95.9	6.6	91.7	9.0	92.1	5.9
W. Scott	Rh	89.6	5.6	88.8	5.6	89.1	5.5
W. Scott	Ch + W	**99.5**	0.3	92.6	11.6	91.9	7.6
W. Scott	Ch + Rh	98.1	2.2	76.9	16.6	89.1	**2.7**
W. Scott	W + Rh	98.1	2.2	**94.2**	7.7	88.7	12.6
W. Scott	All	97.7	3.3	93.5	6.1	**95.2**	4.6
Z. Smith	Ch	94.4	10.0	**86.8**	16.6	82.6	10.5
Z. Smith	W	48.4	0.5	57.2	10.1	59.1	20.4
Z. Smith	Rh	62.0	6.7	62.5	7.8	63.4	**5.4**
Z. Smith	Ch + W	89.0	19.8	83.2	21.0	**87.0**	11.8
Z. Smith	Ch + Rh	90.2	11.4	85.0	14.6	82.7	10.7
Z. Smith	W + Rh	53.5	10.2	67.7	19.1	52.0	5.5
Z. Smith	All	**99.3**	0.6	77.3	13.5	66.4	16.4
N. Gaiman	Ch	91.5	7.7	**77.5**	11.4	64.5	11.2
N. Gaiman	W	64.2	9.6	66.8	12.2	67.5	15.5
N. Gaiman	Rh	81.5	6.7	74.6	10.3	78.7	8.0
N. Gaiman	Ch + W	**94.5**	5.6	68.7	12.4	81.3	**6.5**
N. Gaiman	Ch + Rh	92.5	7.5	69.5	13.8	**82.2**	8.6
N. Gaiman	W + Rh	94.3	8.5	69.1	12.6	74.1	7.3
N. Gaiman	All	90.0	7.5	75.1	12.6	80.8	8.7

Table 3. Verification of Russian authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
M. Bulgakov	Ch	84.2	19.9	66.7	10.5	66.4	15.4
M. Bulgakov	W	76.0	22.4	68.2	11.2	83.8	18.7
M. Bulgakov	Rh	73.7	22.6	68.3	13.0	68.5	17.6
M. Bulgakov	Ch + W	**99.3**	0.5	76.7	12.2	**82.3**	11.9
M. Bulgakov	Ch + Rh	89.0	19.5	74.0	13.1	81.9	**9.6**
M. Bulgakov	W + Rh	75.7	22.0	60.0	8.8	75.2	13.2
M. Bulgakov	All	89.2	19.9	**78.2**	19.3	78.2	18.4
N. Leskov	Ch	**99.3**	0.6	70.7	12.2	69.1	11.5
N. Leskov	W	89.3	19.9	72.8	17.3	75.2	21.8
N. Leskov	Rh	85.0	8.9	78.7	8.4	80.7	11.7
N. Leskov	Ch + W	94.5	10.0	77.5	12.2	87.7	6.7
N. Leskov	Ch + Rh	88.4	13.1	80.1	16.5	87.8	10.8
N. Leskov	W + Rh	82.4	21.0	82.0	16.5	91.4	**5.6**
N. Leskov	All	89.5	20.7	**85.3**	18.1	**96.6**	6.8
A. Prokhanov	Ch	98.0	3.2	98.8	2.2	89.0	19.7
A. Prokhanov	W	96.9	3.4	94.1	4.9	96.0	6.7
A. Prokhanov	Rh	98.1	3.2	**98.9**	1.7	96.3	3.6
A. Prokhanov	Ch + W	98.2	3.3	98.5	2.8	97.9	2.9
A. Prokhanov	Ch + Rh	**99.9**	0.2	98.6	2.9	96.7	3.2
A. Prokhanov	W + Rh	96.1	4.5	95.7	3.7	97.1	2.5
A. Prokhanov	All	96.9	5.6	93.1	8.6	**98.4**	**2.0**

Table 4. Verification of French authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
S. Colette	Ch	95.2	1.4	94.1	2.2	96.3	2.9
S. Colette	W	94.1	3.3	88.3	3.5	92.6	3.8
S. Colette	Rh	92.3	3.7	88.1	3.6	90.2	4.6
S. Colette	Ch + W	97.0	1.8	96.1	2.3	95.6	2.0
S. Colette	Ch + Rh	**99.6**	0.2	95.1	3.9	98.1	2.2
S. Colette	W + Rh	96.0	1.8	94.2	5.2	95.0	3.6
S. Colette	All	**99.6**	0.3	**97.7**	3.4	**98.6**	**1.2**
V. Hugo	Ch	94.9	4.7	81.7	7.2	81.4	6.7
V. Hugo	W	92.4	6.0	73.3	6.7	80.4	**3.9**
V. Hugo	Rh	73.1	8.5	66.1	3.8	70.3	8.4
V. Hugo	Ch + W	**99.0**	0.7	80.3	6.2	83.3	8.1
V. Hugo	Ch + Rh	94.9	5.6	76.6	10.7	**88.6**	4.6
V. Hugo	W + Rh	85.1	14.0	75.9	6.2	82.4	6.8
V. Hugo	All	93.9	5.8	**85.0**	5.8	83.9	11.9
A. Exupery	Ch	96.1	6.4	80.3	16.7	67.6	**10.7**
A. Exupery	W	89.2	20.1	63.8	19.7	65.2	20.6
A. Exupery	Rh	99.3	0.5	69.2	11.7	70.5	12.0
A. Exupery	Ch + W	84.3	20.1	**78.5**	16.6	82.4	18.6
A. Exupery	Ch + Rh	89.6	20.0	73.6	17.0	81.7	12.6
A. Exupery	W + Rh	99.3	0.5	58.3	10.6	65.5	15.9
A. Exupery	All	**99.6**	0.3	75.3	8.5	**91.4**	15.0

Table 5. Verification of Spanish authors

Author	Feature type	Precision	Std dev	Recall	Std dev	F-measure	Std dev
V. Ibáñez	Ch	92.8	2.0	88.5	4.9	90.0	8.0
V. Ibáñez	W	98.2	1.5	94.6	3.7	97.1	2.2
V. Ibáñez	Rh	96.7	3.3	95.1	3.9	93.0	3.2
V. Ibáñez	Ch + W	95.1	4.0	95.6	4.2	95.4	2.2
V. Ibáñez	Ch + Rh	96.9	2.1	94.8	1.3	96.3	3.0
V. Ibáñez	W + Rh	99.3	0.6	**99.2**	1.2	**99.2**	**1.0**
V. Ibáñez	All	**99.4**	0.8	97.4	1.7	98.5	1.8
J. Dicenta	Ch	79.3	24.5	79.8	24.4	71.1	19.7
J. Dicenta	W	79.1	24.4	**85.0**	20.0	**88.3**	**11.7**
J. Dicenta	Rh	82.4	21.3	62.9	19.8	61.1	14.6
J. Dicenta	Ch + W	78.8	24.7	70.0	19.4	79.5	19.9
J. Dicenta	Ch + Rh	89.2	20.0	65.0	20.0	78.2	23.8
J. Dicenta	W + Rh	**89.4**	20.1	70.0	24.5	74.4	20.9
J. Dicenta	All	79.4	24.9	60.0	20.0	71.1	19.7

Tables 2–4 illustrate the typical cases of the authorship verification. Columns "Std dev" contain standard deviations of the measures in the left.

Almost all authors have very high precision of verification 78–99 %. Recall is varied significantly more: 60–98 %.

Several authors are verified with high quality 88–96 % of the F-measure by any feature type: W. Scott, A. Prokhanov, S. Colette, V. Ibáñez. The combination of feature types improves classification even more up to 96–99 %. We can say that the chosen features describe the style of such authors quite effectively.

Several authors are verified with lower quality 66–88 % of the F-measure: Z. Smith, M. Bulgakov, V. Hugo, J. Dicenta. They also have very high standard deviations of 10–20 %. We can point out the feature types that provide quite good precision and recall (usually there are word-level features or combinations with them). But we can conclude that the proposed feature set does not describe common details in the style of such authors.

For some authors the F-measure grows significantly for combination of features. For example, texts of N. Leskov are verified with 69–80 % of F-measure, combinations of two feature types provides the F-measure of 87–91 %, and the combination of all features allows to achieve the best value of 96 %. Besides, the N. Leskov's style is better described by rhythm features than by others, because rhythm features provide higher F-measure of 80 % against 69 % and 75 %. Texts of A. Exupery show the same tendencies.

Several authors are verified significantly better by rhythm features than by statistical ones, for example, N. Gaiman, N. Leskov, A. Exupery.

Thus, all feature types can provide good verification quality. The specific linguistic features — rhythm features — achieve in many cases high precision, recall, and F-measure with small standard deviations. So they are as useful and stable style markers as standard statistical features: character and word level ones.

Verification of particular authors shows that the many authors have the same style in different fragments. They can be successfully separated from others using only one feature type or the combination of standard and rhythm features. Nevertheless, texts of several authors are verified with very high standard deviations, so there are needed other linguistic features to verify reliably their text fragments.

Conclusion

We applied three types of style features: character, word, rhythm-level features, and their combinations to the authorship verification of literary texts in English, Russian, French, and Spanish. Experiments revealed the same tendencies for all four languages. In average combinations of features provide higher classification quality than the single feature type. Moreover, rhythm features are almost as good style markers as popular low-level features.

The more detailed analysis of authorship verification allowed to discover the fact that many authors write text fragments in the same style, so they can be successfully verified by the single feature type or by the combination. But several authors are verified with significantly lower quality than others. The future investigations can be devoted to the error analysis of classification of their texts and search of the larger set of linguistic style markers that help to verify more authors.

References

[1] E. Stamatatos, "A survey of modern authorship attribution methods", *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[2] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, and I. Paramonov, "A survey on stylometric text features", in *Proceedings of the 25th Conference of Open Innovations Association (FRUCT)*, IEEE, 2019, pp. 184–195.

[3] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry techniques and applications", *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–36, 2018.

[4] C.-G. Lim, Y.-S. Jeong, and H.-J. Choi, "Survey of Temporal Information Extraction.", *Journal of Information Processing Systems*, vol. 15, no. 4, pp. 931–956, 2019.

[5] E. Boychuk, I. Paramonov, N. Kozhemyakin, and N. Kasatkina, "Automated approach for rhythm analysis of French literary texts", in *Proceedings of 15th Conference of Open Innovations Association FRUCT*, IEEE, 2014, pp. 15–23.

[6] K. Lagutina, N. Lagutina, E. Boychuk, V. Larionov, and I. Paramonov, "Authorship Verification of Literary Texts with Rhythm Features", in *Proceedings of the 28th Conference of Open Innovations Association FRUCT*, 2021, pp. 240–251. DOI: 10.23919/FRUCT50888.2021.9347649.

[7] N. Potha and E. Stamatatos, "Intrinsic author verification using topic modeling", in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, ACM, 2018, pp. 1–7.

[8] O. Halvani and L. Graner, "Rethinking the evaluation methodology of authorship verification methods", in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2018, pp. 40–51.

[9] O. Halvani, L. Graner, and R. Regev, "TAVeer: an interpretable topic-agnostic authorship verification method", in *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 2020, pp. 1–10.

[10] B. Boenninghoff, R. M. Nickel, S. Zeiler, and D. Kolossa, "Similarity learning for authorship verification in social media", in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2457–2461.

[11] S. Adamovic, V. Miskovic, M. Milosavljevic, M. Sarac, and M. Veinovic, "Automated language-independent authorship verification (for Indo-European languages)", *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 858–871, 2019.

[12] M. A. Al-Khatib and J. K. Al-qaoud, "Authorship verification of opinion articles in online newspapers using the idiolect of author: a comparative study", *Information, Communication & Society*, pp. 1–19, 2020.

[13] T. Stanisz, J. Kwapień, and S. Drożdż, "Linguistic data mining with complex networks: a stylometric-oriented approach", *Information Sciences*, vol. 482, pp. 301–320, 2019.

[14] K. Lagutina, A. Poletaev, N. Lagutina, E. Boychuk, and I. Paramonov, "Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries", in *Proceedings of the 26th Conference of Open Innovations Association FRUCT*, IEEE, 2020, pp. 247–255.

[15] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey", *Information*, vol. 10, no. 4, 150 (1–68), 2019.

[16] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF DATA

Analysis of the Impact of the Stylometric Characteristics of Different Levels for the Verification of Authors of the Prose

A. M. Manakhova[1], N. S. Lagutina[1]

DOI: 10.18255/1818-1015-2021-3-260-279

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received June 25, 2021
After revision August 23, 2021
Accepted August 25, 2021

This article is dedicated to the analysis of various stylometric characteristics combinations of different levels for the quality of verification of authorship of Russian, English and French prose texts. The research was carried out for both low-level stylometric characteristics based on words and symbols and higher-level structural characteristics.
All stylometric characteristics were calculated automatically with the help of the ProseRhythmDetector program. This approach gave a possibility to analyze the works of a large volume and of many writers at the same time. During the work, vectors of stylometric characteristics of the level of symbols, words and structure were compared to each text. During the experiments, the sets of parameters of these three levels were combined with each other in all possible ways. The resulting vectors of stylometric characteristics were applied to the input of various classifiers to perform verification and identify the most appropriate classifier for solving the problem. The best results were obtained with the help of the AdaBoost classifier. The average F-score for all languages turned out to be more than 92 %. Detailed assessments of the quality of verification are given and analyzed for each author. Use of high-level stylometric characteristics, in particular, frequency of using N-grams of POS tags, offers the prospect of a more detailed analysis of the style of one or another author. The results of the experiments show that when the characteristics of the structure level are combined with the characteristics of the level of words and / or symbols, the most accurate results of verification of authorship for literary texts in Russian, English and French are obtained. Additionally, the authors were able to conclude about a different degree of impact of stylometric characteristics for the quality of verification of authorship for different languages.

Keywords: stylometry; stylometric characteristics; authorship verification; natural language processing

INFORMATION ABOUT THE AUTHORS

Alla Mikhajlovna Manakhova correspondence author	orcid.org/0000-0001-7429-3529. E-mail: al.mnkhv@yandex.ru MSc student.
Nadezhda Stanislavovna Lagutina	orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru Associate Professor, PhD in Physics and Mathematics.

For citation: A. M. Manakhova and N. S. Lagutina, “Analysis of the Impact of the Stylometric Characteristics of Different Levels for the Verification of Authors of the Prose”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 260-279, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Анализ влияния стилометрических характеристик разного уровня на верификацию авторов художественных произведений

А. М. Манахова[1], Н. С. Лагутина[1]

DOI: 10.18255/1818-1015-2021-3-260-279

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 25 июня 2021 г.
После доработки 23 августа 2021 г.
Принята к публикации 25 августа 2021 г.

Данная статья посвящена анализу влияния различных комбинаций стилометрических характеристик разного уровня на качество верификации авторства русских, английских и французских прозаических текстов. Исследование проводилось как для низкоуровневых стилометрических характеристик, основанных на словах и символах, так и для более высокоуровневых – структурных.

Подсчёт всех стилометрических характеристик был выполнен автоматически с помощью программы ProseRhythmDetector. Такой подход позволил провести анализ произведений большого объёма и многих писателей одновременно. В ходе работы каждому тексту были сопоставлены векторы стилометрических характеристик уровня символов, слов и структуры. При проведении экспериментов наборы параметров этих трёх уровней были скомбинированы между собой всеми возможными способами. Полученные векторы стилометрических характеристик были поданы на вход различным классификаторам для выполнения верификации и выявления наиболее подходящего классификатора для решения поставленной задачи. Лучшие результаты были получены с помощью классификатора AdaBoost. Средняя F-мера для всех языков оказалась более 92%. Детальные оценки качества верификации приведены для каждого автора и проанализированы. Использование высокоуровневых стилометрических характеристик, в частности, частоты использования N-грамм POS-тегов открывает перспективу более детального анализа стиля того или иного автора. Результаты экспериментов показывают, что при соединении характеристик уровня структуры с характеристиками уровня слов и/или символов получаются наиболее точные результаты верификации авторства для художественных текстов на русском, английском и французском языках. Дополнительно авторам удалось сделать вывод о разной степени влияния стилометрических характеристик на качество верификации авторства для различных языков.

Ключевые слова: стилометрия; стилометрические характеристики; верификация авторства; обработка естественного языка

ИНФОРМАЦИЯ ОБ АВТОРАХ

Алла Михайловна Манахова автор для корреспонденции	orcid.org/0000-0001-7429-3529. E-mail: al.mnkhv@yandex.ru магистрант.
Надежда Станиславовна Лагутина	orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru доцент, кандидат физико-математических наук.

Для цитирования: A. M. Manakhova and N. S. Lagutina, “Analysis of the Impact of the Stylometric Characteristics of Different Levels for the Verification of Authors of the Prose”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 260-279, 2021.

Введение

Одной из проблем извлечения информации из неструктурированных данных является верификация автора текста на естественном языке [1]. Решение этой задачи заключается в определении принадлежности текста заданному автору. Основная часть исследований в этой области посвящена верификации авторов электронных писем и сообщений в социальных сетях [2]. Аналогичной задачей является разрешение споров об авторских правах [3]. Кроме анализа современных текстов, актуальной проблемой остаётся верификация авторов художественных произведений [4—6].

Основным методом автоматической верификации авторства является классификация текстов с помощью векторов стилометрических характеристик [7]. В первую очередь такими характеристиками являются простые текстовые функции, например, частоты слов и символов, n-граммы символов и слов, длины слов и предложений. В последнее время всё больше авторов обращают внимание на определение особенностей синтаксиса и грамматики. Количество таких параметров очень велико, может достигать нескольких тысяч. Однако не все характеристики вносят одинаковый вклад в решение задачи. Наиболее сложным для анализа стилем обладают художественные произведения. Поэтому авторы поставили перед собой задачу исследовать влияние различных групп стилометрических параметров на верификацию авторов повестей на английском, французском и русском языках.

1. Современное состояние исследований в области верификации авторства текстов

Задача верификации авторства может рассматриваться как математическая задача бинарной классификации, принадлежит ли рассматриваемый документ определённому классу или нет. Для текстов на естественном языке формируются характеристические числовые векторы, затем применяется один из методов классификации, которым очень часто является метод машинного обучения. Многие исследователи для вычисления вектора признаков документа используют, ставшие классическими в компьютерной лингвистике, параметры: частоты символов и слов, n-граммы слов, эмбеддинги слов [7]. Популярные низкоуровневые текстовые функции: униграммы слов и n-граммы символов использовали авторы статьи [8]. Они разработали метод проверки на основе внутреннего профиля автора, который создаёт текстовую модель, представляющую все тексты одного автора как общий вектор. Исследователи экспериментировали с корпусами прозы, газетными статьями, обзорами и другими жанрами из соревнований PAN-2014 и PAN-2015 на четырех языках: голландском, английском, греческом и испанском. Разработанный подход показал в экспериментах точность (accuracy) более 80 % и позиционировался авторами как независимый от языка. Ещё более высокий результат 90 % показал метод, описанный в работе [9]. Авторы выделили параметры стиля текста, не зависящие от языка и использовали метод выбора признаков SVM-RFE (Support Vector Machine Based on Recursive Feature Elimination) для удаления избыточных и нерелевантных характеристик из процесса обучения. Метод был применён к верификации авторства статей на четырёх языках: английском, греческом, испанском и немецком.

Однако использование параметрических векторов из простых характеристик имеет свои ограничения и недостатки. Авторы работы [10] обращают внимание, что надежность использования таких параметров в алгоритмах машинного обучения значительно снижается для коротких и тематически разнообразных текстов в социальных сетях. Авторы решают эту проблему попыткой визуализировать процесс принятия решений нейронной сетью. Для верификации авторства коротких обзоров Amazon исследователи используют сиамские нейронные сети. При обсуждении результатов авторы проводят лингвистический анализ внутренних весов сети с целью привязать результат к некоторым традиционным лингвистическим категориям. В данной работе используется корпус текстов большого объёма 9052606 отзывов, написанных 784649 авторами, что, конечно, способствует повыше-

нию качества решения задачи. Другие исследователи использовали простое моделирование текста с помощью эмбеддинга на основе векторного представления слов Word2vec [11] для верификации авторства коротких статей на английском языке. Они добились увеличения качества классификации в модели машинного обучения, основанной на слиянии трех разных архитектур: сверточных нейронных сетей, рекуррентно-сверточных нейронных сетей и машинных классификаторов опорных векторов. Окончательное решение получается путем объединения результатов трех моделей с использованием метода голосования. В результате экспериментов точность (accuracy) оказалась от 91 % до 97 %.

Однако большое количество современных исследований проблемы верификации авторов идёт по пути совершенствования характеристических векторов текстов. В качестве параметров добавляются более сложные текстовые функции.

Ли и соавторы [12] предприняли попытку применить функции специфичные для предметной области. Они использовали 233 функции, включая 227 стилометрических функций и шесть новых специфических для социальных сетей функций. Стилометрические функции включали уровень символов: частоты отдельных букв, прописных букв, специальных символов; уровень слов: общее количество слов, средняя длина слова, количество слов с одним символом и т. д .; синтаксические: количество знаков препинания и функциональных слов, общее количество предложений. Набор специфичных для социальных сетей функций включал смайлики, сокращения, начало предложения без заглавной буквы, окончание предложения без знака препинания и отсутствие упоминания «Я» или «Мы» в сообщении. Они разработали алгоритмы и исследовали различные классификаторы для определения подлинности коротких сообщений социальной сети Facebook. Результаты экспериментов по проверке, является ли указанный пользователь автором данного сообщения, показали среднюю точность 79,6 % для 30 пользователей и 9259 сообщений. Это качество было достигнуто за счет стилометрических характеристик. Функции на основе предложений показали худшую производительность с точностью 53,6 %. Особенности социальных сетей не улучшили классификацию. Тот факт, что особенности, основанные на предложениях, не повлияли на качество классификации, можно объяснить особенностью коротких сообщений в социальных сетях, поскольку они редко состоят из большого количества предложений. Более интересно, что специальные символы социальных сетей влияют на решение данной конкретной задачи гораздо меньше, чем лексические и синтаксические особенности.

Определение авторства художественных произведений связывается в первую очередь со сложными стилистическими характеристиками текста. При решении вопроса о подлинности текстов Плиния [13], оценки принадлежности спорного произведения Данте [5], решения проблемы проверки авторства Гёте по отношению к анонимным статьям [4] рассматриваются стилометрические параметры основанные на специфических фразах, характерных для автора или времени написания. Однако авторы подчёркивают неоднозначность полученных результатов и необходимость продолжения исследований. Важность выбора релевантных характеристик подчёркивают авторы работы, посвящённой верификации античных авторов [14].

Исследователи используют параметры, вычисляемые на основе структуры текста. В статье [15] представлен подход к решению проблемы проверки авторства в области судебной медицины. Разработанный метод использует графы для представления лексических и синтаксических аспектов текстов. На основе этих структур данных вычисляются лингвистические функции, которые позволяют выявить стиль письма автора. Предлагаемый метод применяется к англоязычным документам.

Следует отметить, что исследователи национальных языков также обращают внимание на эффективность использования структурных характеристик текста [16]. В этой работе задача верификации авторов решается для книг на классическом арабском языке. Метод, предлагаемый учёными основан на сходстве лингвистических особенностей текста и применяет ряд лексических, мор-

фологических и синтаксических признаков и ансамблей признаков. В результате экспериментов точность достигла 87,1 %, однако корпус текстов состоял всего из 31 книги.

Следует обратить внимание, что появление сложных текстовых функций позволяет не только получить ответ классификатора, как чёрного ящика, но и проводить анализ результатов в рамках предметной области и лингвистики. Такие исследования могут оказаться очень полезными, так как дают возможность собирать информацию о деталях стиля написания текста, проводить качественный анализ ошибок, выявлять ограничения методов, прогнозировать возможность их применения в предметных областях. Например, авторам работы [17] удалось показать, что у каждого писателя есть свой стиль, который проявляется в использовании нескольких типов стилометрических характеристик, уникальных для отдельных авторов. Таким образом, выделение нескольких групп параметров текста на разных уровнях: символов, слов, структуры предложений, и исследование их влияния на качество верификации авторства является актуальной задачей в области автоматической обработки естественного языка.

2. Обзор характеристик

Стилометрический анализ текста включает в себя поиск и подсчёт различных стилометрических характеристик. Выбор этих характеристик текста является важнейшим этапом его исследования. Среди таких характеристик можно выделить несколько категорий:

1. Уровень символов:
   (a) количество отдельных букв;
   (b) общее количество букв;
   (c) количество отдельных символов;
   (d) общее количество символов
   (e) средняя длина предложения в символах;
2. Уровень слов:
   (a) количество слов;
   (b) количество предложений;
   (c) средняя длина предложений по количеству слов;
   (d) средняя длина слова;
   (e) частота встречаемости N-грамм из одного, двух и трёх слов;
3. Структурный уровень:
   (a) частота встречаемости POST N-грамм из одного, двух, трёх и четырёх слов.

Для проведения исследований в области верификации авторства авторами были выбраны именно эти стилометрические характеристики, потому что они позволяют наиболее точно определить авторский стиль произведения [18].

3. Классификация текстов

Классификация является одной из важных задач в рамках обработки естественного языка. Она решается с помощью специальных аналитических моделей, называемых классификаторами. В настоящее время существует большое количество различных видов классификаторов, для построения которых используются как статистические методы (логистическая регрессия), так и методы машинного обучения (нейронные сети, деревья решений, метод k-ближайших соседей, машины опорных векторов). Одним из наиболее важных этапов в задаче классификации текста является выбор классификатора. Без этого не возможно определить наиболее эффективную модель для алгоритма классификации текста. Это обусловлено тем, что решаемые задачи могут иметь особенности, связанные с числом классов или с объёмом и качеством исходных данных.

Для решения поставленной задачи авторами были использованы следующие классификаторы:

1. Классификатор DecisionTree: Одним из популярных алгоритмов классификации для анализа текста и данных является дерево решений [19]. Структура этого метода представляет собой иерархическую декомпозицию пространства признаков. Каждый лист дерева представляет собой значение целевой переменной, каждый внутренний узел соответствует одному из признаков. Дерево может быть построено разделением исходных наборов характеристик на подмножества, основанные на проверке значений этих признаков. Каждый узел дерева содержит условие ветвления по одному из признаков.
2. Классификатор Random Forest: Random Forest является ансамблем множества деревьев решений [20]. Это позволяет повысить точность классификации по сравнению с одним деревом. Результат классификации получается в итоге агрегирования ответов множества деревьев.
3. Классификатор SVM: Задача бинарной классификации с помощью метода опорных векторов (SVM) состоит в построении оптимальной разделяющей гиперплоскости в пространстве признаков текстов высокой размерности. Задача классификации на несколько классов с помощью метода опорных векторов заключается в переходе от задачи классификации на множество классов к множественной задаче разбиения на два класса. Первый вариант перехода соответствует стратегии «один против всех». Обучается несколько классификаторов, в соответствии с количеством классов. Классификатор с самым лучшим значением функции выхода присваивает текст к соответствующему классу. Второй вариант перехода соответствует стратегии «один против одного». Также обучается несколько классификаторов по количеству классов. Текст классифицируется в соответствии с тем к какому классу его отнесло большинство классификаторов.
4. Классификатор Gaussian Naive Bayes: Метод наивного байесовского классификатора, применяемого для классификации текстов [21], основан на теореме Байеса:

$$P(c|d) = \frac{P(d|c) * P(c)}{P(d)},$$

где, $P(c|d)$ — вероятность, что текст d принадлежит классу c; $P(d|c)$ — вероятность встретить текст d среди всех текстов класса c; $P(c)$ — безусловная вероятность встретить текст класса c в корпусе текстов; $P(d)$ — безусловная вероятность текста d в корпусе текстов.
5. Классификатор AdaBoost Классификатор AdaBoost используется с целью повышения точности алгоритмов классификации. Он лучше всего работает с обучающими алгоритмами, которые показывают наименее точные результаты (слабые обучающие алгоритмы). Наиболее распространенными алгоритмами, используемыми с AdaBoost, являются одноуровневые деревья решений. Кроме того, слабые классификаторы легко вычисляются, и поэтому появляется возможность объединять много сущностей алгоритма, для создания более сильного классификатора с помощью бустинга.

Векторы стилометрических характеристик были поданы на вход каждому из классификаторов для последующего анализа.

С целью проверки устойчивости классификаторов была применена техника пятикратной кроссвалидации. Тексты были разделены на пять частей, 80 % из которых составляли тренировочную выборку, а оставшиеся 20 % — тестовую. Оценка качества выполнялась с помощью таких параметров, как точность, полнота и F-мера.

4. Эксперименты

4.1. Корпус текстов

Для проведения экспериментов использовались корпуса художественной литературы на русском языке (724 фрагмента за период с 1832 до 2020 года), на английском языке (664 фрагмента за

период с 1816 до 2019 года) и на французском языке (500 фрагментов за период с 1823 года до 2019 года). В каждом из корпусов были представлены произведения 20 авторов. Размеры фрагментов варьируются от 15,000 до 20,000 слов.

4.2. Постановка экспериментов

Расчёт стилометрических характеристик для последующей классификации текстов выполнялся автоматически. Порядок проведения экспериментов был следующим. С помощью алгоритмов, разработанных ранее в рамках проекта ProseRhythmDetector [22], для текстов были подсчитаны стилометрические характеристики разного уровня, а затем записаны в csv-таблицы.

Csv-таблицы были скомбинированны между собой различными способами:

- Характеристики уровня символов;
- Характеристики уровня слов;
- Характеристики структурного уровня;
- Характеристики уровня символов и слов;
- Характеристики уровня символов и структурного уровня;
- Характеристики уровня слов и структурного уровня;
- Характеристики уровня символов, слов и структурного уровня.

С помощью алгоритма, разработанного ранее в рамках проекта ProseRhythmDetector, была проведена классификация текстов по авторам. Результаты экспериментов были представлены в виде таблиц.

4.3. Результаты экспериментов

Для обозначения стилометрических характеристик различного уровня в таблицах с результатами были приняты следующие условные обозначения: характеристики уровня символов – Ch, характеристики уровня слов – W, характеристики структурного уровня – St.

Для выявления наилучшего классификатора было проведено сравнение результатов, показанных всеми вышеописанными классификаторами на основе совокупности характеристик уровня символов, слов и структуры. Сравнение осуществлялось по показателям точности, полноты и F-меры.

Точность — это число правильно положительно классифицированных текстов, поделенное на число всех положительно классифицированных текстов:

$$P = \frac{TP}{TP + FP}.$$

Полнотой называют число правильно классифицированных текстов, поделенное на число всех подходящих текстов:

$$R = \frac{TP}{TP + FN}.$$

F-мера — это среднее гармоническое точности и полноты:

$$F = \frac{2PR}{P + R}.$$

Сравнение проводилось по усреднённым показателям. Ниже представлены получившиеся таблицы.

На основе таблиц 1, 2, 3 можно сделать вывод, что лучшие результаты показал классификатор AdaBoost с наивысшими значениями по всем показателям для всех исследуемых языков. Исходя из этого он был выбран для получения более развёрнутых результатов.

Рассмотрим подробнее результаты, полученные после работы классификатора AdaBoost. В таблицах 4, 5, 6 представлены результаты работы классификатора на основе различных комбинаций

Table 1. Russian (Ch + W + St, average metrics)

Таблица 1. Русский язык (Ch + W + St, средние метрики)

Классификатор	Ср. точность	Ср. полнота	Ср. F-мера
AdaBoost	**97.2**	**90.4**	**93.7**
DecisionTree	78.9	77.8	78.4
GaussianNB	64.2	77.7	70.3
RandomForest	72.4	60.0	65.6
SVM	94.4	80.4	86.9

Table 2. English (Ch + W + St, average metrics)

Таблица 2. Английский язык (Ch + W + St, средние метрики)

Классификатор	Ср. точность	Ср. полнота	Ср. F-мера
AdaBoost	**97.2**	**87.6**	**92.2**
DecisionTree	79.8	79.8	79.8
GaussianNB	64.1	79.5	71.0
RandomForest	68.5	58.9	63.3
SVM	94.1	81.8	87.5

Table 3. French (Ch + W + St, average metrics)

Таблица 3. Французский язык (Ch + W + St, средние метрики)

Классификатор	Ср. точность	Ср. полнота	Ср. F-мера
AdaBoost	**98.0**	**88.6**	**93.1**
DecisionTree	78.7	77.8	78.2
GaussianNB	67.1	80.6	73.2
RandomForest	70.4	61.4	65.6
SVM	96.7	85.5	90.8

характеристик уровня символов, слов и структуры, описанных в предыдущем разделе. В таблицах отражены средние показатели по каждому из трёх параметров.

Первое, на что стоит обратить внимание, это высокие (больше 90) значения точности для каждого набора характеристик во всех трёх таблицах. Единственным исключением из этого правила является значение уровня символов, рассчитанное на основе корпуса английских текстов. Здесь значение уровня символов равняется 85,3. Средние значения полноты для всех трёх языков находятся в пределах от 80,0 до 90,0 за редким исключением.

Наиболее важным показателем является F-мера, демонстрирующая баланс точности и полноты. Этот показатель во всех трёх таблицах имеет значение не менее 79,5, что уверенно можно назвать хорошим результатом.

Анализируя таблицы можно сделать вывод, что наиболее высокое значение F-меры достигается при соединении характеристик на всех трёх уровнях. В случае с русским и французским языками уровни символов, слов и структуры по отдельности вносят примерно одинаковый вклад в общий результат, однако английский язык не подчиняется данной тенденции: основной вклад в итоговое значение вносит уровень слов, а наименьший (с разницей практически в 10 пунктов) уровень символов.

Результаты, полученные при классификации на основе попарного соединения характеристик уровня символов, слов и структуры, также не позволяют прийти к однозначному заключению сразу для трёх языков, касаемо вопроса о наиболее эффективной комбинации характеристик различного уровня. Для русского и французского языков наилучшее значение было получено при соединении символьного и структурного уровня (92,0 и 92,2 соответственно), однако для английского языка лучший показатель F-меры был достигнут при соединении структурного уровня с уровнем слов (91,7).

Однако стоит отметить, что для всех трёх языков различные попарные комбинации характеристик дали примерно одинаковые результаты.

Table 4. Russian (AdaBoost, all metrics)

Таблица 4. Русский язык (AdaBoost, все метрики)

Уровни	Ср. точность	Ср. полнота	Ср. F-мера
Ch	90.4	83.4	86.8
W	94.6	83.0	88.4
St	93.2	82.4	87.5
Ch + St	96.0	88.3	92.0
Ch + W	95.4	86.8	90.9
W + St	96.2	86.5	91.1
Ch + W + St	97.2	90.4	93.7

Table 5. English (AdaBoost, all metrics)

Таблица 5. Английский язык (AdaBoost, все метрики)

Уровни	Ср. точность	Ср. полнота	Ср. F-мера
Ch	85.3	74.4	79.5
W	92.7	84.0	88.1
St	90.6	81.6	85.9
Ch + St	93.1	86.7	89.8
Ch + W	94.5	85.3	89.7
W + St	95.2	88.4	91.7
Ch + W + St	97.2	87.6	92.2

Table 6. French (AdaBoost, all metrics)

Таблица 6. Французский язык (AdaBoost, все метрики)

Уровни	Ср. точность	Ср. полнота	Ср. F-мера
Ch	93.7	84.6	88.9
W	90.5	82.2	86.2
St	91.2	79.6	85.0
Ch + St	96.4	88.4	92.2
Ch + W	95.9	87.3	91.4
W + St	95.3	86.1	90.5
Ch + W + St	98.0	88.6	93.1

Следующим шагом рассмотрим расширенные таблицы, демонстрирующие все вышеописанные комбинации характеристик разного уровня на примере конкретных авторов. Помимо значений трёх основных параметров, в этих таблицах представленно и значение стандартного отклонения для каждого из них.

Таблицы 7, 8, 9, содержащие результаты исследования на материале русскоязычного корпуса, позволяют сделать вывод о том, что комбинация характеристик структурного уровня с уровнем символов или уровнем слов и позволяет получить если не самое высокое значение F-меры, то близкое к этому. Сочетание всех трёх уровней дало лучший результат для фрагментов авторства Тургенева, Алексея Толстого, Лескова, Пелевина, Горького, Пикуля, Достоевского, Рубанова и Стругацких. А при попарном сочетании характеристик структурного уровня с уровнем слов или уровнем символов, лучший результат был достигнут для фрагментов авторства Набокова, Маканина, Аксенова, Славниковой, Льва Толстого, Водолазкина, Солженицина и Гоголя.

К тому же следует обратить внимание на минимальность стандартного отклонения F-меры при комбинировании характеристик сразу трёх уровней: в большинстве случаев оно не превышает 5,0.

Table 7. Russian (AdaBoost, all metrics, all authors)

Таблица 7. Русский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	И С Тургенев	95.1	4.7	79.4	10.5	86.5	6.4
W	И С Тургенев	87.8	10.5	81.8	10.5	84.3	5.2
St	И С Тургенев	92.0	5.1	86.6	4.8	88.2	2.0
Ch + St	И С Тургенев	94.8	5.2	90.8	8.1	85.9	9.7
Ch + W	И С Тургенев	95.0	6.1	86.8	8.8	92.5	5.0
W + St	И С Тургенев	92.6	6.3	86.9	8.5	91.4	5.9
Ch + W + St	И С Тургенев	95.9	6.3	91.9	7.5	93.5	1.3
Ch	А К Толстой	91.8	8.9	83.1	13.8	84.1	3.2
W	А К Толстой	94.2	4.5	81.4	7.7	88.4	5.0
St	А К Толстой	92.2	8.7	86.1	10.9	83.6	8.0
Ch + St	А К Толстой	91.5	8.5	85.7	2.6	90.2	6.5
Ch + W	А К Толстой	95.9	4.2	84.3	11.5	91.3	5.0
W + St	А К Толстой	91.9	4.0	87.1	11.4	92.9	4.7
Ch + W + St	А К Толстой	92.5	6.0	90.7	8.3	93.5	2.3
Ch	Н С Лесков	77.4	16.8	71.6	12.9	74.8	13.2
W	Н С Лесков	89.2	12.6	78.2	8.9	79.7	3.2
St	Н С Лесков	84.6	18.3	74.5	5.7	75.7	15.2
Ch + St	Н С Лесков	91.9	9.9	76.3	7.5	75.4	5.4
Ch + W	Н С Лесков	99.3	0.3	81.0	6.7	83.9	5.2
W + St	Н С Лесков	87.6	19.2	71.9	12.7	89.7	12.6
Ch + W + St	Н С Лесков	94.4	10.2	77.5	9.0	93.0	4.5
Ch	В О Пелевин	91.5	12.8	91.0	5.2	87.8	8.5
W	В О Пелевин	96.6	4.9	80.1	8.9	86.7	6.2
St	В О Пелевин	88.8	3.3	83.9	2.2	84.7	7.4
Ch + St	В О Пелевин	94.9	4.0	89.9	7.1	92.0	5.6
Ch + W	В О Пелевин	97.0	3.3	94.6	6.6	97.3	2.8
W + St	В О Пелевин	93.0	3.5	89.5	4.2	89.8	7.3
Ch + W + St	В О Пелевин	96.5	3.9	95.1	4.1	96.7	1.9
Ch	В В Набоков	91.5	6.3	81.5	7.6	85.7	4.1
W	В В Набоков	96.4	3.8	73.2	11.5	71.5	4.6
St	В В Набоков	86.1	9.6	74.8	10.0	85.9	5.5
Ch + St	В В Набоков	95.4	4.8	76.3	14.5	90.3	6.5
Ch + W	В В Набоков	91.9	7.0	82.2	6.5	89.1	5.8
W + St	В В Набоков	93.5	7.3	72.4	5.5	82.1	13.1
Ch + W + St	В В Набоков	93.3	7.4	89.5	7.0	88.0	7.5
Ch	М Горький	95.9	6.7	87.8	5.5	90.9	6.1
W	М Горький	95.2	5.9	80.9	8.7	89.8	4.3
St	М Горький	99.3	0.4	84.5	5.9	94.3	4.3
Ch + St	М Горький	98.2	2.8	93.7	5.3	92.0	2.9
Ch + W	М Горький	99.8	0.2	91.8	8.5	94.3	4.6
W + St	М Горький	99.1	0.9	88.3	6.5	93.5	1.7
Ch + W + St	М Горький	99.6	0.4	91.9	5.4	96.3	2.5
Ch	М А Булгаков	86.0	11.5	75.9	14.8	69.0	13.8
W	М А Булгаков	96.9	4.9	78.0	17.4	81.2	10.7
St	М А Булгаков	82.1	20.9	72.6	17.3	71.8	7.0
Ch + St	М А Булгаков	86.7	8.6	83.0	11.0	74.3	15.2
Ch + W	М А Булгаков	96.0	6.5	81.4	17.6	88.1	10.5
W + St	М А Булгаков	99.2	0.3	78.0	12.9	77.9	15.1
Ch + W + St	М А Булгаков	99.4	0.3	82.6	9.5	87.7	11.7

Рассмотрим таблицы 10, 11, 12, в которых отражены результаты исследования, полученные на материале англоязычного корпуса. Анализ полученных значений F-меры для каждой из возмож-

Table 8. Russian (AdaBoost, all metrics, all authors)

Таблица 8. Русский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	В С Пикуль	94.3	4.2	88.9	7.0	88.8	7.4
W	В С Пикуль	96.9	5.1	87.1	11.4	92.7	4.7
St	В С Пикуль	99.5	0.4	93.7	5.8	96.2	4.7
Ch + St	В С Пикуль	98.2	2.9	94.3	5.1	94.9	2.9
Ch + W	В С Пикуль	97.0	3.3	90.5	5.8	95.2	4.1
W + St	В С Пикуль	98.6	2.0	89.5	5.8	96.5	2.4
Ch + W + St	В С Пикуль	98.6	1.7	92.7	9.8	97.5	3.0
Ch	Д И Рубина	94.9	3.2	81.8	10.1	85.0	7.4
W	Д И Рубина	93.4	7.0	86.7	6.9	89.1	1.4
St	Д И Рубина	93.9	6.7	72.0	7.3	82.0	5.8
Ch + St	Д И Рубина	96.1	3.8	81.2	4.3	85.0	9.3
Ch + W	Д И Рубина	95.7	5.3	91.5	5.2	94.2	3.0
W + St	Д И Рубина	99.0	0.4	89.0	4.7	86.2	7.1
Ch + W + St	Д И Рубина	97.8	2.7	87.6	4.9	92.9	4.2
Ch	В С Маканин	96.9	4.7	90.8	9.4	91.9	4.7
W	В С Маканин	97.9	3.2	88.4	4.3	95.9	5.1
St	В С Маканин	99.4	0.3	88.1	7.3	88.2	6.1
Ch + St	В С Маканин	98.6	2.1	95.0	4.5	97.6	4.8
Ch + W	В С Маканин	99.6	0.3	91.8	4.1	95.7	4.7
W + St	В С Маканин	98.2	3.3	94.3	5.3	95.9	3.7
Ch + W + St	В С Маканин	99.6	0.4	94.3	6.8	96.4	3.3
Ch	В П Аксенов	82.3	14.0	73.9	9.7	71.9	8.0
W	В П Аксенов	88.2	20.3	73.7	11.1	74.4	6.4
St	В П Аксенов	85.6	13.6	68.0	7.8	78.3	8.2
Ch + St	В П Аксенов	95.5	6.3	79.3	7.7	82.4	7.4
Ch + W	В П Аксенов	84.0	17.9	76.0	15.4	83.3	3.1
W + St	В П Аксенов	98.8	0.7	79.1	8.5	87.4	6.6
Ch + W + St	В П Аксенов	99.0	0.6	88.6	8.2	85.4	4.8
Ch	Ф М Достоевский	86.7	7.0	90.0	7.4	86.4	6.8
W	Ф М Достоевский	88.8	7.9	77.7	3.7	89.0	4.3
St	Ф М Достоевский	94.6	3.2	94.5	3.1	93.2	2.4
Ch + St	Ф М Достоевский	94.9	4.6	95.7	4.6	94.1	6.0
Ch + W	Ф М Достоевский	88.8	7.9	77.7	3.7	89.0	4.3
W + St	Ф М Достоевский	95.3	3.9	94.5	1.7	93.9	5.0
Ch + W + St	Ф М Достоевский	98.5	2.2	97.7	3.0	95.6	1.6
Ch	А В Рубанов	97.4	2.7	92.3	4.4	95.5	4.6
W	А В Рубанов	98.1	3.0	93.5	5.6	91.2	5.5
St	А В Рубанов	97.1	3.8	88.8	3.0	93.9	3.7
Ch + St	А В Рубанов	99.8	0.4	91.3	1.6	95.5	5.0
Ch + W	А В Рубанов	98.1	2.6	93.2	9.3	96.4	3.5
W + St	А В Рубанов	98.1	3.0	93.5	5.6	91.2	5.5
Ch + W + St	А В Рубанов	99.8	0.3	95.8	3.6	97.2	3.5

ных комбинаций стилометрических характеристик показывает, что использование характеристик структурного уровня часто позволяет улучшить результат. Это видно на примере сочинений Генти (G A Henty), Моэма (W S Maugham), Честертона (G K Chesterton), Мойес (J Moyes), Элиот (G Eliot), Коллинза (W Collins), Троллопа (A Trollope), Лэнга (A Lang), Пратчетта (T Pratchett), Смит (Z Smith), Геймана (N Gaiman), Джеймса (H James), Харди (T Hardy), Роулинг (J K Rowling) и Макьюэна (I McEwan).

Сочетание характеристик всех трёх уровней тоже позволяет добиться улучшения результата верификации, хотя это прослеживается не так явно, как для русскоязычных авторов. Для текстов англоязычного корпуса наилучшие результаты при комбинации характеристик всех трёх уровней удалось получить для текстов 6 авторов: Генти, Моэма, Элиота, Смит, Харди и Макьюэна. К тому же,

Table 9. Russian (AdaBoost, all metrics, all authors)

Таблица 9. Русский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	О А Славникова	98.4	2.9	93.1	5.5	96.3	4.4
W	О А Славникова	99.6	0.2	88.7	11.3	97.3	2.5
St	О А Славникова	99.7	0.4	96.7	4.4	98.8	1.5
Ch + St	О А Славникова	99.9	0.1	96.6	4.3	98.4	2.0
Ch + W	О А Славникова	99.4	0.3	94.7	6.9	95.9	3.8
W + St	О А Славникова	99.6	0.2	88.7	11.3	97.3	2.5
Ch + W + St	О А Славникова	99.9	0.2	97.3	3.3	98.0	2.5
Ch	Стругацкие	83.6	3.9	81.4	13.3	86.5	7.9
W	Стругацкие	91.7	3.9	82.8	3.4	87.3	6.1
St	Стругацкие	91.9	8.4	68.5	7.5	73.4	12.7
Ch + St	Стругацкие	95.0	5.4	87.6	8.4	87.0	5.7
Ch + W	Стругацкие	95.6	5.2	88.5	9.5	88.3	7.9
W + St	Стругацкие	91.7	3.9	82.8	3.4	87.3	6.1
Ch + W + St	Стругацкие	95.2	3.8	92.4	4.1	89.5	5.8
Ch	Л Н Толстой	71.7	15.7	63.8	13.0	66.6	10.0
W	Л Н Толстой	87.3	11.0	73.2	8.6	83.1	13.6
St	Л Н Толстой	94.0	6.5	71.1	4.8	83.8	10.5
Ch + St	Л Н Толстой	94.4	9.9	77.6	7.4	87.9	5.0
Ch + W	Л Н Толстой	83.6	18.7	74.4	8.9	68.5	12.7
W + St	Л Н Толстой	87.3	11.0	73.2	8.6	83.1	13.6
Ch + W + St	Л Н Толстой	88.9	20.2	75.0	13.9	87.5	4.4
Ch	Е Г Водолазкин	96.9	4.9	87.2	5.1	92.5	1.8
W	Е Г Водолазкин	99.7	0.4	94.1	3.0	90.0	7.1
St	Е Г Водолазкин	94.0	5.3	90.6	7.2	91.8	5.7
Ch + St	Е Г Водолазкин	99.7	0.1	100.0	0.0	97.5	2.2
Ch + W	Е Г Водолазкин	99.7	0.3	87.0	10.5	96.5	2.0
W + St	Е Г Водолазкин	99.7	0.4	94.1	3.0	90.0	7.1
Ch + W + St	Е Г Водолазкин	99.9	0.2	97.1	5.7	96.8	3.9
Ch	А А Проханов	95.4	4.8	96.8	3.8	95.2	4.9
W	А А Проханов	96.4	3.2	98.3	3.3	97.2	1.6
St	А А Проханов	96.4	4.8	93.0	4.7	95.2	5.9
Ch + St	А А Проханов	97.2	3.1	98.7	2.5	96.1	4.7
Ch + W	А А Проханов	99.9	0.2	97.5	3.2	100.0	0.0
W + St	А А Проханов	96.4	3.2	98.3	3.3	97.2	1.6
Ch + W + St	А А Проханов	99.7	0.3	98.1	3.2	97.9	2.7
Ch	А И Солженицын	91.4	6.6	86.0	9.1	83.5	4.0
W	А И Солженицын	98.3	2.1	90.9	5.3	94.6	3.0
St	А И Солженицын	93.2	4.9	85.2	9.4	90.7	4.4
Ch + St	А И Солженицын	98.0	2.9	89.8	5.8	92.1	4.0
Ch + W	А И Солженицын	94.2	4.6	91.7	5.6	93.4	4.2
W + St	А И Солженицын	98.3	2.1	90.9	5.3	94.6	3.0
Ch + W + St	А И Солженицын	96.6	3.4	87.7	4.6	91.8	3.8
Ch	Н В Гоголь	89.1	9.7	72.4	5.0	78.4	2.7
W	Н В Гоголь	99.4	0.5	71.7	12.5	76.6	18.0
St	Н В Гоголь	99.4	0.4	75.5	12.8	87.9	6.4
Ch + St	Н В Гоголь	99.5	0.7	83.7	10.7	88.3	6.9
Ch + W	Н В Гоголь	97.1	5.2	76.4	12.6	85.8	11.4
W + St	Н В Гоголь	99.4	0.5	71.7	12.5	76.6	18.0
Ch + W + St	Н В Гоголь	99.4	0.3	84.7	13.8	81.6	17.2

стандартное отклонение F-меры для комбинации характеристик всех трёх уровней в большинстве случаев не превышает 5,0.

Таблицы 13, 14, 15 демонстрируют результаты экспериментов, проведённых с корпусами французских текстов. В результате анализа этих таблиц удалось определить, что стилометрические

Table 10. English (AdaBoost, all metrics, all authors)

Таблица 10. Английский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	G A Henty	93.9	4.4	78.2	5.5	82.1	5.0
W	G A Henty	94.9	5.1	81.4	10.8	84.3	10.0
St	G A Henty	81.3	17.3	88.0	11.9	86.4	5.5
Ch + St	G A Henty	89.8	8.7	87.8	6.7	88.4	7.4
Ch + W	G A Henty	92.7	6.3	83.2	13.3	86.4	3.3
W + St	G A Henty	93.3	7.0	82.0	8.3	88.4	5.9
Ch + W + St	G A Henty	95.9	3.8	94.0	5.5	88.7	4.4
Ch	W S Maugham	87.7	8.2	78.2	9.0	77.4	6.4
W	W S Maugham	92.1	5.8	90.8	10.8	88.2	8.7
St	W S Maugham	92.8	3.0	86.3	6.1	82.1	9.3
Ch + St	W S Maugham	93.1	5.2	86.2	7.5	82.0	8.8
Ch + W	W S Maugham	96.0	4.1	86.5	9.2	84.9	5.6
W + St	W S Maugham	97.0	3.3	87.7	7.9	92.4	5.4
Ch + W + St	W S Maugham	96.5	3.4	86.6	8.4	93.1	4.3
Ch	G K Chesterton	83.7	11.8	62.2	4.2	68.6	6.5
W	G K Chesterton	98.4	2.3	85.1	5.3	93.8	3.8
St	G K Chesterton	97.9	2.9	83.7	12.2	92.8	2.4
Ch + St	G K Chesterton	92.6	9.8	91.1	8.6	84.3	8.8
Ch + W	G K Chesterton	97.0	3.2	90.7	5.8	90.6	6.8
W + St	G K Chesterton	99.5	0.3	89.4	6.5	96.2	3.2
Ch + W + St	G K Chesterton	97.9	3.0	89.4	9.4	93.1	4.6
Ch	J Moyes	74.4	16.2	69.3	9.8	65.9	12.3
W	J Moyes	83.4	18.0	83.9	5.2	84.2	10.2
St	J Moyes	97.2	4.2	81.3	7.0	82.3	8.8
Ch + St	J Moyes	97.1	3.8	84.0	9.1	89.5	6.6
Ch + W	J Moyes	96.7	5.2	81.9	5.4	87.7	5.0
W + St	J Moyes	97.1	4.4	85.5	6.5	92.2	6.6
Ch + W + St	J Moyes	97.6	3.4	87.0	7.9	91.7	4.8
Ch	G Eliot	66.6	16.5	60.6	8.1	64.6	8.6
W	G Eliot	88.7	12.2	73.8	11.4	85.3	6.7
St	G Eliot	92.0	9.7	68.4	9.7	81.5	6.4
Ch + St	G Eliot	91.4	9.9	80.2	11.1	89.6	4.2
Ch + W	G Eliot	99.1	0.6	76.5	9.8	81.8	7.2
W + St	G Eliot	96.4	6.4	85.2	11.4	78.7	6.2
Ch + W + St	G Eliot	94.0	7.4	85.9	6.1	92.8	7.3
Ch	W Collins	91.6	6.2	75.2	5.9	84.3	1.6
W	W Collins	98.2	2.1	86.6	3.8	92.4	6.3
St	W Collins	93.3	4.9	82.3	8.0	93.0	6.4
Ch + St	W Collins	96.3	5.9	87.6	6.8	88.1	9.4
Ch + W	W Collins	97.7	3.2	84.2	18.4	91.5	3.6
W + St	W Collins	98.2	3.1	92.0	7.0	92.0	5.0
Ch + W + St	W Collins	97.3	4.3	89.1	4.2	92.2	6.8
Ch	A Trollope	96.7	3.1	83.2	6.1	95.1	3.0
W	A Trollope	96.9	4.7	91.8	5.2	96.4	1.8
St	A Trollope	97.7	3.8	93.3	6.2	95.3	2.9
Ch + St	A Trollope	96.7	3.6	94.2	6.0	95.7	2.3
Ch + W	A Trollope	98.3	2.4	94.6	5.0	94.7	5.4
W + St	A Trollope	99.8	0.2	94.5	2.9	97.3	1.6
Ch + W + St	A Trollope	98.4	2.2	90.9	3.2	96.6	1.9

характеристики структурного уровня вносят значительный вклад в улучшение качества верификации авторства. Это подтверждают результаты, полученные на основе произведений Франса (A France), Гара (R Gard), Колетт (Colette), Панколь (K Pancol), Мопассана (G Maupassant), Золя (É Zola), Гюго (V Hugo), Роллана (R Rolland), Жида (A Gide), Леви (M Levy), Бальзака (H Balzac)

Table 11. English (AdaBoost, all metrics, all authors)

Таблица 11. Английский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	H Bindloss	90.4	6.6	78.9	11.6	84.6	8.8
W	H Bindloss	98.5	1.7	88.1	3.5	89.1	3.6
St	H Bindloss	97.1	3.6	91.4	7.9	88.5	3.7
Ch + St	H Bindloss	99.7	0.3	91.6	2.7	93.9	4.0
Ch + W	H Bindloss	98.2	2.3	92.5	6.3	96.6	3.2
W + St	H Bindloss	97.8	2.9	91.0	6.2	92.9	4.7
Ch + W + St	H Bindloss	99.5	0.5	94.9	4.5	95.2	4.2
Ch	K Atkinson	94.3	6.8	86.1	10.1	89.5	9.2
W	K Atkinson	99.3	0.7	93.8	5.2	98.5	3.0
St	K Atkinson	96.9	5.7	80.0	17.5	86.4	12.0
Ch + St	K Atkinson	99.5	0.4	89.1	8.1	95.9	3.4
Ch + W	K Atkinson	99.9	0.2	98.8	2.5	97.0	2.5
W + St	K Atkinson	99.8	0.2	94.0	8.0	95.6	4.0
Ch + W + St	K Atkinson	99.9	0.2	95.1	6.1	93.4	7.0
Ch	Sir W Scott	92.7	6.7	81.6	8.6	89.8	6.2
W	Sir W Scott	93.2	9.8	88.3	8.3	91.3	7.3
St	Sir W Scott	95.4	7.9	87.5	7.3	84.6	7.3
Ch + St	Sir W Scott	93.1	6.0	89.3	9.7	94.5	4.7
Ch + W	Sir W Scott	97.7	4.1	92.6	4.7	95.7	3.9
W + St	Sir W Scott	97.8	3.3	97.0	3.6	95.2	4.2
Ch + W + St	Sir W Scott	97.1	4.8	91.7	5.8	94.1	3.4
Ch	C Kingsley	87.5	8.4	77.2	7.9	79.7	7.9
W	C Kingsley	93.0	8.2	91.0	7.4	94.9	3.1
St	C Kingsley	91.1	4.9	83.4	6.0	84.4	8.0
Ch + St	C Kingsley	96.6	3.6	90.3	5.9	92.2	3.9
Ch + W	C Kingsley	98.2	3.1	87.3	9.1	93.7	8.8
W + St	C Kingsley	97.9	3.3	97.1	5.9	91.1	8.9
Ch + W + St	C Kingsley	99.5	0.5	94.3	5.0	93.7	4.9
Ch	A Lang	71.6	20.3	73.3	12.4	78.5	13.3
W	A Lang	87.8	9.7	71.8	3.5	78.4	6.5
St	A Lang	90.2	7.5	72.4	6.8	84.5	8.1
Ch + St	A Lang	96.8	4.0	76.6	8.4	89.9	5.2
Ch + W	A Lang	85.9	8.3	74.8	6.2	79.6	9.4
W + St	A Lang	89.0	21.0	83.3	10.9	86.7	4.3
Ch + W + St	A Lang	99.1	0.4	79.3	8.6	84.1	10.1
Ch	T Parsons	90.0	12.2	85.9	6.5	87.6	7.3
W	T Parsons	99.1	0.3	84.2	12.1	92.3	8.1
St	T Parsons	95.1	5.2	72.7	18.6	85.1	6.2
Ch + St	T Parsons	84.5	18.6	97.2	3.5	94.4	4.9
Ch + W	T Parsons	99.6	0.2	96.0	5.0	97.3	3.3
W + St	T Parsons	99.5	0.2	88.6	9.1	92.7	2.0
Ch + W + St	T Parsons	99.6	0.5	89.1	9.6	92.2	11.6
Ch	T Pratchett	92.0	6.6	88.2	7.4	86.4	9.3
W	T Pratchett	98.9	1.9	97.3	3.1	95.1	6.0
St	T Pratchett	100.0	0.0	93.0	5.6	98.7	1.7
Ch + St	T Pratchett	99.9	0.2	99.9	0.2	96.1	5.3
Ch + W	T Pratchett	98.8	1.9	94.8	6.7	95.3	2.0
W + St	T Pratchett	97.7	3.1	99.1	1.6	97.3	1.6
Ch + W + St	T Pratchett	97.4	3.3	99.1	1.6	97.7	2.9

и Пруста (M Proust). Комбинация сразу трёх уровней стилометрических характеристик помогла получить лучшие результаты для 3 авторов: Франса, Золя, Роллана и Леви, что меньше, чем для русскоязычных и англоязычных писателей. Среднее отклонение для значений F-меры по корпусу

Table 12. English (AdaBoost, all metrics, all authors)

Таблица 12. Английский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	Z Smith	83.3	18.1	64.0	9.9	75.5	8.2
W	Z Smith	95.7	6.8	79.9	13.6	89.8	7.4
St	Z Smith	75.7	22.4	76.6	12.5	75.0	8.4
Ch + St	Z Smith	99.3	0.7	81.5	12.1	87.4	8.5
Ch + W	Z Smith	93.6	7.0	84.4	14.1	90.8	5.3
W + St	Z Smith	95.5	7.8	85.0	9.4	85.9	7.9
Ch + W + St	Z Smith	99.5	0.3	78.3	18.0	94.3	7.0
Ch	N Gaiman	86.2	8.6	75.1	7.7	80.3	7.0
W	N Gaiman	91.1	6.7	74.9	8.2	74.8	11.1
St	N Gaiman	87.3	6.0	79.4	8.7	81.3	6.6
Ch + St	N Gaiman	84.9	5.9	78.9	8.1	88.1	5.5
Ch + W	N Gaiman	89.6	9.3	76.2	7.9	81.1	10.2
W + St	N Gaiman	92.9	7.9	84.9	13.3	82.8	7.1
Ch + W + St	N Gaiman	92.3	5.2	79.6	8.8	75.8	14.0
Ch	H James	64.2	20.0	52.8	6.7	49.3	0.2
W	H James	64.2	20.2	73.7	14.0	74.7	13.1
St	H James	59.2	20.0	71.7	12.5	69.6	19.0
Ch + St	H James	59.2	20.2	76.6	20.7	76.7	17.4
Ch + W	H James	79.2	24.9	55.0	10.0	76.7	17.3
W + St	H James	69.2	24.7	74.4	21.3	67.0	14.4
Ch + W + St	H James	89.3	19.9	66.6	18.3	70.3	11.2
Ch	T Hardy	88.4	10.1	74.5	3.7	81.1	5.9
W	T Hardy	91.5	9.7	82.2	5.1	81.2	11.1
St	T Hardy	86.7	5.0	84.4	7.4	83.9	8.7
Ch + St	T Hardy	97.7	2.6	82.8	7.0	87.8	5.8
Ch + W	T Hardy	83.1	17.6	82.1	4.2	87.6	6.0
W + St	T Hardy	95.7	4.6	83.8	12.7	80.1	5.6
Ch + W + St	T Hardy	95.4	4.5	81.6	9.7	90.0	3.5
Ch	J K Rowling	90.2	5.2	65.7	9.8	70.8	12.9
W	J K Rowling	93.7	9.8	78.3	9.9	78.3	9.8
St	J K Rowling	90.7	7.5	80.1	11.6	91.4	3.1
Ch + St	J K Rowling	99.3	1.0	89.5	3.3	91.8	5.2
Ch + W	J K Rowling	92.3	8.0	82.9	7.1	86.3	7.4
W + St	J K Rowling	96.6	5.4	83.7	5.9	89.6	8.6
Ch + W + St	J K Rowling	99.4	0.5	87.8	4.2	91.2	3.4
Ch	I McEwan	81.4	5.7	78.5	8.4	79.0	7.7
W	I McEwan	95.4	6.5	82.9	10.7	85.6	5.1
St	I McEwan	94.7	5.2	76.9	10.8	83.6	6.4
Ch + St	I McEwan	94.2	5.6	79.5	4.8	86.7	6.3
Ch + W	I McEwan	96.4	3.1	90.8	3.2	88.8	8.9
W + St	I McEwan	94.1	10.5	90.7	6.1	89.0	6.8
Ch + W + St	I McEwan	99.3	1.0	90.8	6.2	91.2	3.9

франкоязычных текстов оказалось несколько выше, чем для текстов, рассмотренных ранее: здесь в большинстве случаев оно попадало в промежуток от 5,0 до 10,0.

5. Заключение

На основе проведённого исследования можно сделать вывод о высокой значимости стилометрических характеристик структурного уровня в решении задачи верификации авторства. Соединение характеристик структурного уровня с характеристиками уровня слов и/или символов позволило получить наиболее точные результаты во время экспериментов на корпусе русскоязычных (85,0 %), англоязычных (75,0 %) и франкоязычных (60,0 %) художественных текстов (значения в процентах рассчитано как отношение количества авторов, для которых лучший результат удалось получить

Table 13. French (AdaBoost, all metrics, all authors)

Таблица 13. Французский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	G Flaubert	99.1	0.7	89.7	6.7	92.5	5.5
W	G Flaubert	97.9	4.0	97.9	4.0	95.5	4.1
St	G Flaubert	98.0	3.4	95.5	5.9	92.7	4.7
Ch + St	G Flaubert	99.7	0.3	90.7	8.7	92.0	8.7
Ch + W	G Flaubert	96.6	4.2	95.1	6.5	98.3	2.3
W + St	G Flaubert	99.9	0.2	98.0	4.0	95.2	4.4
Ch + W + St	G Flaubert	99.9	0.2	96.6	6.6	97.9	2.6
Ch	A France	97.1	4.4	77.5	14.0	83.3	10.4
W	A France	95.7	7.0	89.9	8.7	83.5	9.6
St	A France	95.4	5.3	81.8	9.7	87.2	7.6
Ch + St	A France	99.8	0.3	91.4	8.0	94.7	3.4
Ch + W	A France	97.4	3.7	89.5	6.6	89.7	8.3
W + St	A France	99.3	0.5	95.2	6.0	91.5	7.3
Ch + W + St	A France	99.4	0.6	88.7	11.5	96.0	3.7
Ch	F Cusset	83.9	19.7	76.8	13.0	73.4	16.2
W	F Cusset	89.0	19.7	67.7	11.0	81.4	8.1
St	F Cusset	83.7	20.5	68.8	5.4	72.6	3.5
Ch + St	F Cusset	96.5	4.5	73.3	16.5	84.7	7.6
Ch + W	F Cusset	76.9	23.7	79.1	12.9	88.1	8.2
W + St	F Cusset	84.8	19.3	75.8	15.6	77.9	10.1
Ch + W + St	F Cusset	94.0	7.1	83.7	10.7	85.3	9.3
Ch	F Beigbeder	92.3	9.9	92.3	6.8	94.9	2.8
W	F Beigbeder	97.0	4.3	87.1	6.6	75.7	14.5
St	F Beigbeder	97.3	3.6	82.5	9.6	85.1	10.0
Ch + St	F Beigbeder	99.6	0.5	90.4	7.7	92.1	4.1
Ch + W	F Beigbeder	99.7	0.4	94.2	7.3	97.7	4.6
W + St	F Beigbeder	97.4	4.0	84.6	12.7	92.2	5.7
Ch + W + St	F Beigbeder	97.6	4.1	90.6	8.1	95.3	4.4
Ch	R Gard	94.2	6.6	85.4	8.4	86.7	8.8
W	R Gard	75.3	9.8	77.4	9.2	78.0	4.9
St	R Gard	93.3	6.9	76.4	3.2	85.1	5.4
Ch + St	R Gard	99.0	0.9	88.8	9.5	90.4	2.8
Ch + W	R Gard	97.0	3.6	83.6	7.4	86.5	6.8
W + St	R Gard	96.6	4.4	82.1	12.7	85.6	7.3
Ch + W + St	R Gard	93.3	5.0	80.3	9.0	85.5	5.6
Ch	Colette	99.6	0.4	92.1	6.6	97.2	4.0
W	Colette	87.5	13.0	76.0	13.0	79.7	11.5
St	Colette	91.0	12.7	89.9	5.1	76.5	11.2
Ch + St	Colette	99.7	0.3	94.6	4.9	99.2	1.6
Ch + W	Colette	97.1	5.1	92.1	7.0	91.0	5.6
W + St	Colette	99.3	0.5	80.4	8.3	94.3	5.1
Ch + W + St	Colette	99.4	0.6	91.6	7.1	91.0	7.5
Ch	K Pancol	93.5	7.2	93.6	4.3	88.7	6.3
W	K Pancol	98.0	2.4	81.5	7.7	87.3	6.0
St	K Pancol	92.2	4.4	85.2	3.1	86.8	10.7
Ch + St	K Pancol	95.7	7.8	94.7	4.8	97.1	2.4
Ch + W	K Pancol	99.3	0.4	91.1	2.5	95.0	6.4
W + St	K Pancol	94.3	5.0	88.4	9.7	91.9	6.6
Ch + W + St	K Pancol	99.8	0.3	96.1	5.1	93.8	2.7

путём соединения структурных характеристик с характеристиками уровня символов и/или слов к общему числу авторов в корпусе). Кроме того, значения среднего отклонения для параметра F-меры в большинстве случаев не превышает отметку 5.0 для русских и английских текстов и 10.0 для французских. Таким образом, полученные результаты позволяют выдвинуть гипотезу о разной

Table 14. French (AdaBoost, all metrics, all authors)

Таблица 14. Французский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	A Nothomb	97.4	4.0	88.4	10.3	82.1	17.4
W	A Nothomb	91.3	10.2	90.2	14.3	95.0	6.9
St	A Nothomb	92.0	10.0	85.5	14.9	80.3	10.7
Ch + St	A Nothomb	96.2	6.7	80.0	12.3	90.0	11.2
Ch + W	A Nothomb	99.7	0.4	84.1	9.1	96.8	4.2
W + St	A Nothomb	96.1	6.9	89.1	9.6	92.0	7.8
Ch + W + St	A Nothomb	99.8	0.3	90.0	8.2	86.4	10.9
Ch	G Maupassant	89.7	12.7	77.4	10.2	83.5	7.8
W	G Maupassant	92.8	7.8	79.0	13.8	76.4	9.6
St	G Maupassant	83.4	21.1	74.7	7.4	79.3	8.2
Ch + St	G Maupassant	95.4	6.4	80.3	10.2	81.6	17.2
Ch + W	G Maupassant	78.9	25.0	78.3	12.5	83.7	17.8
W + St	G Maupassant	86.8	11.7	79.8	4.3	90.3	6.0
Ch + W + St	G Maupassant	99.1	0.5	84.6	15.6	84.1	9.5
Ch	É Zola	99.5	0.3	83.0	12.0	87.2	19.2
W	É Zola	84.1	20.1	77.8	13.6	77.1	7.4
St	É Zola	94.0	9.8	81.4	11.0	87.1	13.7
Ch + St	É Zola	99.7	0.3	91.7	10.5	91.8	8.5
Ch + W	É Zola	99.5	0.6	81.2	17.6	84.5	18.4
W + St	É Zola	99.2	0.4	75.7	10.1	87.9	4.3
Ch + W + St	É Zola	98.0	3.2	89.8	19.9	92.2	5.6
Ch	J G Verne	99.5	0.5	90.3	9.3	89.0	19.7
W	J G Verne	97.2	3.9	93.2	6.0	89.5	7.6
St	J G Verne	92.1	8.0	79.1	5.0	81.3	15.0
Ch + St	J G Verne	99.7	0.3	96.0	8.0	96.5	7.0
Ch + W	J G Verne	99.4	0.8	93.5	5.4	97.0	4.1
W + St	J G Verne	99.0	0.3	81.3	12.0	95.8	5.2
Ch + W + St	J G Verne	99.5	0.8	91.6	5.5	96.9	4.0
Ch	J P Modiano	85.6	20.5	74.9	13.4	91.2	4.5
W	J P Modiano	99.7	0.4	87.3	12.6	94.9	4.6
St	J P Modiano	89.2	19.9	83.3	11.8	79.0	17.4
Ch + St	J P Modiano	99.5	0.6	87.4	12.4	83.7	19.4
Ch + W	J P Modiano	99.6	0.6	90.8	7.6	93.6	6.3
W + St	J P Modiano	99.6	0.4	85.2	9.9	91.0	10.7
Ch + W + St	J P Modiano	99.6	0.2	94.7	6.9	91.9	9.5
Ch	V Hugo	86.7	9.6	66.6	12.2	83.2	10.9
W	V Hugo	88.6	9.6	69.9	10.7	72.8	15.1
St	V Hugo	91.8	8.2	66.3	16.4	74.5	14.3
Ch + St	V Hugo	75.1	22.9	86.5	17.7	74.6	14.3
Ch + W	V Hugo	96.6	4.8	64.3	15.2	74.1	10.7
W + St	V Hugo	98.7	0.6	83.7	5.5	84.8	7.1
Ch + W + St	V Hugo	95.6	6.7	74.9	3.2	77.1	10.8
Ch	G Musso	99.6	0.4	96.1	5.1	93.6	6.6
W	G Musso	78.9	24.4	78.2	14.5	81.5	9.6
St	G Musso	75.7	23.0	66.5	13.9	87.2	7.5
Ch + St	G Musso	99.7	0.3	86.4	18.8	91.6	7.6
Ch + W	G Musso	99.9	0.2	90.1	9.4	90.7	10.6
W + St	G Musso	86.4	20.0	94.0	8.0	88.0	7.2
Ch + W + St	G Musso	96.5	6.6	85.5	10.9	91.6	5.6

степени влияния стилометрических характеристик на качество верификации авторства для различных языков. Это означает, что для каждого языка необходимы самостоятельные исследования для получения наиболее эффективных алгоритмов решения задач определения авторского стиля.

Table 15. French (AdaBoost, all metrics, all authors)

Таблица 15. Французский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	A St Exupery	89.3	20.4	76.6	19.9	77.3	10.3
W	A St Exupery	96.1	7.1	80.9	18.7	83.9	19.2
St	A St Exupery	87.3	14.3	64.0	18.7	75.0	16.2
Ch + St	A St Exupery	89.4	20.2	86.7	8.1	87.1	10.9
Ch + W	A St Exupery	89.7	20.1	85.0	13.3	95.4	6.7
W + St	A St Exupery	99.1	0.4	87.6	10.5	86.6	13.3
Ch + W + St	A St Exupery	99.4	0.7	84.1	12.2	86.5	19.5
Ch	R Rolland	99.4	0.6	81.0	16.8	82.5	9.4
W	R Rolland	80.4	18.3	68.1	8.1	80.8	13.5
St	R Rolland	93.5	11.0	71.2	13.8	83.0	12.4
Ch + St	R Rolland	99.2	0.4	80.0	8.3	78.1	17.8
Ch + W	R Rolland	99.0	0.6	83.8	18.3	83.3	0.7
W + St	R Rolland	86.7	19.9	82.8	12.0	84.1	12.1
Ch + W + St	R Rolland	97.3	3.7	81.3	6.3	87.0	8.2
Ch	A Gide	99.6	0.6	93.6	5.9	94.3	3.1
W	A Gide	87.1	9.4	81.2	2.7	87.0	8.9
St	A Gide	95.9	4.0	86.3	3.4	85.9	7.7
Ch + St	A Gide	95.2	5.5	96.2	4.7	97.1	3.8
Ch + W	A Gide	96.4	6.8	97.1	3.7	93.2	6.6
W + St	A Gide	89.8	9.5	87.1	6.5	82.1	7.7
Ch + W + St	A Gide	99.7	0.4	92.6	4.9	91.9	7.2
Ch	M Levy	84.2	18.5	82.5	9.3	86.5	4.6
W	M Levy	82.9	9.7	76.6	16.1	72.5	13.6
St	M Levy	84.6	18.6	73.8	12.5	83.2	6.9
Ch + St	M Levy	89.0	20.2	81.8	6.7	90.9	8.4
Ch + W	M Levy	99.4	0.6	92.6	6.2	84.7	17.9
W + St	M Levy	95.8	6.8	78.7	13.0	88.4	6.4
Ch + W + St	M Levy	96.0	7.1	80.6	9.1	91.4	5.8
Ch	H Balzac	87.1	8.7	75.6	11.8	68.6	12.3
W	H Balzac	92.4	10.0	87.8	9.2	90.2	7.0
St	H Balzac	93.6	7.1	81.5	5.7	88.9	4.9
Ch + St	H Balzac	99.5	0.6	91.2	5.3	93.2	2.1
Ch + W	H Balzac	95.6	6.9	84.8	11.0	89.7	5.9
W + St	H Balzac	99.3	0.3	94.2	5.2	95.3	4.7
Ch + W + St	H Balzac	97.5	3.5	94.3	4.8	93.4	6.0
Ch	M Proust	97.8	3.4	98.6	2.9	98.8	1.5
W	M Proust	98.9	1.5	96.1	5.2	94.9	4.9
St	M Proust	99.7	0.4	98.3	3.3	99.5	1.0
Ch + St	M Proust	100.0	0.0	100.0	0.0	95.2	5.7
Ch + W	M Proust	99.9	0.2	95.2	4.1	98.5	1.9
W + St	M Proust	98.3	3.3	99.0	2.0	100.0	0.0
Ch + W + St	M Proust	99.7	0.6	100.0	0.0	97.3	5.4

Использование высокоуровневых стилометрических характеристик открывает перед учёными широкую перспективу для исследований в области автоматической обработки текстов на естественном языке. Анализ частоты использования N-грамм POS-тегов является шагом в сторону построения структурных шаблонов, которые могут быть использованы для более детального анализа стиля того или иного автора.

References

[1] N. P. Tuchkova and O. M. Ataeva, "Podhody k izvlecheniyu znanij v nauchnyh predmetnyh oblastyah", *Informacionnye i matematicheskie tekhnologii v nauke i upravlenii*, no. 2 (18), pp. 5–18, 2020.

[2] A. Altamimi, N. Clarke, S. Furnell, and F. Li, "Multi-platform authorship verification", in *Proceedings of the Third Central European Cybersecurity Conference*, 2019, pp. 1–7.

[3] O. Halvani, L. Graner, and R. Regev, "Taveer: An interpretable topic-agnostic authorship verification method", in *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 2020, pp. 1–10.

[4] M. Kestemont, G. Martens, and T. Ries, "A computational approach to authorship verification of johann wolfgang goethe's contributions to the frankfurter gelehrte anzeigen (1772–73)", *Journal of European Periodical Studies*, vol. 4, no. 1, pp. 115–143, 2019.

[5] S. Corbara, A. Moreo, F. Sebastiani, and M. Tavoni, "The epistle to cangrande through the lens of computational authorship verification", in *International Conference on Image Analysis and Processing*, Springer, 2019, pp. 148–158.

[6] V. A. Drozdov, "Ob avtorstve poemy «'Ushshak-name» s tochki zreniya akademicheskogo vostokovedeniya i novejshih komp'yuternyh tekhnologij", *Orientalistika*, vol. 3, no. 5, pp. 1360–1378, 2020.

[7] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, and B. Stein, "Overview of the cross-domain authorship verification task at pan 2020", in *CLEF*, 2020.

[8] N. Potha and E. Stamatatos, "Intrinsic author verification using topic modeling", in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, ACM, 2018, pp. 1–7.

[9] S. Adamovic, V. Miskovic, M. Milosavljevic, M. Sarac, and M. Veinovic, "Automated language-independent authorship verification (for indo-european languages)", *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 858–871, 2019.

[10] B. Boenninghoff, S. Hessler, D. Kolossa, and R. M. Nickel, "Explainable authorship verification in social media via attention-based similarity learning", in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 36–45.

[11] N. E. Benzebouchi, N. Azizi, M. Aldwairi, and N. Farah, "Multi-classifier system for authorship verification task using word embeddings", in *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, IEEE, 2018, pp. 1–6.

[12] J. S. Li, L.-C. Chen, J. V. Monaco, P. Singh, and C. C. Tappert, "A comparison of classifiers and features for authorship authentication of social networking messages", *Concurrency and Computation: Practice and Experience*, vol. 29, no. 14, e3918, 2017.

[13] E. Tuccinardi, "An application of a profile-based method for authorship verification: Investigating the authenticity of pliny the younger's letter to trajan concerning the christians", *Digital Scholarship in the Humanities*, vol. 32, no. 2, pp. 435–447, 2017.

[14] P. B. Reddy, T. M. Mohan, P. V. K. Raja, and T. R. Reddy, "A novel approach for authorship verification", in *Data Engineering and Communication Technology*, Springer, 2020, pp. 441–448.

[15] E. Castillo, O. Cervantes, and D. Vilarino, "Authorship verification using a graph knowledge discovery approach", *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 6, pp. 6075–6087, 2019.

[16] H. Ahmed, "The role of linguistic feature categories in authorship verification", *Procedia computer science*, vol. 142, pp. 214–221, 2018.

[17] M. A. Al-Khatib and J. K. Al-qaoud, "Authorship verification of opinion articles in online newspapers using the idiolect of author: A comparative study", *Information, Communication & Society*, pp. 1–19, 2020.

[18] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, and I. Paramonov, "A survey on stylometric text features", in *Proceedings of the 25th Conference of Open Innovations Association (FRUCT)*, IEEE, 2019, pp. 184–195.

[19] Y. Polin, T. Zudilova, I. Ananchenko, and T. Vojtyuk, "Derevya reshenij v zadachah klassifikacii: osobennosti primeneniya i metody povysheniya kachestva klassifikacii", *Sovremennye naukoemkie tekhnologii*, no. 9, pp. 59–63, 2020.

[20] B. Xu, X. Guo, Y. Ye, and J. Cheng, "An improved random forest classifier for text categorization.", *JCP*, vol. 7, no. 12, pp. 2913–2920, 2012.

[21] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification", *IEEE transactions on knowledge and data engineering*, vol. 18, no. 11, pp. 1457–1466, 2006.

[22] K. Lagutina, A. Poletaev, N. Lagutina, E. Boychuk, and I. Paramonov, "Automatic extraction of rhythm figures and analysis of their dynamics in prose of 19th-21st centuries", in *Proceedings of the 26th Conference of Open Innovations Association (FRUCT)*, IEEE, 2020, pp. 247–255.

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF DATA

Text Classification by Genre Based on Rhythm Features

K. V. Lagutina[1], N. S. Lagutina[1], E. I. Boychuk[2]

DOI: 10.18255/1818-1015-2021-3-280-291

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.
[2]Yaroslavl State Pedagogical University named after K. D.Ushinsky, 108/1 Respublikanskaya str., Yaroslavl 150000, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received August 20, 2021
After revision August 30, 2021
Accepted September 1, 2021

The article is devoted to the analysis of the rhythm of texts of different genres: fiction novels, advertisements, scientific articles, reviews, tweets, and political articles. The authors identified lexico-grammatical figures in the texts: anaphora, epiphora, diacope, aposiopesis, etc., that are markers of the text rhythm. On their basis, statistical features were calculated that describe quantitatively and structurally these rhythm features.
The resulting text model was visualized for statistical analysis using boxplots and heat maps that showed differences in the rhythm of texts of different genres. The boxplots showed that almost all genres differ from each other in terms of the overall density of rhythm features. Heatmaps showed different rhythm patterns across genres. Further, the rhythm features were successfully used to classify texts into six genres. The classification was carried out in two ways: a binary classification for each genre in order to separate a particular genre from the rest genres, and a multi-class classification of the text corpus into six genres at once. Two text corpora in English and Russian were used for the experiments. Each corpus contains 100 fiction novels, scientific articles, advertisements and tweets, 50 reviews and political articles, i.e. a total of 500 texts. The high quality of the classification with neural networks showed that rhythm features are a good marker for most genres, especially fiction. The experiments were carried out using the ProseRhythmDetector software tool for Russian and English languages. Text corpora contains 300 texts for each language.

Keywords: stylometry; natural language processing; rhythm features; genres; text classification

INFORMATION ABOUT THE AUTHORS

Ksenia Vladimirovna Lagutina correspondence author	orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru postgraduate student.
Nadezhda Stanislavovna Lagutina	orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru PhD, associate professor.
Elena Igorevna Boychuk	orcid.org/0000-0001-6600-2971. E-mail: elena-boychouk@rambler.ru PhD, associate professor.

Funding: The reported study was funded by RFBR, project number 19-07-00243.

For citation: K. V. Lagutina, N. S. Lagutina, and E. I. Boychuk, “Text Classification by Genre Based on Rhythm Features”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 280-291, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Классификация текстов по жанрам на основе ритмических характеристик

К. В. Лагутина[1], Н. С. Лагутина[1], Е. И. Бойчук[2]

DOI: 10.18255/1818-1015-2021-3-280-291

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.
[2]Ярославский государственный педагогический университет им. К. Д. Ушинского, ул. Республиканская, д. 108/1, г. Ярославль, 150000 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 20 августа 2021 г.
После доработки 30 августа 2021 г.
Принята к публикации 1 сентября 2021 г.

Статья посвящена анализу ритма текстов различных жанров: художественных романов, рекламы, научных статей, отзывов, твитов и политических статей. Авторы выделили в текстах лексико-грамматические средства: анафору, эпифору, диакопу, апозиопезу и т. п., которые являются маркерами ритма текста. На их основе были подсчитаны статистические характеристики, описывающие количественно и структурно данные ритмические средства.
Полученная модель текста была визуализирована для статистического анализа с помощью диаграмм размаха и тепловых карт, которые показали отличия в ритме текстов различных жанров. Диаграммы размаха показали, что практически все жанры отличаются друг от друга по общей плотности ритмических характеристик. Тепловые карты показали различную структуру ритма у жанров.
Далее ритмические характеристики успешно использовались для классификации текстов по шести жанрам. Высокое качество классификации показало, что ритмические характеристики являются хорошим маркером для большинства жанров, в особенности для художественной литературы. Эксперименты проводились с помощью программного инструмента ProseRhythmDetector для русского и английского языков. Корпуса текстов содержат по 300 текстов для каждого языка.

Ключевые слова: стилометрия; обработка естественного языка; ритмические характеристики; жанры; классификация текстов

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ксения Владимировна Лагутина
автор для корреспонденции
orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
аспирант.

Надежда Станиславовна Лагутина
orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru
канд. физ.-мат. наук, доцент.

Елена Игоревна Бойчук
orcid.org/0000-0001-6600-2971. E-mail: elena-boychouk@rambler.ru
доктор фил. наук, доцент.

Финансирование: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00243.

Для цитирования: K. V. Lagutina, N. S. Lagutina, and E. I. Boychuk, “Text Classification by Genre Based on Rhythm Features”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 280-291, 2021.

Введение

Развитие методов автоматической обработки естественного языка позволяет исследователям ставить и решать сложные задачи на уровне дискурса, характеризующего смысловую организацию текстов. Идентификация жанра текста, как его функциональной характеристики, позволяет сделать акцент на иерархической природе текста в противоположность моделированию языка в виде плоской последовательности или неупорядоченного множества слов или букв [1].

Определение жанра текста является важной задачей как языкознания, так и создания корпусов текстов, без которых невозможно решение задач компьютерной лингвистики [2]. Этот факт отмечают и российские учёные [3]. Классификация текстов по жанру используется в исследованиях классической литературы и литературного языка [4], оказывается актуальной для анализа и извлечения информации из интернет-ресурсов [5], играет существенную роль для качественного машинного перевода текстов [6].

Чаще всего исследователи рассматривают функциональные стили текста, соответствующие научному, художественному, официально-деловому и публицистическому жанрам. Кроме того, учёные решают более специфические задачи по определению жанров художественной литературы [4, 7] или web-страниц [5]. В этих задачах методы решения можно условно отнести к одному из двух основных подходов: статистическому анализу стилометрических характеристик текста [8] и классификации на основе машинного обучения [7]. Однако в обоих случаях учёные подчёркивают, что самая важная часть работы связана с отбором релевантных параметров текста и исследованием роли различных типов характеристик для автоматической классификации по жанрам с конечной целью выявления наиболее эффективных по качеству.

В своих работах авторы данного исследования предложили комплекс высокоуровневых параметров художественного текста, основанных на фигурах речи образуемых повторением слов и словосочетаний [9, 10]. Эти параметры описывают ритм текста, который позволяет выявить уникальный авторский стиль и успешно классифицировать тексты по времени и авторам. Поэтому в данной работе была поставлена задача проанализировать влияние ритмических характеристик текстов, как нового типа стилометрических параметров, на определение жанра. Для этого выполняется статистический анализ ритмических характеристик и классификация текстов по жанрам: художественные романы, научные статьи, политические статьи, рекламные статьи, отзывы, твиты.

Обзор смежных работ

Наиболее распространённый подход к классификации текстов, в том числе по жанрам, основан на подборе и адаптации методов машинного обучения. В статье [11] проведён сравнительный анализ пяти различных базовых алгоритмов классификации (наивный байесовский классификатор, метод опорных векторов, логистическая регрессия, метод k-ближайших соседей и алгоритм «случайный лес») в сочетании с методами ансамблевого обучения (такими как Boosting, Bagging и Random Subspace). На основе эмпирического анализа представлена схема классификации ансамблей, которая объединяет Random Subspace и случайного леса с четырьмя типами признаков (признаки, используемые в атрибуции авторства, n-граммы символов, n-граммы части речи и частота редких слов). Для корпуса текстов LFA наивысшая средняя прогностическая эффективность, полученная по предложенной схеме, составляет 94,43 %. Однако немаловажную часть этой работы занимает сравнение и подбор подходящих характеристик текста.

На анализе подходящих методов машинного обучения базируются многие исследователи национальных языков. Учёные провели классификацию арабских текстов по одному из четырех жанров: реклама, новости, личные и научные документы [12]. Они использовали те же классификаторы, что и в исследовании [11]. Качество определения жанра оказалось очень высоким: F-мера более 90 %. Авторы отметили, что подход к вычислению характеристик текста на уровне «мешок слов»

дал низкую эффективность, поэтому они использовали более сложные параметры стиля текста. Интересно, что предсказание текстов личного и научного жанра более точно, чем прогнозирование рекламы и новостей. Самый лучший результат определения жанра дали деревья решений, но они были построены на основе статистического анализа конкретного корпуса текстов, поэтому высокое качество ожидаемо, но не носит обобщающий характер.

Более универсальный подход к анализу текстов с точки зрения определения их жанровой принадлежности основан на применении сверточных нейронных сетей [13]. Авторами разработана архитектура сверточной нейронной сети с использованием векторного представления слов на основе модели word2vec. Эффективность работы построенной модели проверена для пяти жанров: история, детективы, детская литература, поэзия, фантастика. Точность классификации составила 78,64 %. В качестве обучающих данных был выбран корпус русскоязычных текстов Максима Мошкова.

Эксперименты по определению стилей и жанров поэтических текстов (ода, элегия, баллада, эпиграмма и т. д.) с использованием корпуса текстов лицейской лирики А. С. Пушкина описаны в работе [14]. Авторы выбирали наиболее точный алгоритм классификации с использованием известных приемов ансамблирования базовых алгоритмов, таких как взвешенное голосование, бустинг и стекинг. В качестве характеристических признаков текстов использовались униграммы, биграммы и триграммы слов. Было установлено, что даже с помощью простых классификаторов на основе этих лексических признаков можно получить хороший результат решения задачи. Лучшая точность оказалась у многослойного персептрона 99 % при использовании в качестве характеристик текстов триграмм слов. Авторы отметили важность применения такого анализа поэзии для эксперта-лингвиста.

Хотя стандартные алгоритмы классификации показывают высокое качество определения жанров текста, все авторы описанных выше работ в большей или меньшей степени обращают внимание на выбор характеристик текста. Обобщая их результаты можно обратить внимание на то, что использование более сложных стилометрических параметров, таких как n-граммы слов, даёт лучшее качество решения задачи. К такому же выводу приходят авторы работы [5], которые проанализировали вклад различных типов характеристик в решение проблемы определения жанра в компьютерной лингвистике. Аналогичная задача решается в исследовании [8]. В нём отмечается, что ключевую роль играют синтаксические особенности текста, влияние которых, различается в зависимости от жанра.

Исследователи русского языка, обращаясь к проблеме определения жанра текста, выявляют сложные лексические особенности стиля, отличающие разные жанры. В работе [15] выдвигается гипотеза о том, что коэффициенты соотношения частот семантически противопоставленных предлогов русского языка могут указывать на стилевую принадлежность текстов. Материалом для экспериментов послужили корпусы текстов разных функциональных стилей и разной тематики: общий, художественный, публицистический, нехудожественный, устный из Национального корпуса русского языка (НКРЯ), корпусов Araneum Russicum Russicum и Araneum Russicum Externum, корпуса текстов из социальных сетей Facebook и Twitter, корпуса художественных текстов с сайта Либрусек. Эксперименты подтвердили значимость ряда коэффициентов в диагностике стиля и типа текстов. Так же была получена информация о семантическом наполнении предложных конструкций, которая важна для определения стилевых и жанровых характеристик текстов.

Другие учёные предлагают вместе со стилометрическими параметрами использовать дополнительные статистические числовые характеристики. Авторы работы [16] классифицируют по жанрам фрагменты научных и научно-популярных текстов академика Александра Евгеньевича Ферсмана, выдающегося ученого и популяризатора науки. Они используют характеристики уровня символов, индексы на основе частоты гласных букв, частот биграмм и триграмм символов, индексы

энтропии и сжимаемости текстов. Однако выбранный для экспериментов корпус очень мал, всего 44 фрагмента, что скорее всего обусловлено трудоёмкостью сбора и качественной разметки таких ресурсов.

В статье [17] проводились эксперименты на материале русскоязычных корпусов текстов, принадлежащих четырём функциональным стилям: научному, художественному, официально-деловому и публицистическому. В качестве характеристик текста использовались частоты морфологических параметров, частоты pos-тегов, некоторые биграммы, длины слов и предложений. Кроме того, рассчитывался комбинированный параметр β, отражающий соотношение динамичности и статичности текстов коллекции [18]. На основе статистического анализа была определена система правил для классификации текстов по жанрам и получены высокие результаты. F–мера составила 99 % для художественного и делового стилей, 83 % для научного, 70 % для публицистического. Результаты исследования аналогичны прогнозированию жанра текста на основе дерева решений [12], однако объём использованного корпуса текстов значительно меньше, по 65 текстов для каждого стиля.

Таким образом, анализ использования сложных стилометрических характеристик текста для определения его жанра является перспективной и актуальной задачей. Определение степени влияния различных типов параметров позволит строить эффективные системы извлечения информации из текстов на естественном языке и проведения лингвистических исследований.

Ритмические характеристики

Числовые ритмические характеристики основываются на ритмических средствах, непосредственно появляющихся в тексте. В данном исследовании используются лексико-грамматические средства: анафора, эпифора, симплока, анадиплозис, эпаналепсис, многосоюзие, диакопа, эпизевксис, хиазм, апозиопеза, повторяющиеся вопросительные и восклицательные предложения. Определения ритмических средств и алгоритмы их поиска приведены в предыдущих работах авторов [9, 10]. Апозиопеза и повторяющиеся вопросительные и восклицательные предложения основываются на появлениях знаков препинания. Остальные ритмические средства состоят из повторяющихся слов или словосочетаний.

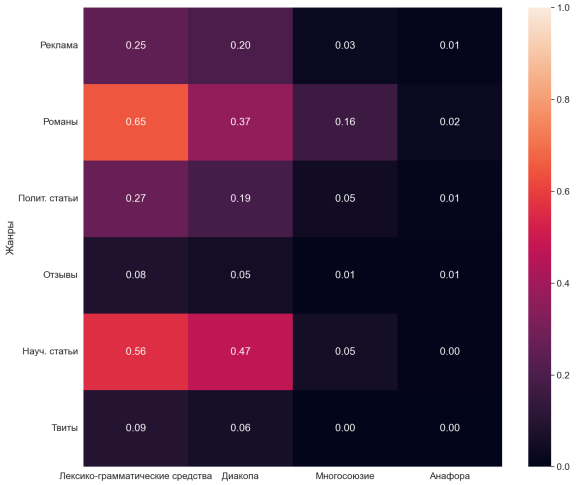
Для лексико-грамматических средств были вычислены следующие стилометрические характеристики:

1. количество появлений в тексте конкретного средства, делённое на количество предложений;
2. количество появлений в тексте всех средств, делённое на количество предложений;
3. доля уникальных слов среди всех, составляющих средства, т. е. тех, которые повторяются только один раз;
4. доли существительных, прилагательных, глаголов и наречий среди слов, составляющих средства.

Данные ритмические характеристики описывают как относительную частоту появления лексико-грамматических средств, т. е. их плотность, так и статистику для структуры лексических средств.

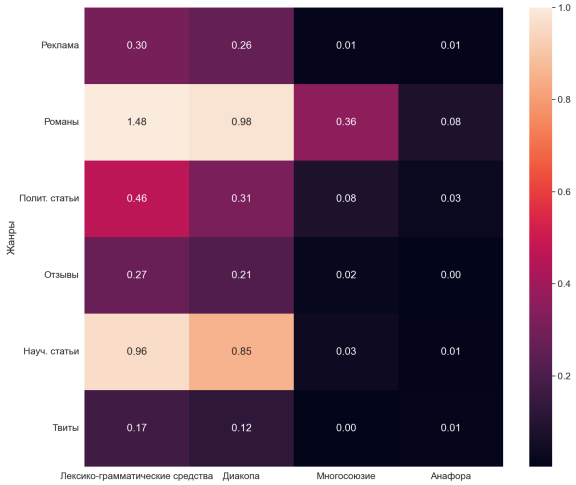
Статистический анализ ритмических характеристик

Для статистического анализа ритмических характеристик были собраны корпуса текстов в шести жанрах на русском и английском языках. Каждый корпус содержит по 100 художественных романов, научных статей, рекламных текстов и твитов, по 50 отзывов и политических статей, т. е. суммарно 500 текстов. Романы были взяты из корпуса из предыдущих исследований авторов [10]. Научные статьи были собраны из журналов Грамота, Диалог, International journal of digital evidence и Philosophical Transactions of the Royal Society of London. Рекламные тексты были взяты с сайтов auto.ru, detmir.ru и smartmedicalbuyer.com. Отзывы были собраны с сайта tripadvisor.com. Политические статьи представляют собой текстовые расшифровки речей президентов и министров России и США.



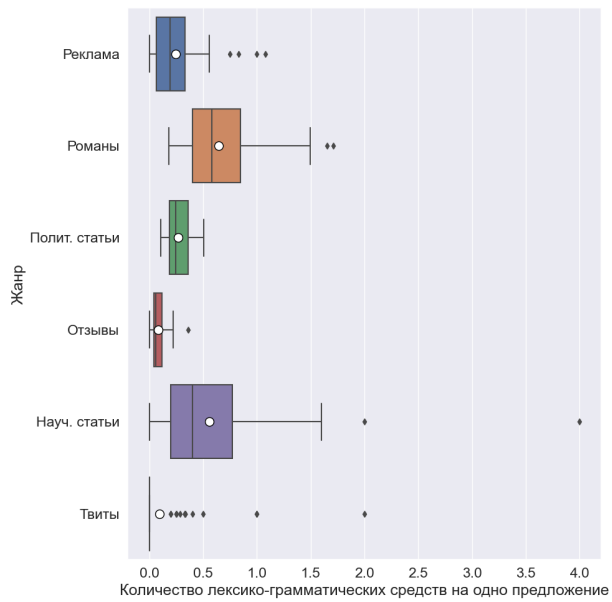
a)

Fig. 1. Heat map of mean values for genres for frequent features a) in Russian-language texts, b) in English-language texts



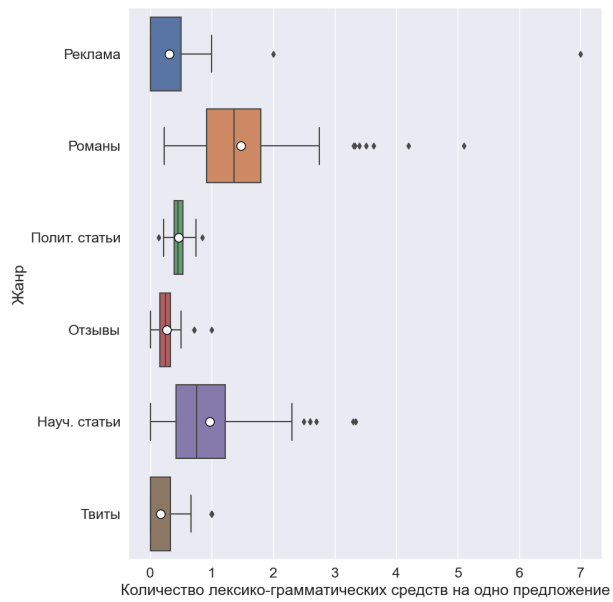
b)

Рис. 1. Тепловая карта средних значений по жанрам для часто встречающихся средств а) в русскоязычных текстах, b) в англоязычных текстах



a)

Fig. 2. Boxplot for lexico-grammatical features a) in Russian-language texts, b) in English-language texts



b)

Рис. 2. Диаграмма размаха лексико-грамматических средств а) в русскоязычных текстах, b) в англоязычных текстах

Ритмические характеристики вычисляются независимо для каждого текста и визуализируются с помощью тепловых карт и диаграмм размаха.

Тепловые карты представлены на рис. 1. Строки соответствуют жанрам, столбцы — самым часто встречающимся ритмическим средствам. В ячейках указано, сколько раз в среднем данное средство появляется в одном предложении в текстах данного жанра — плотность ритмического средства.

В текстах лексико-грамматические средства появляются чаще всего в романах и научных статьях, реже всего — в рекламе, отзывах и твитах. Самые популярные лексико-грамматические средства — диакопа, многосоюзие и анафора. Романы содержат больше разнообразных ритмических средств, чем тексты других жанров, в научных статьях обычно появляется диакопа и очень редко — другие средства. Твиты практически не содержат ритмических средств.

Отзывы на английском языке содержат в среднем в несколько больше маркеров ритма, чем на русском, и по плотности ритмических средств близки к рекламе. В остальном жанры на разных языках похожи по тенденциям ритма с поправкой на то, что англоязычные тексты содержат в среднем больше ритмических средств, чем русскоязычные тексты в тех же жанрах.

Более подробно распределение плотности лексико-грамматических средств представлено на диаграммах размаха 2. Прямоугольник показывает границы первого и третьего квартилей распределения, чёрная вертикальная линия внутри него — медианное значение, белый круг — среднее значение. Чёрная горизонтальная линия показывает граничные значения распределения, чёрные ромбы — выбросы.

Диаграммы показывают, что отзывы и политические статьи достаточно однородны по количеству ритмических средств. Среди русскоязычных текстов наиболее разнообразны по плотности ритма твиты и научные статьи, среди англоязычных — реклама и романы.

В целом графики и диаграммы показывают, что каждый жанр имеет свои особенности ритма, и по совокупности характеристик жанры отличаются друг от друга. Это означает, что ритмические характеристики могут быть хорошими маркерами жанра, что должна подтвердить классификация текстов.

Классификация по жанрам

Описанный корпус текстов был классифицирован по жанрам на основе числовых ритмических характеристик. Для классификации были взяты все ритмические характеристики, кроме количества появлений в тексте всех средств, делённого на количество предложений, поскольку она является суммой других характеристик.

Классификация проводилась двумя способами:

- бинарная классификация для каждого жанра, когда тексты классифицировались на принадлежащие и не принадлежащие конкретному жанру;
- мультиклассовая классификация на шесть жанров: художественные романы, научные статьи, политические статьи, рекламные статьи, отзывы, твиты.

Для обоих способов применялись одни и те же классификаторы:

- классификатор AdaBoost — мета-алгоритм машинного обучения, который объединяет результаты 50 классификаторов-деревьев решений, корректирующих неправильно классифицированные тексты;
- двунаправленная LSTM — рекуррентная нейронная сеть со слоем двунаправленной долгой краткосрочной памяти (LSTM) с 64 блоками и полносвязным выходным слоем, использующим функцию активации Softmax для мультиклассовой классификации и Sigmoid для бинарной;
- GRU — рекуррентная нейронная сеть со слоем Gated Recurrent Unit (GRU) с 4 блоками и полносвязным выходным слоем, использующим функцию активации Softmax для мультиклассовой классификации и Sigmoid для бинарной.

Table 1. Binary text classification by genres for Russian language

Таблица 1. Бинарная классификация текстов по жанрам для русского языка

Жанр	Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
Реклама	GRU	44.8	1.3	50.0	0.0	47.3	0.4
Романы	GRU	**94.2**	2.8	**92.5**	2.3	**94.1**	1.9
Полит. статьи	GRU	44.6	1.6	50.0	0.0	47.1	0.9
Отзывы	GRU	71.1	18.0	69.6	10.8	66.1	17.4
Научн. статьи	GRU	79.3	2.7	76.1	6.9	77.5	5.1
Твиты	GRU	83.3	3.2	81.0	5.2	81.1	3.3
Реклама	LSTM	**68.5**	21.3	52.7	3.5	52.5	7.4
Романы	LSTM	**96.6**	1.9	**94.6**	2.6	**96.7**	1.9
Полит. статьи	LSTM	85.3	6.7	67.9	11.5	68.0	16.0
Отзывы	LSTM	81.0	11.2	**87.3**	12.1	**86.9**	6.3
Научн. статьи	LSTM	80.5	2.8	76.1	4.6	78.5	2.2
Твиты	LSTM	85.3	7.3	83.3	7.1	86.2	2.4
Реклама	AdaBoost	58.1	7.2	**56.5**	5.4	**56.7**	6.3
Романы	AdaBoost	**97.0**	1.5	**97.9**	1.0	**97.4**	1.0
Полит. статьи	AdaBoost	**95.7**	4.8	**92.1**	5.6	**93.3**	3.9
Отзывы	AdaBoost	**83.5**	17.1	79.4	15.5	81.2	16.0
Научн. статьи	AdaBoost	**84.6**	1.3	**84.3**	3.3	**84.1**	1.7
Твиты	AdaBoost	**90.9**	2.7	**89.5**	4.1	**89.9**	3.0

Для обучения нейронных сетей LSTM и GRU применяется категориальная кросс-энтропия как функция потерь и алгоритм оптимизации Adam.

Данные классификаторы уже доказали свое качество в решении современных задач обработки ритма текстов [9, 10], поэтому они были выбраны для экспериментов.

Для классификации корпус был разделён случайным образом на обучающую и тестовую выборки в отношении 4:1. Это позволило провести пятикратную кросс-валидацию для анализа стабильности результатов. Оценка качества выполнялась с помощью трёх стандартных мер: точность, полнота и F-мера [19], а также их стандартные отклонения.

Алгоритмы визуализации ритмических характеристик, классификации по жанрам и оценки результатов реализованы в инструменте ProseRhythmDetector, который доступен в Интернете по адресу https://github.com/text-processing/prose-rhythm-detector. Он написан на языке программирования Python и использует библиотеки StanfordNLP 0.2.0, Scikit-Learn 0.23.2 и Keras 2.4.3.

Результаты бинарной классификации представлены в таблицах 1 и 2 для русского и английского языка соответственно. Для точности, полноты и F-меры справа указаны стандартные отклонения при кросс-валидации.

Среди всех жанров лучше всего отделяются от остальных художественные романы с F-мерой более 97 % и политические статьи с F-мерой более 92 %. Отзывы, научные статьи и твиты тоже хорошо классифицируются (F-мера более 76 %). Следует отметить, что научные статьи на английском языке классифицируются лучше, чем на русском, а твиты — наоборот. Русскоязычная реклама отделяется от других жанров хуже остальных.

Стандартные отклонения в большинстве случаев низки: менее 5 %, что говорит о высокой стабильности классификации.

Среди трёх классификаторов лучших значений точности, полноты и F-меры чаще достигает AdaBoost за исключением русскоязычных отзывов, где LSTM превосходит его по F-мере на 5 % и показывает стандартное отклонение ниже на 10 %. Романы классифицируются очень хорошо всеми классификаторами: F-мера более 94 %.

Table 2. Binary text classification by genres for English language

Таблица 2. Бинарная классификация текстов по жанрам для английского языка

Жанр	Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
Реклама	GRU	73.4	2.2	67.6	2.4	68.4	4.3
Романы	GRU	**93.4**	1.7	**94.3**	4.1	**94.3**	1.9
Полит. статьи	GRU	82.5	18.8	64.8	9.2	63.8	20.8
Отзывы	GRU	45.4	0.8	51.0	2.0	47.4	0.5
Научн. статьи	GRU	87.1	2.0	83.5	3.1	83.0	3.1
Твиты	GRU	**79.1**	5.8	72.9	3.0	76.8	4.9
Реклама	LSTM	**78.3**	3.3	72.9	5.1	71.8	3.7
Романы	LSTM	**96.5**	2.6	**95.8**	2.5	**95.8**	1.9
Полит. статьи	LSTM	87.7	7.6	80.5	12.1	83.0	6.2
Отзывы	LSTM	68.1	16.5	58.0	6.7	61.5	11.6
Научн. статьи	LSTM	88.2	5.3	85.5	3.5	86.4	3.8
Твиты	LSTM	**82.4**	5.7	73.9	3.8	74.8	5.4
Реклама	AdaBoost	75.6	4.4	**73.3**	4.3	**74.3**	4.4
Романы	AdaBoost	**98.3**	2.0	**98.0**	2.3	**98.1**	2.1
Полит. статьи	AdaBoost	**94.3**	2.7	**92.1**	6.6	**92.7**	3.8
Отзывы	AdaBoost	**79.3**	11.8	**74.2**	8.1	**76.3**	9.3
Научн. статьи	AdaBoost	**91.6**	3.9	**88.3**	3.5	**89.7**	3.6
Твиты	AdaBoost	78.5	4.0	**76.4**	3.1	**77.1**	3.1

Table 3. Multi-class text classification by genres for Russian language

Таблица 3. Мультиклассовая классификация текстов по жанрам для русского языка

Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
GRU	71.5	11.2	69.7	4.0	65.5	2.9
LSTM	**77.1**	5.6	**77.5**	5.5	**77.5**	6.8
AdaBoost	43.3	14.0	43.3	9.0	36.7	12.1

Таблицы 3 и 4 демонстрируют результаты для мультиклассовой классификации. Здесь уже классификатор на основе нейросети LSTM существенно превосходит остальные: он достигает F-меры более 74 %, тогда как AdaBoost не показывает и 40 % F-меры.

В целом результаты у мультиклассовой классификации ниже, чем в лучших случаях у бинарной, а стандартные отклонения такие же низкие. Тем не менее точность, полнота и F-мера значительно высоки: более 72 %. Мультиклассовая классификация выполняется более эффективно для русского языка: для него точность, полнота и F-мера составляют 77 %.

Для того, чтобы обнаружить и проанализировать ошибки классификации, для мультиклассовой классификации алгоритмом LSTM были собраны неверные предсказания алгоритма. Они агрегированы в таблицах 5 и 6 как пример ошибок из одного раунда классификации при делении корпусов текстов случайным образом на обучающую и тестовую выборки в отношении 4:1.

Строки таблицы соответствуют исходным жанрам текстов, а столбцы — жанрам, предсказанным неверно. В ячейках указывается, сколько текстов исходного жанра было ошибочно причислено к

Table 4. Multi-class text classification by genres for English language

Таблица 4. Мультиклассовая классификация текстов по жанрам для английского языка

Классификатор	Точность	Стд. откл.	Полнота	Стд. откл.	F-мера	Стд. откл.
GRU	71.0	4.6	71.7	3.4	68.9	5.1
LSTM	**72.4**	2.7	**75.3**	2.2	**74.1**	3.7
AdaBoost	42.7	8.0	46.6	2.7	39.5	3.5

Table 5. Errors of the multi-class text classification by genres for Russian language

Таблица 5. Ошибки мультиклассовой классификации текстов по жанрам на русском языке

Исходный жанр	Реклама	Романы	Полит. статьи	Отзывы	Научн. статьи	Твиты
Реклама	-	0	5	0	4	3
Романы	0	-	0	0	0	0
Полит. статьи	0	0	-	0	1	0
Отзывы	0	0	0	-	0	2
Научн. статьи	0	0	3	0	-	3
Твиты	1	0	0	0	2	-

Table 6. Errors of the multi-class text classification by genres for English language

Таблица 6. Ошибки мультиклассовой классификации текстов по жанрам на английском языке

Исходный жанр	Реклама	Романы	Полит. статьи	Отзывы	Научн. статьи	Твиты
Реклама	-	0	0	2	0	4
Романы	0	-	2	0	0	0
Полит. статьи	0	1	-	0	0	0
Отзывы	3	0	0	-	2	0
Научн. статьи	1	0	1	1	-	0
Твиты	7	0	0	0	2	-

жанру, указанному в столбце. Например, пять рекламных текстов на русском языке были приняты за политические статьи.

Из результатов агрегирования ошибок можно сделать вывод, что классификатор нередко причисляет рекламу к любому жанру, кроме художественных романов. Твиты и реклама часто смешиваются между собой. Ошибки в остальных жанрах достаточно случайны и, вероятно, вызываются особенностями конкретных текстов. Например, два англоязычных романа, которые были ошибочно классифицированы как политические статьи, содержат мало ритмических средств, что обычно не характерно для их жанра.

Обсуждение результатов с лингвистической точки зрения

Полученные результаты по количеству ритмических средств в текстах разных жанров позволяет судить о ритмической специфике того или иного жанра в рамках одного языка. В русском языке, в котором по сравнению с другими языками, как показали ранние исследования [9], частотность ритмических средств ниже, чем в других языках, что подтверждают показатели по жанрам. Это безусловно связано с языковой спецификой, типологическими особенностями языка, в особенности с критериями частеречной классификации. Например, в английском языке значим как морфологический, так и синтаксический критерий в частеречной классификации, но для одноморфемных слов важен синтаксический критерий, их позиция в предложении. Кроме того, английский язык отличается большей степенью номинативности в сопоставлении с русским. Номинативность английского языка увеличивается также за счет герундия. Такое стремление к номинативности обусловлено этносоциокультурными факторами, а именно исторически сформировавшимся в рамках британской культуры уважением к факту и научной точности в разговоре [20]. Русский язык по сравнению с английским обладает большей степенью глагольности, при этом довольно много в русском языке глаголов, передающих эмоциональное состояние (волноваться, гневаться, раздражаться, радоваться и т. д.), что подчеркивает значимость эмоционального начала [21].

Данные особенности важны для характеристики ритмических средств, поскольку большая их часть выражается при помощи существительных, что обусловлено грамматической структурой

предложения, в котором второстепенные члены реже всего выражены глаголом, но чаще именем существительным, а также прилагательным или наречием. Таким образом, английский язык в целом обладает большей степенью ритмизации только на основе своей структуры, а именно на основе стремления к номинативности в морфологии.

Что касается различных показателей в различных жанрах, то несомненным является преобладание ритмических средств в художественной литературе, поскольку поэтичность изложения, творческий подход позволяют сосредоточиться на образах, которые могут реализовываться через различные типы повторов. Объяснением того, что научный текст близок к художественному, может быть определенная структурированная, отлаженная терминосистема, характерная для научных текстов в целом. В этом случае повторы обусловлены необходимостью оперирования конкретной для каждой дисциплины и отрасли знания лексики.

Что касается твитов, которые содержат наименьшее количество средств, то в качестве основной причины этого можно отметить их прагматическую функцию — выражение собственного мнения, четкое, краткое, не поэтизированное. То же и для отзывов, которые в русском языке также имеют низкий уровень ритмических средств. Однако в англоязычных отзывах, как и в рекламе в обоих языках, ритмические средства более активны, что также может объясняться большей номинативностью английского языка.

Заключение

В статье оценивалось влияние ритмических характеристик текстов на определение жанра. Задача была выполнена в два этапа: статистический анализ ритмических характеристик и классификация текстов по шести жанрам: художественные романы, научные статьи, политические статьи, рекламные статьи, отзывы, твиты. Визуализация статистических данных о ритмических характеристиках показала, что тексты различных жанров отличаются по маркерам стиля. При классификации текстов по жанрам с помощью этих характеристик и современных алгоритмов AdaBoost и LSTM были достигнуты достаточно высокие значения метрик качества: не менее 76 % F-меры для всех жанров, кроме рекламы. С наивысшим качеством около 98 % точности, полноты и F-меры были классифицированы художественные романы.

Перспективным направлением дальнейших исследований будет анализ ошибок классификации, который позволит лучше изучить ритмические особенности текстов и учесть их в моделях текстов.

References

[1] J. Worsham and J. Kalita, “Genre identification and the compositional effect of genre in literature”, in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1963–1973.

[2] M. N. Melissourgou and K. T. Frantzi, “Genre identification based on SFL principles: The representation of text types and genres in English language teaching material”, *Corpus Pragmatics*, vol. 1, no. 4, pp. 373–392, 2017.

[3] L. A. Kochetova and V. V. Popov, “Research of Axiological Dominants in Press Release Genre based on Automatic Extraction of Key Words from Corpus”, *Nauchnyi dialog*, no. 6, 2019, In Russian.

[4] S. E. Murphy, “Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560-1640”, *ICAME Journal*, pp. 59–82, 2019.

[5] R. Malhotra and A. Sharma, “Quantitative evaluation of web metrics for automatic genre classification of web pages”, *International Journal of System Assurance Engineering and Management*, vol. 8, no. 2, pp. 1567–1579, 2017.

[6] D. DEJICA, "Understanding Technical and Scientific Translation: A Genre-based Approach", *Scientific Bulletin of the Politehnica University of Timisoara. Transactions on Modern Languages/Buletinul Stiintific al Universitatii Politehnica din Timisoara. Seria Limbi Moderne*, vol. 19, no. 1, pp. 56–66, 2020.

[7] V. Thakur and A. C. Patel, "An Improved Dictionary Based Genre Classification Based on Title and Abstract of E-book Using Machine Learning Algorithms", in *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, Springer, 2021, pp. 323–337.

[8] A. Cimino, M. Wieling, F. Dell'Orletta, S. Montemagni, and G. Venturi, "Identifying predictive features for textual genre classification: the key role of syntax", *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, pp. 107–112, 2017.

[9] K. Lagutina, A. Poletaev, N. Lagutina, E. Boychuk, and I. Paramonov, "Automatic extraction of rhythm figures and analysis of their dynamics in prose of 19th-21st centuries", *Proceedings of the 26th Conference of Open Innovations Association FRUCT*, pp. 247–255, 2020.

[10] K. Lagutina, N. Lagutina, E. Boychuk, V. Larionov, and I. Paramonov, "Authorship verification of literary texts with rhythm features", *Proceedings of the 28th Conference of Open Innovations Association FRUCT*, pp. 240–251, 2021.

[11] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification", *Journal of Information Science*, vol. 44, no. 1, pp. 28–47, 2018.

[12] A. M. El-Halees, "Arabic Text Genre Classification", *Journal of Engineering Research and Technology*, vol. 4, no. 3, pp. 105–109, 2017.

[13] I. A. Batraeva, A. D. Nartsev, and A. S. Lezgyan, "Using the analysis of semantic proximity of words in solving the problem of determining the genre of texts within deep learning", *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie vychislitelnaja tehnika i informatika*, no. 50, pp. 14–22, 2020, In Russian.

[14] V. B. Barahnin, O. Y. Kozhemyakina, E. V. Rychkova, I. S. Pastushkov, and Y. S. Borzilova, "Izvlechenie leksicheskih i metroritmicheskih priznakov, harakternyh dlya zhanra i stilya i ih kombinacij v processe avtomatizirovannoj obrabotki tekstov na russkom yazyke", *Sovremennye informacionnye tekhnologii i IT-obrazovanie*, vol. 14, no. 4, pp. 888–895, 2018, In Russian.

[15] O. A. Mitrofanova and A. D. Moskvina, "On the Role of Prepositional Statistics for Genre Identification of Russian texts", *International Journal of Open Information Technologies*, vol. 8, no. 11, pp. 91–96, 2020, In Russian.

[16] L. G. Gorbich and A. A. Zhivoderov, "Using statistical indexes to distinguish between scientific and popular science texts on the example of the works of A. E. Fersman", *Software & Systems*, vol. 33, no. 4, pp. 720–725, 2020, In Russian.

[17] A. R. Dubovik, "Automatic text style identification in terms of statistical parameters", *Komp'yuternaya lingvistika i vychislitel'nye ontologii*, no. 1, pp. 29–45, 2017, In Russian.

[18] A. Y. Antonova, E. S. Klyshinskij, and E. V. YAgunova, "Opredelenie stilevyh i zhanrovyh harakteristik kollekcij tekstov na osnove chasterechnoj sochetaemosti", *Otkrytye sistemy*, vol. 3, pp. 80–85, 2011, In Russian.

[19] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[20] L. Kozlova, "Sravnitel'naya tipologiya anglijskogo i russkogo yazykov", *Barnaul: AltGPU*, no. 20019, p. 180, 2019, In Russian.

[21] A. Wierzbicka, *The semantics of grammar*. John Benjamins Publishing, 1988, vol. 18, p. 617.

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF DATA

Word-embedding Based Text Vectorization Using Clustering

V. I. Yuferev[1], N. A. Razin[2]

DOI: 10.18255/1818-1015-2021-3-292-311

[1]Department of Information Technologies of the Central Bank of the Russian Federation, Laboratory of innovations "Novosibirsk", 12 Neglinnaya str., Moscow 107016, Russia.
[2]Department of Counteraction to Unfair Practices, the Central Bank of the Russian Federation, 12 Neglinnaya str., Moscow 107016, Russia.

MSC2020: 97R40, 68T50
Research article
Full text in Russian

Received June 23, 2021
After revision August 16, 2021
Accepted August 25, 2021

It is known that in the tasks of natural language processing, the representation of texts by vectors of fixed length using word-embedding models makes sense in cases where the vectorized texts are short.
The longer the texts being compared, the worse the approach works. This situation is due to the fact that when using word-embedding models, information is lost when converting the vector representations of the words that make up the text into a vector representation of the entire text, which usually has the same dimension as the vector of a single word.
This paper proposes an alternative way for using pre-trained word-embedding models for text vectorization. The essence of the proposed method consists in combining semantically similar elements of the dictionary of the existing text corpus by clustering their (dictionary elements) embeddings, as a result of which a new dictionary is formed with a size smaller than the original one, each element of which corresponds to one cluster. The original corpus of texts is reformulated in terms of this new dictionary, after which vectorization is performed on the reformulated texts using one of the dictionary approaches (TF-IDF was used in the work). The resulting vector representation of the text can be additionally enriched using the vectors of words of the original dictionary obtained by decreasing the dimension of their embeddings for each cluster.
A series of experiments to determine the optimal parameters of the method is described in the paper, the proposed approach is compared with other methods of text vectorization for the text ranking problem – averaging word embeddings with TF-IDF weighting and without weighting, as well as vectorization based on TF-IDF coefficients.

Keywords: word embedding; Fasttext; TF-IDF; averaging; clustering; text similarity; distance; text ranking

INFORMATION ABOUT THE AUTHORS

Vitaly I. Yuferev
correspondence author
orcid.org/0000-0003-3245-6240. E-mail: YuferevVI@mail.cbr.ru
Chief expert, Master of science.

Nikolai A. Razin
orcid.org/0000-0002-7669-776X. E-mail: razinna@cbr.ru
Head of division, PhD.

For citation: V. I. Yuferev and N. A. Razin, "Word-embedding Based Text Vectorization Using Clustering", *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 292-311, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Векторизация текстов на основе word-embedding моделей с использованием кластеризации

В. И. Юферев[1], Н. А. Разин[2]

DOI: 10.18255/1818-1015-2021-3-292-311

[1]Департамент информационных технологий Центрального банка Российской Федерации, Инновационная лаборатория «Новосибирск», ул. Неглинная, д. 12, г. Москва, 107016 Россия.
[2]Департамент противодействия недобросовестным практикам, Центральный банк Российской Федерации, ул. Неглинная, д. 12, г. Москва, 107016 Россия.

УДК 004.8
Научная статья
Полный текст на русском языке

Получена 23 июня 2021 г.
После доработки 16 августа 2021 г.
Принята к публикации 25 августа 2021 г.

Известно, что в задачах обработки естественного языка представление текстов векторами фиксированной длины с использованием word-embedding моделей оправдано в тех случаях, когда векторизуемые тексты являются короткими. Чем сравниваемые тексты длиннее, тем подход работает хуже. Такая ситуация обусловлена тем, что при использовании word-embedding моделей происходит потеря информации при преобразовании векторных представлений слов, составляющих текст, в векторное представление всего текста, имеющее обычно ту же размерность, что и вектор отдельного слова.

В настоящей работе предлагается альтернативный способ использования предобученных word-embedding моделей для векторизации текстов. Суть предлагаемого способа заключается в объединении семантически близких элементов словаря имеющегося корпуса текстов путем кластеризации их (элементов словаря) эмбеддингов, в результате чего формируется новый словарь размером меньше исходного, каждый элемент которого соответствует одному кластеру. Исходный корпус текстов переформулируется в терминах этого нового словаря, после чего на переформулированных текстах выполняется векторизация одним из словарных подходов (в работе применялся TF-IDF). Полученное векторное представление текста дополнительно может обогащаться с использованием векторов слов исходного словаря, полученных путем уменьшения размерности их эмбеддингов по каждому кластеру. В работе описана серия экспериментов по определению оптимальных параметров предлагаемого подхода; для задачи ранжирования текстов приведено сравнение подхода с другими способами векторизации – усреднением эмбеддингов слов со взвешиванием по TF-IDF и без взвешивания, а также с векторизацией на основе TF-IDF коэффициентов.

Ключевые слова: эмбеддинговые модели; Fasttext; TF-IDF; усреднение; кластеризация; семантическое сходство текстов; определение расстояний; ранжирование текстов

ИНФОРМАЦИЯ ОБ АВТОРАХ

Виталий Иванович Юферев
автор для корреспонденции
orcid.org/0000-0003-3245-6240. E-mail: YuferevVI@mail.cbr.ru
Главный эксперт, магистр техники и технологий.

Николай Алексеевич Разин
orcid.org/0000-0002-7669-776X. E-mail: razinna@cbr.ru
Начальник отдела, кандидат физ.-мат. наук.

Для цитирования: V. I. Yuferev and N. A. Razin, “Word-embedding Based Text Vectorization Using Clustering”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 292-311, 2021.

Введение

Распространенной задачей в области обработки естественного языка является ранжирование текстов, то есть определение, какой из двух текстов T_1 или T_2 семантически ближе к тексту T.

Базовый подход для определения семантического сходства текстов состоит из двух основных этапов: представление сравниваемых текстов в векторном виде, отражающем семантику текста, и последующее определение расстояний между полученными векторами [1].

Существуют различные способы представления текстов в векторном виде: с использованием словарей, с использованием word-embedding моделей, с использованием языковых моделей, основанных на архитектуре Transformer.

К недостаткам словарных подходов, в частности основанных на TF-IDF [2], можно отнести отсутствие учета семантики слов [3] при определении близости текстов, если составляющие их слова не пересекаются, то они будут определены как далекие друг от друга, независимо от того, содержат ли тексты слова, похожие по смыслу. Применение современных языковых моделей, основанных на архитектуре Transformer, также имеет некоторые ограничения: высокие требования к вычислительным ресурсам и предельный размер входа [4—6].

Применение word-embedding моделей для векторизации текстов выглядит оправданным в случаях, когда имеют место указанные выше ограничения других подходов. Особенно актуально применение предобученных word-embedding моделей, когда имеющийся корпус текстов страдает от недостатка данных [7].

Наиболее простой и часто применяемый подход по формированию векторного представления текста с использованием word-embedding моделей заключается в формировании эмбеддингов входящих в текст слов с последующим формированием вектора текста путем усреднения (сложения) полученных эмбеддингов слов [8, 9]. В качестве улучшения подхода дополнительно может выполняться взвешивание полученных векторов эмбеддингов слов, например, по TF-IDF [10].

Решение прикладных задач обработки естественного языка подтверждает, что применение подхода с векторизацией текстов при помощи усреднения (сложения) эмбеддингов слов, как правило, даёт приемлемое качество (конкретные метрики и их значения зависят от задачи) на коротких текстах (на предложениях и меньше) [11] и по мере увеличения длины текстов качество снижается до неприемлемого.

Низкое качество при усреднении (сложении) на длинных текстах можно объяснить следующим. Эмбеддинг слова представляет его семантику в виде вектора. Усреднение (сумма) двух эмбеддингов (обозначим A и B) также представляет собой вектор (обозначим C) той же размерности. При этом возникает неопределенность, связанная с тем, что вектор C может быть получен указанной комбинацией (усреднение, сумма) как исходных векторов A и B, так и некоторых других векторов D и E, отражающих семантику, отличную от семантики, кодируемой векторами A и B. Соответственно, чем больше исходных эмбеддингов усредняется, тем выше у результирующего вектора неопределенность относительно семантики исходных слов.

Попытка увеличения размерности вектора текста по сравнению с векторами слов приводится в [12—14]. Как и в описываемом в настоящей работе подходе в этих работах предлагается выполнить кластеризацию на словаре корпуса текстов. Однако указанные подходы имеют следующие ограничения: необходимость наличия достаточно большого корпуса текстов для обучения wod2vec модели, отсутствие описания возможности применения для n-грамм.

1. Описание подхода

Чтобы обойти ограничения алгоритмов по векторизации текстов – словарного и основанного на word-embedding – предлагается совместить данные подходы.

Попытки совмещения word-embedding и TF-IDF предпринимались и ранее. Суть таких улучшений состоит в использовании при усреднении эмбеддингов слов весовых коэффициентов, соответствующих TF-IDF этих слов. Однако описанная выше во Введении неоднозначность не устраняется, поскольку размерность итогового вектора остается неизменной.

Далее приводится алгоритм формирования векторного представления текста в рамках предлагаемого подхода. При этом необходимо учитывать, что векторизация текста обычно выполняется в рамках решения какой-либо прикладной задачи. Приведенный алгоритм актуален для решения задачи ранжирования текстов (описана во Введении).

1. Имеется исходное множество текстов T, на которых требуется выполнять ранжирование, то есть для заданного текста t_i из T упорядочить множество T по степени близости к t_i. Также имеется предобученная word-embedding модель M.
2. На множестве T строится словарь V всех слов, входящих в тексты t_i из T.
3. Для каждого слова v_i из словаря V получаем его эмбеддинг при помощи предобученной модели M: $e_i = M(v_i)$. Все e_i в совокупности составляют множество эмбеддингов E.
4. Выполняется кластеризация на множестве E, в результате которой получается кластеризующая модель C, которая по эмбеддингу слова выдает кластер, к которому он относится. Для каждого e_i из E определяется его кластер $c_i=C(e_i)$. Множество всех кластеров c_i модели C также обозначим символом C.
5. Для каждого текста t_i из T получаем новый текст t^c_i следующим способом:
   5.1. Копируем t_i в t^c_i;
   5.2. Для каждого слова w_j из текста t^c_i
      a) определяем его кластер $c_j = C(M(w_j))$;
      b) заменяем в тексте t^c_i слово w_j на номер соответствующего кластера c_j.

   В результате путем замены всех t_i из T на t^c_i получено новое множество T_c. Все тексты этого множества состоят из номеров кластеров с символами-разделителями между ними.
6. Выполняется векторизация текстов T_c при помощи TF-IDF на n-граммах слов. В результате чего получается:
   - множество X_c TF-IDF-векторов x^c_i для каждого t^c_i из T_c,
   - словарь n-грамм V_c, а также
   - N_{min} и N_{max} — заданные в качестве входных параметров алгоритма минимальная и максимальная длины n-грамм, используемых для построения словаря TF-IDF.
7. Для каждого t_i из T x^c_i из X_c далее рассматривается как векторное представление текста t_i.

Поскольку объединенные в один кластер одни слова исходного словаря могут быть ближе друг к другу, чем другие, обогащение векторного представления информацией о взаимной близости слов кластера может повысить качество этого векторного представления. Чтобы учесть в векторном представлении текстов взаимную близость слов друг к другу в рамках одного кластера, предлагается следующее улучшение подхода в виде дополнительных шагов алгоритма.

Формирование обогащающих векторов слов.

8. По каждому кластеру c_i из C.
   8.1. Для каждого относящегося к c_i эмбеддинга e_j слова w_j исходного словаря V выполняется снижение размерности до одного (располагаются на одной числовой оси). Полученные в результате числовые значения обозначим через e^r_j.
   8.2. С использованием min-max–нормализации выполняется масштабирование числовых представлений e^r_j эмбеддингов e_j слов w_j кластера c_i на отрезок [0;1]. Отрезок [0;1] разбивается на D-1 частей (D — размерность обогащающего вектора слова, задается в качестве одного из входных параметров алгоритма), пронумерованных от 1 до D-1.

8.3. Для каждого слова w_j исходного словаря V выполняется формирование обогащающего вектора e^e_j следующим образом. Нулевая позиция вектора всегда заполняется значением «1». Далее по каждому из D-1 отрезков, на которые разбит интервал [0;1], если e^r_j попадает в k-й интервал, то k-я позиция вектора заполняется значением «1», иначе «0».

Формирование обогащенных векторов текстов.

9. Для каждого текста t^c_i из T_c.

Каждая позиция (обозначим индексом j) TF-IDF вектора x^c_i соответствует TF-IDF-коэффициенту x^c_{ij} для n-граммы w^c_j из словаря V_c. Для каждой j-й позиции вектора x^c_i сформируем обогащающий вектор e^{xc}_j следующим образом.

9.1. На основе текста t_i из T сформируем обогащающие векторы всех входящих в t_i n-грамм длиной от N_{min} до N_{max} путем конкатенации обогащающих векторов входящих в них слов (полученных на шаге 8.3). Максимальная длина каждого такого вектора равна N_{max}*D. Если вектор получен из n-граммы длиной меньше N_{max}, вектор дополняется справа нулями до максимальной длины.

9.2. Каждой n-грамме из t_i соответствует некоторая n-грамма из V_c. Выполним усреднение обогащающих векторов n-грамм из t_i по соответствующим им n-граммам из V_c. Таким образом, получены обогащающие векторы для тех позиций вектора x^c_i, для которых соответствующие n-граммы w^c_j из V_c входят в t^c_i.

9.3. Если соответствующие j-й позиции из x^c_i n-граммы отсутствуют в t^c_j, то соответствующие обогащающие векторы состоят из N_{max}*D нулевых элементов.

9.4. Обогащенный вектор x^{ce}_i текста t_i вычисляется конкатенацией векторов x^c_{ij} *e^{xc}_j (произведение TF-IDF-коэффициента n-граммы w^c_j на ее обогащающий вектор).

Таким образом, получено множество обогащенных векторных представлений X_{ce} текстов T.

2. Апробация подхода

2.1. Условия проведения апробации

Проверка качества подхода по векторизации текстов осуществлялась в контексте решения задачи ранжирования текстов по семантической близости.

Для проверки качества подхода имелось в распоряжении 13772 примера вида: (T_1, T_2, T), где T_1, T_2 и T — тексты, такие что T_1 ближе к T, чем T_2.

Все тексты из 13772 примеров сформированы на основе 940 текстов, представляющих собой внутреннюю переписку в Банке России.

Качество подхода по ранжированию текстов определяется по точности (Accuracy) как отношение количества корректно определенных примеров к общему их количеству.

Распределение длин текстов корпуса представлено на рисунке 1. По горизонтальной оси отложены длины текстов в символах. По вертикальной — количество текстов заданной длины.

Тексты корпуса предварительно были приведены к нижнему регистру, из них были отфильтрованы символы, не являющиеся пробелом или буквами русского или латинского алфавитов.

Размер словаря корпуса текстов составляет 73731 слово.

В качестве меры, при помощи которой определяется близость между векторными представлениями текстов, использовалась косинусная близость.

2.2. Определение оптимальных параметров

2.2.1. Параметры алгоритма

В рамках предлагаемого подхода:

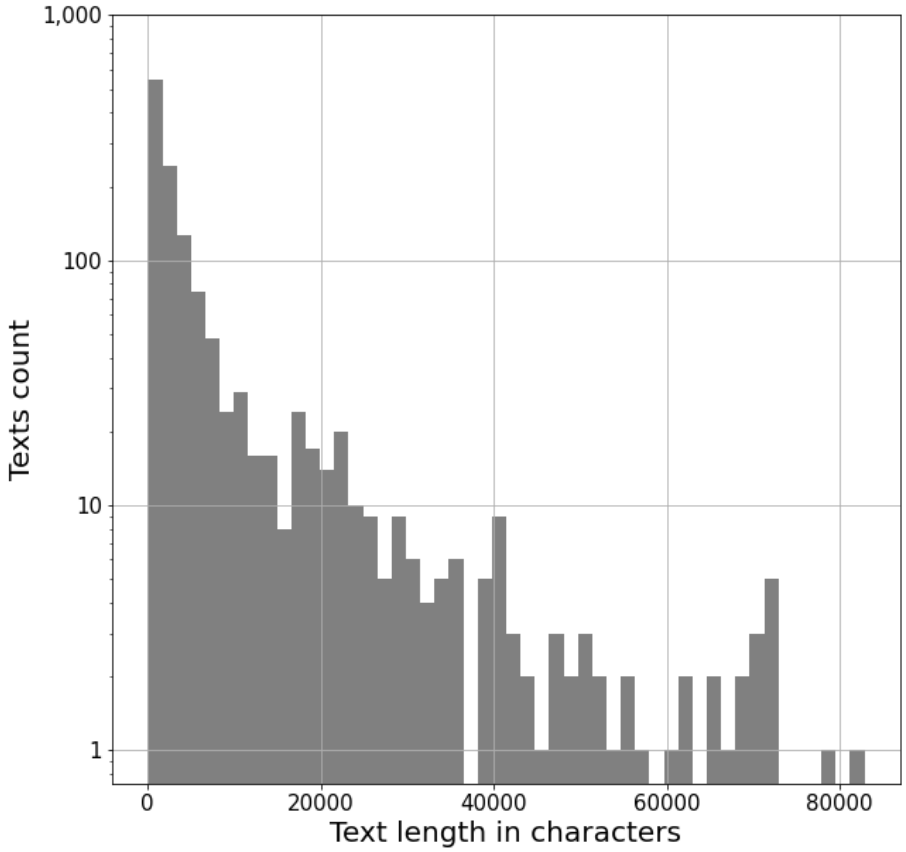


Fig. 1. Distribution of text lengths in the corpus

Рис. 1. Распределение длин текстов в корпусе

- в качестве word-embedding модели использовалась FastText модель от DeepPavlov[1] , обученная на русскоязычной части Wikipedia совместно с набором данных Lenta ru news. Размерность выходного вектора 300. Использованная в работе модель Fasttext формирует эмбеддинг слова с использованием входящих в это слово символьных последовательностей, что позволяет применять ее к словам out-of-vocabulary, то есть таким, которые в обучении FastText модели не участвовали.

- в качестве метода кластеризации использован Kmeans,

- в качестве метода снижения размерности на шаге 8.1 использован t-SNE,

- при векторизации TF-IDF устанавливается фильтр на минимальное количество документов, в которых встречается n-грамма, равная двум.

Изменяемыми параметрами для алгоритма являются: количество кластеров, диапазон n-грамм, размерность обогащающего вектора слова.

Применим предлагаемый подход для всех возможных комбинаций параметров из представленных в Таблице 1, выполнив серию экспериментов по три для каждой комбинации параметров. Здесь размерность обогащающего вектора слова, равная нулю, означает, что обогащение не производится, а используются непосредственно TF-IDF-векторы.

[1]http://files.deeppavlov.ai/embeddings/ft_native_300_ru_wiki_lenta_lower_case/ft_native_300_ru_wiki_lenta_lower_case.bin

Table 1. The values of the parameters to be checked

Таблица 1. Проверяемые значения параметров алгоритма

Параметр	**Значения**
Количество кластеров	100, 1000, 3000, 5000, 10000, 15000, 20000, 25000, 45000, 65000
Размерность обогащающего вектора слова	0, 2, 5, 10, 15, 20
Диапазон n-грамм	(1,1), (1,2), (1,3), (1,4)

2.2.2. Результаты

Таблица с полными результатами экспериментов приведена в Приложении А.

Значения параметров, на которых получены лучшие показатели точности для различных диапазонов n-грамм, представлены в Таблице 2.

Из таблицы видно, что лучшие результаты получены для следующей комбинации параметров: количество кластеров 25000, размер обогащающего вектора слова 2, диапазон n-грамм (1, 4).

Table 2. Parameters that give the best results

Таблица 2. Параметры с лучшими результатами

диапазон n-gram	**Количество кластеров**	**Размерность обогащающего вектора слова**	**Точность**	**Стандартное отклонение**
(1, 1)	20000	0	0.934	1.321
(1, 2)	25000	0	0,940	1.151
(1, 3)	25000	2	0.944	1.156
(1, 4)	25000	2	**0.947**	1.16

2.2.3. Интерпретация результатов

Отметим, что значение в 25000 кластеров, на котором получено лучшее значение точности, составляет приблизительно одну третью часть от размера словаря корпуса текстов (73731 слово).

Ниже приведены графики зависимости точности от параметров алгоритма.

На Рисунке 2 приведен график зависимости точности от количества кластеров для разной размерности обогащающего вектора слова. Значения точности для графика получены усреднением точности для всех диапазонов n-грамм.

Из графика видно, что увеличение количества кластеров до определенного значения приводит к значительному повышению точности. При количестве кластеров 25000 точность достигает максимального значения, после чего начинает убывать.

Завершается график горизонтальным участком, что можно объяснить следующим: чем ближе параметр «количество кластеров» при кластеризации к количеству кластеризуемых объектов, тем больше получается «пустых» кластеров, то есть, несмотря на увеличение значения параметра «количество кластеров», словарь V_c описанного алгоритма растет незначительно.

Зависимость точности от размерности обогащающего вектора слова имеет разный характер, в зависимости от количества кластеров, в связи с чем для зависимости точности от размерности обогащающего вектора слова приводятся два графика: график для количества кластеров до 20000 приведен на Рисунке 3, а график для количества кластеров 20000 и более приведен на Рисунке 4. Значения точности для графиков получены усреднением точности для всех диапазонов n-грамм.

Из графиков видно, что чем меньше количество кластеров, тем больший прирост точности дает процедура обогащения, а начиная с определенного количества кластеров, в целом, обогащение понижает точность. При этом наиболее существенный прирост точности наблюдается на размерности, равной двум. Также видно, что при использовании обогащения максимальный прирост точности на малом количестве кластеров больше, чем максимальное снижение на большом.

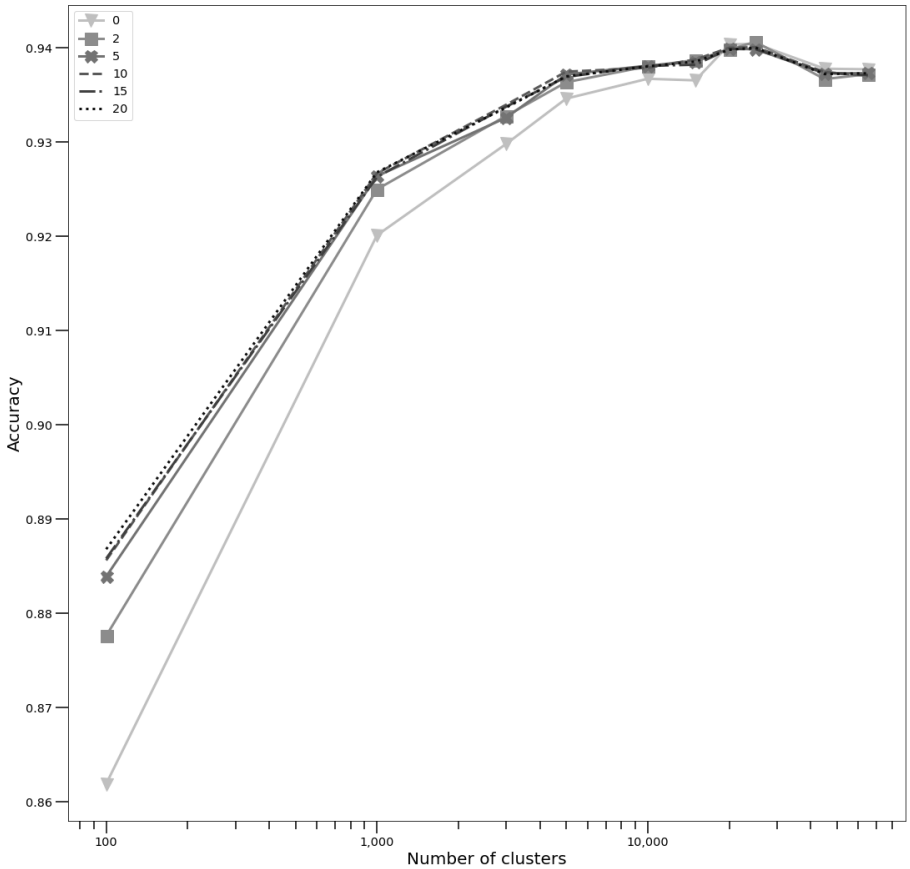


Fig. 2. Dependence of accuracy on the number of clusters

Рис. 2. Зависимость точности от количества кластеров

График зависимости точности от количества кластеров для различных диапазонов используемых n-грамм приведен на Рисунке 5. Значения точности для графика получены усреднением точности для всех размерностей обогащающего вектора. Из графика видно, что чем больше используемый диапазон n-грамм, тем точность выше.

2.3. Сравнение с baseline-подходами

Сравниваются следующие подходы:
- предлагаемый в настоящей работе (обогащенные TF-IDF-векторы на номерах кластеров),
- TF-IDF на текстах корпуса,
- TF-IDF на лемматизированных текстах корпуса,
- усреднение эмбеддингов,
- усреднение эмбеддингов, взвешенных по TF-IDF.

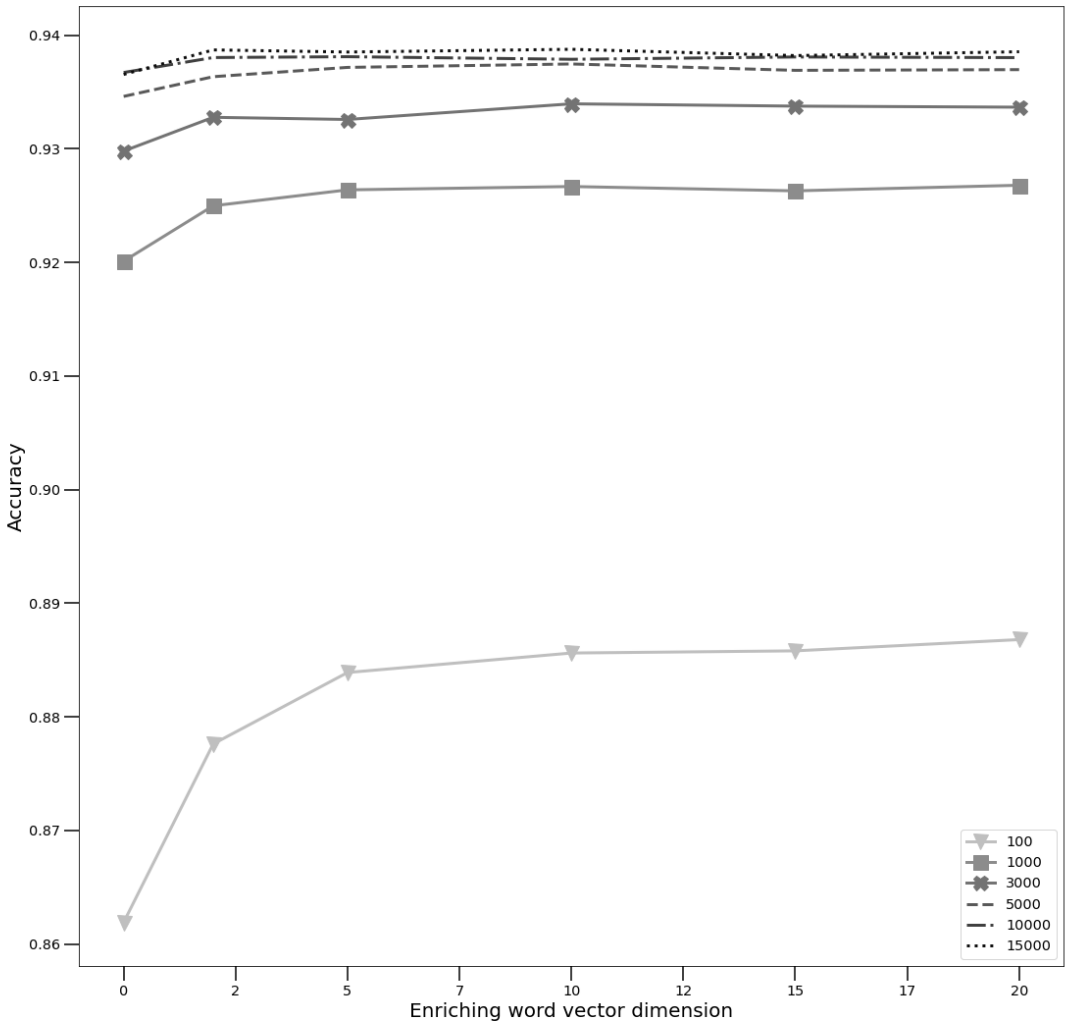


Fig. 3. Dependence of accuracy on enriching word vector for the number of clusters less than 20,000

Рис. 3. Зависимость точности от размерности обогащающего вектора слова для количества кластеров менее 20000

2.3.1. TF-IDF

Векторизация текстов на основе TF-IDF для следующих диапазонов n-грамм: (1, 1), (1, 2), (1, 3), (1, 4).

Векторизация TF-IDF осуществлялась с использованием библиотеки sklearn.

При векторизации TF-IDF установлен фильтр на минимальное количество документов, в которых встречается n-грамма, равная двум.

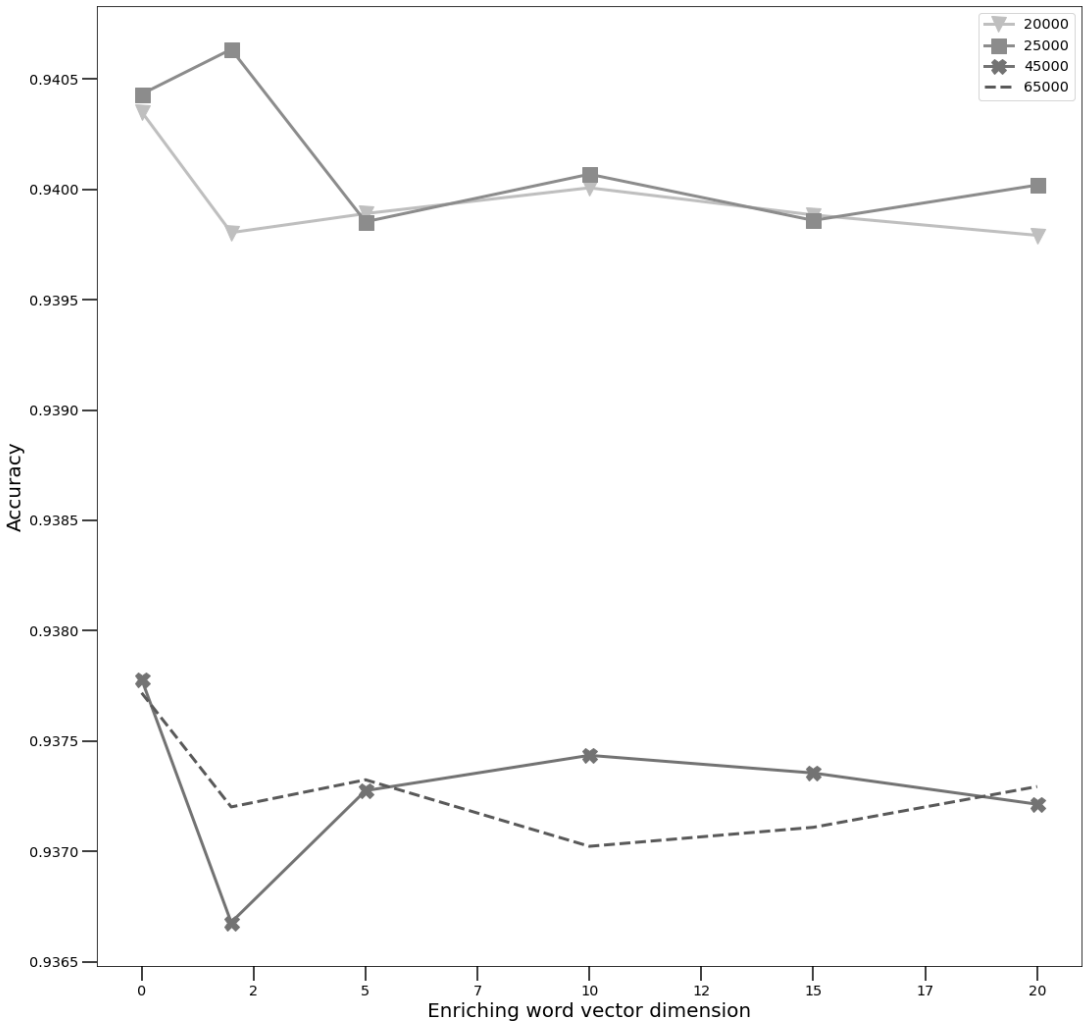


Fig. 4. Dependence of accuracy on the dimension of the enriching word vector for the number of clusters of 20,000 or more

Рис. 4. Зависимость точности от размерности обогащающего вектора слова для количества кластеров 20000 и более

2.3.2. TF-IDF с лемматизацией

Практика решения задач обработки естественного языка показывает, что поскольку в русском языке разные формы слова часто имеют разное написание, непосредственное применение подхода TF-IDF для векторизации текстов в большинстве случаев работает хуже, чем в случае, если выполнено предварительное приведение слов текста в нормальную форму.

В эксперименте приведение слов в нормальную форму выполнялось путем лемматизации при помощи библиотеки Mystem от компании Яндекс.

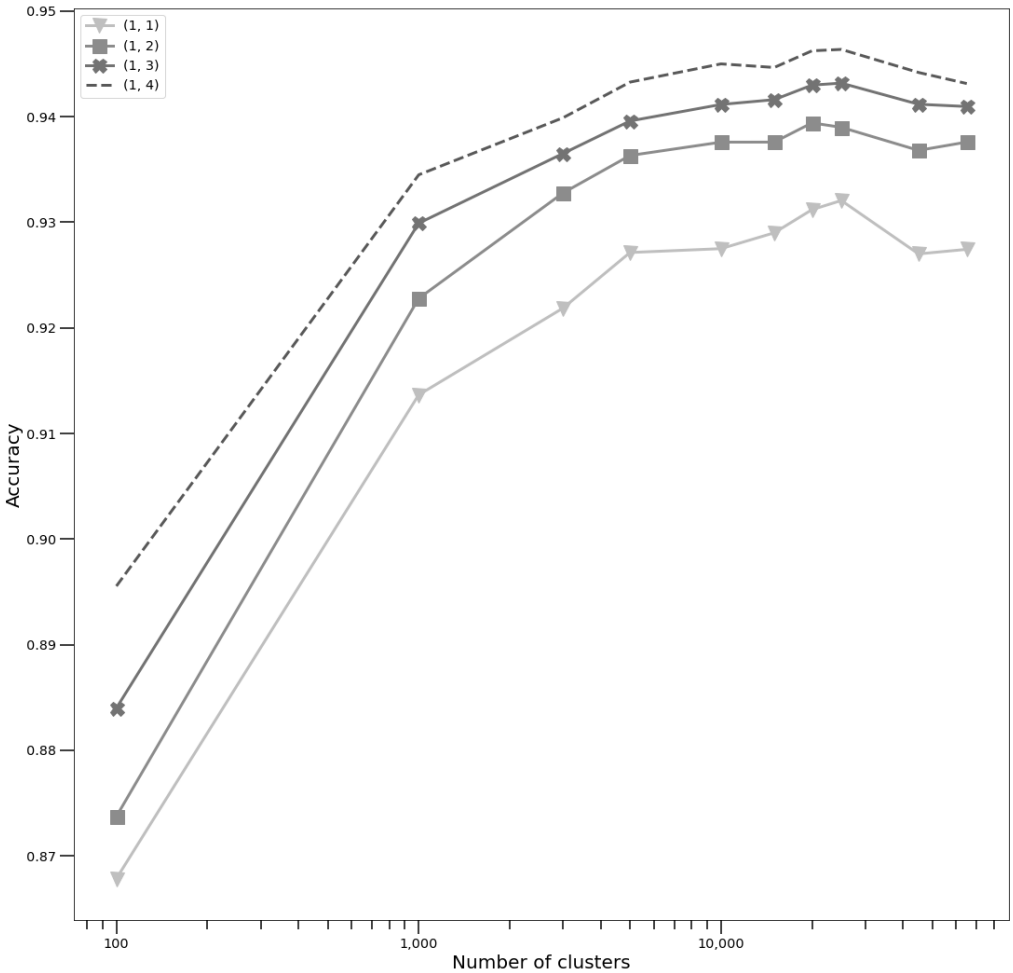


Fig. 5. Dependence of accuracy on the number of clusters for different n-gram ranges

Рис. 5. Зависимость точности от количества кластеров для разных диапазонов n-грамм

Векторизация текстов на основе TF-IDF выполнялась для следующих видов n-грамм: (1, 1), (1, 2), (1, 3), (1, 4).

Векторизация TF-IDF осуществлялась с использованием библиотеки sklearn.

При векторизации TF-IDF установлен фильтр на минимальное количество документов, в которых встречается n-грамма, равная двум.

2.3.3. Усреднение эмбеддингов

Усреднение эмбеддингов для текста заключается в преобразовании каждого слова текста в векторное представление при помощи word-embedding модели и последующем усреднении эмбеддингов слов с получением итогового вектора текста той же размерности, что и векторы эмбеддингов.

2.3.4. Усреднение эмбеддингов, взвешенных по TF-IDF

Подход отличается от простого усреднения эмбеддингов тем, что при усреднении эмбеддинги слов берутся с весовыми коэффициентами, соответствующими их TF-IDF-коэффициентам, рассчитанным на имеющемся корпусе текстов [10]. В настоящей работе коэффициенты TF-IDF рассчитывались на всем корпусе документов.

2.3.5. Результаты сравнения подходов

Результаты сравнения подходов приведены в таблице 3.

Table 3. Methods comparison **Таблица 3.** Сравнение подходов

Диапазон n-грамм	**TF-IDF**	**TF-IDF со стеммингом**	**Усреднение эмбеддингов Fasttext**	**Усреднение эмбеддингов Fasttext со взвешиванием по TD-IDF**	**Обогащенные TF-IDF векторы на кластерах**
(1,1)	0,9287	0,9279	0,779	0,8679	0.9302
(1,2)	0,9333	0,9364	не применимо	не применимо	0.9405
(1,3)	0,9366	0,9396	не применимо	не применимо	0.9459
(1,4)	0,9381	0,94	не применимо	не применимо	0.947

Заключение

В настоящей работе предложен подход по векторизации текстов при помощи word-embedding моделей, экспериментально определены оптимальные параметры для решения задачи ранжирования текстов, представляющих собой 940 писем внутренней переписки Банка России, выполнено сравнение предложенного подхода с распространенными подходами.

Предлагаемый подход в сравнении с представленными baseline-подходами показывает лучшие результаты.

Использовавшийся в предлагаемом подходе алгоритм Kmeans для кластеризации эмбеддингов элементов словаря в качестве входного параметра принимает количество кластеров. При этом остается открытым вопрос, насколько компактными получаются кластеры, что может влиять на качество представления текста при помощи кластеров.

Для решения данной проблемы видятся перспективными следующие направления дальнейших исследований:

- Анализ компактности получаемых кластеров для улучшения представления текстов в виде номеров кластеров, например, фильтрация элементов кластера по порогу расстояния до центроида.
- Использование такого подхода для кластеризации, при котором параметры определяют расстояния между объектами, а не количество кластеров. Однако, по сравнению с Kmeans, такие подходы более требовательны к вычислительным ресурсам. Соответственно, в контексте данного направления актуально решение задачи поиска эффективного способа кластеризации элементов словаря.

Предложенный подход для векторизации апробирован в рамках решения задачи ранжирования текстов. Необходимо исследовать подход на применимость для решения других задач обработки естественного языка.

Appendix A. Experimental results
Приложение А. Результаты экспериментов

№ п/п	Количество кластеров	Диапазон n-грамм	Размерность вектора обогащения слова	Точность	Стандартное отклонение
1	100	(1, 1)	0	0,858	1,051
2	100	(1, 1)	2	0,863	1,058
3	100	(1, 1)	5	0,87	1,066
4	100	(1, 1)	10	0,871	1,067
5	100	(1, 1)	15	0,871	1,067
6	100	(1, 1)	20	0,873	1,069
7	100	(1, 2)	0	0,86	1,054
8	100	(1, 2)	2	0,871	1,066
9	100	(1, 2)	5	0,877	1,074
10	100	(1, 2)	10	0,878	1,075
11	100	(1, 2)	15	0,878	1,075
12	100	(1, 2)	20	0,879	1,076
13	100	(1, 3)	0	0,862	1,056
14	100	(1, 3)	2	0,882	1,08
15	100	(1, 3)	5	0,888	1,087
16	100	(1, 3)	10	0,89	1,09
17	100	(1, 3)	15	0,891	1,091
18	100	(1, 3)	20	0,892	1,092
19	100	(1, 4)	0	0,867	1,062
20	100	(1, 4)	2	0,895	1,096
21	100	(1, 4)	5	0,901	1,104
22	100	(1, 4)	10	0,903	1,106
23	100	(1, 4)	15	0,903	1,106
24	100	(1, 4)	20	0,904	1,107
25	1000	(1, 1)	0	0,91	1,115
26	1000	(1, 1)	2	0,912	1,118
27	1000	(1, 1)	5	0,915	1,121
28	1000	(1, 1)	10	0,915	1,12
29	1000	(1, 1)	15	0,915	1,12
30	1000	(1, 1)	20	0,915	1,12
31	1000	(1, 2)	0	0,918	1,125
32	1000	(1, 2)	2	0,923	1,13
33	1000	(1, 2)	5	0,924	1,131
34	1000	(1, 2)	10	0,924	1,132
35	1000	(1, 2)	15	0,924	1,131
36	1000	(1, 2)	20	0,924	1,132
37	1000	(1, 3)	0	0,924	1,132
38	1000	(1, 3)	2	0,93	1,139
39	1000	(1, 3)	5	0,931	1,14

40	1000	(1, 3)	10	0,932	1,141
41	1000	(1, 3)	15	0,931	1,14
42	1000	(1, 3)	20	0,931	1,141
43	1000	(1, 4)	0	0,928	1,136
44	1000	(1, 4)	2	0,935	1,145
45	1000	(1, 4)	5	0,936	1,146
46	1000	(1, 4)	10	0,936	1,147
47	1000	(1, 4)	15	0,936	1,146
48	1000	(1, 4)	20	0,937	1,147
49	3000	(1, 1)	0	0,92	1,126
50	3000	(1, 1)	2	0,921	1,128
51	3000	(1, 1)	5	0,922	1,129
52	3000	(1, 1)	10	0,923	1,131
53	3000	(1, 1)	15	0,923	1,13
54	3000	(1, 1)	20	0,923	1,13
55	3000	(1, 2)	0	0,929	1,138
56	3000	(1, 2)	2	0,933	1,142
57	3000	(1, 2)	5	0,932	1,142
58	3000	(1, 2)	10	0,934	1,144
59	3000	(1, 2)	15	0,934	1,144
60	3000	(1, 2)	20	0,934	1,144
61	3000	(1, 3)	0	0,934	1,143
62	3000	(1, 3)	2	0,937	1,148
63	3000	(1, 3)	5	0,936	1,147
64	3000	(1, 3)	10	0,937	1,148
65	3000	(1, 3)	15	0,938	1,148
66	3000	(1, 3)	20	0,937	1,148
67	3000	(1, 4)	0	0,937	1,147
68	3000	(1, 4)	2	0,94	1,151
69	3000	(1, 4)	5	0,94	1,151
70	3000	(1, 4)	10	0,941	1,152
71	3000	(1, 4)	15	0,941	1,152
72	3000	(1, 4)	20	0,941	1,152
73	5000	(1, 1)	0	0,926	1,134
74	5000	(1, 1)	2	0,927	1,135
75	5000	(1, 1)	5	0,927	1,136
76	5000	(1, 1)	10	0,928	1,137
77	5000	(1, 1)	15	0,927	1,135
78	5000	(1, 1)	20	0,927	1,136
79	5000	(1, 2)	0	0,934	1,144
80	5000	(1, 2)	2	0,936	1,147
81	5000	(1, 2)	5	0,937	1,148
82	5000	(1, 2)	10	0,937	1,148
83	5000	(1, 2)	15	0,937	1,147
84	5000	(1, 2)	20	0,936	1,147

85	5000	(1, 3)	0	0,937	1,148
86	5000	(1, 3)	2	0,94	1,151
87	5000	(1, 3)	5	0,94	1,151
88	5000	(1, 3)	10	0,941	1,152
89	5000	(1, 3)	15	0,94	1,151
90	5000	(1, 3)	20	0,94	1,151
91	5000	(1, 4)	0	0,941	1,152
92	5000	(1, 4)	2	0,943	1,155
93	5000	(1, 4)	5	0,944	1,156
94	5000	(1, 4)	10	0,944	1,156
95	5000	(1, 4)	15	0,944	1,156
96	5000	(1, 4)	20	0,944	1,156
97	10000	(1, 1)	0	0,928	1,137
98	10000	(1, 1)	2	0,927	1,135
99	10000	(1, 1)	5	0,928	1,137
100	10000	(1, 1)	10	0,927	1,136
101	10000	(1, 1)	15	0,927	1,136
102	10000	(1, 1)	20	0,927	1,136
103	10000	(1, 2)	0	0,936	1,147
104	10000	(1, 2)	2	0,938	1,149
105	10000	(1, 2)	5	0,938	1,148
106	10000	(1, 2)	10	0,938	1,149
107	10000	(1, 2)	15	0,938	1,149
108	10000	(1, 2)	20	0,938	1,148
109	10000	(1, 3)	0	0,94	1,151
110	10000	(1, 3)	2	0,942	1,153
111	10000	(1, 3)	5	0,941	1,153
112	10000	(1, 3)	10	0,941	1,153
113	10000	(1, 3)	15	0,941	1,153
114	10000	(1, 3)	20	0,942	1,153
115	10000	(1, 4)	0	0,943	1,155
116	10000	(1, 4)	2	0,945	1,158
117	10000	(1, 4)	5	0,945	1,158
118	10000	(1, 4)	10	0,945	1,158
119	10000	(1, 4)	15	0,946	1,158
120	10000	(1, 4)	20	0,945	1,158
121	15000	(1, 1)	0	0,928	1,136
122	15000	(1, 1)	2	0,93	1,139
123	15000	(1, 1)	5	0,929	1,138
124	15000	(1, 1)	10	0,929	1,138
125	15000	(1, 1)	15	0,929	1,138
126	15000	(1, 1)	20	0,929	1,138
127	15000	(1, 2)	0	0,936	1,146
128	15000	(1, 2)	2	0,938	1,149
129	15000	(1, 2)	5	0,938	1,149

130	15000	(1, 2)	10	0,938	1,149
131	15000	(1, 2)	15	0,938	1,148
132	15000	(1, 2)	20	0,938	1,149
133	15000	(1, 3)	0	0,94	1,151
134	15000	(1, 3)	2	0,942	1,154
135	15000	(1, 3)	5	0,942	1,154
136	15000	(1, 3)	10	0,942	1,154
137	15000	(1, 3)	15	0,942	1,153
138	15000	(1, 3)	20	0,942	1,154
139	15000	(1, 4)	0	0,943	1,155
140	15000	(1, 4)	2	0,945	1,157
141	15000	(1, 4)	5	0,945	1,157
142	15000	(1, 4)	10	0,945	1,158
143	15000	(1, 4)	15	0,945	1,157
144	15000	(1, 4)	20	0,945	1,158
145	20000	(1, 1)	0	0,934	1,143
146	20000	(1, 1)	2	0,931	1,14
147	20000	(1, 1)	5	0,931	1,14
148	20000	(1, 1)	10	0,931	1,14
149	20000	(1, 1)	15	0,931	1,14
150	20000	(1, 1)	20	0,93	1,14
151	20000	(1, 2)	0	0,94	1,151
152	20000	(1, 2)	2	0,939	1,15
153	20000	(1, 2)	5	0,939	1,15
154	20000	(1, 2)	10	0,94	1,151
155	20000	(1, 2)	15	0,939	1,15
156	20000	(1, 2)	20	0,939	1,15
157	20000	(1, 3)	0	0,943	1,154
158	20000	(1, 3)	2	0,943	1,155
159	20000	(1, 3)	5	0,943	1,155
160	20000	(1, 3)	10	0,943	1,155
161	20000	(1, 3)	15	0,943	1,155
162	20000	(1, 3)	20	0,943	1,155
163	20000	(1, 4)	0	0,945	1,158
164	20000	(1, 4)	2	0,946	1,159
165	20000	(1, 4)	5	0,946	1,159
166	20000	(1, 4)	10	0,946	1,159
167	20000	(1, 4)	15	0,946	1,159
168	20000	(1, 4)	20	0,946	1,159
169	25000	(1, 1)	0	0,933	1,143
170	25000	(1, 1)	2	0,932	1,142
171	25000	(1, 1)	5	0,932	1,141
172	25000	(1, 1)	10	0,932	1,141
173	25000	(1, 1)	15	0,931	1,141
174	25000	(1, 1)	20	0,932	1,141

175	25000	(1, 2)	0	0,94	1,151
176	25000	(1, 2)	2	0,939	1,15
177	25000	(1, 2)	5	0,938	1,149
178	25000	(1, 2)	10	0,939	1,15
179	25000	(1, 2)	15	0,939	1,15
180	25000	(1, 2)	20	0,939	1,15
181	25000	(1, 3)	0	0,943	1,154
182	25000	(1, 3)	2	0,944	1,156
183	25000	(1, 3)	5	0,943	1,155
184	25000	(1, 3)	10	0,943	1,155
185	25000	(1, 3)	15	0,943	1,155
186	25000	(1, 3)	20	0,943	1,155
187	25000	(1, 4)	0	0,946	1,158
188	25000	(1, 4)	2	0,947	1,16
189	25000	(1, 4)	5	0,946	1,159
190	25000	(1, 4)	10	0,947	1,16
191	25000	(1, 4)	15	0,946	1,159
192	25000	(1, 4)	20	0,946	1,159
193	45000	(1, 1)	0	0,93	1,139
194	45000	(1, 1)	2	0,926	1,134
195	45000	(1, 1)	5	0,927	1,135
196	45000	(1, 1)	10	0,927	1,135
197	45000	(1, 1)	15	0,927	1,135
198	45000	(1, 1)	20	0,927	1,135
199	45000	(1, 2)	0	0,938	1,148
200	45000	(1, 2)	2	0,936	1,146
201	45000	(1, 2)	5	0,937	1,148
202	45000	(1, 2)	10	0,937	1,147
203	45000	(1, 2)	15	0,937	1,147
204	45000	(1, 2)	20	0,937	1,147
205	45000	(1, 3)	0	0,941	1,153
206	45000	(1, 3)	2	0,941	1,153
207	45000	(1, 3)	5	0,941	1,153
208	45000	(1, 3)	10	0,941	1,153
209	45000	(1, 3)	15	0,941	1,153
210	45000	(1, 3)	20	0,941	1,153
211	45000	(1, 4)	0	0,943	1,155
212	45000	(1, 4)	2	0,944	1,156
213	45000	(1, 4)	5	0,944	1,157
214	45000	(1, 4)	10	0,945	1,157
215	45000	(1, 4)	15	0,945	1,157
216	45000	(1, 4)	20	0,944	1,157
217	65000	(1, 1)	0	0,93	1,139
218	65000	(1, 1)	2	0,927	1,135
219	65000	(1, 1)	5	0,927	1,135

220	65000	(1, 1)	10	0,927	1,135
221	65000	(1, 1)	15	0,927	1,135
222	65000	(1, 1)	20	0,927	1,136
223	65000	(1, 2)	0	0,938	1,148
224	65000	(1, 2)	2	0,938	1,148
225	65000	(1, 2)	5	0,938	1,149
226	65000	(1, 2)	10	0,937	1,148
227	65000	(1, 2)	15	0,937	1,148
228	65000	(1, 2)	20	0,938	1,148
229	65000	(1, 3)	0	0,941	1,152
230	65000	(1, 3)	2	0,941	1,152
231	65000	(1, 3)	5	0,941	1,153
232	65000	(1, 3)	10	0,941	1,152
233	65000	(1, 3)	15	0,941	1,152
234	65000	(1, 3)	20	0,941	1,153
235	65000	(1, 4)	0	0,943	1,154
236	65000	(1, 4)	2	0,943	1,155
237	65000	(1, 4)	5	0,943	1,155
238	65000	(1, 4)	10	0,943	1,155
239	65000	(1, 4)	15	0,943	1,155
240	65000	(1, 4)	20	0,943	1,155

References

[1] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability", vol. 1, 2019, pp. 1–4. DOI: 10.1109/AITB48515.2019.8947433.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008, ISBN: 0521865719.

[3] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, "Representation Learning for Very Short Texts Using Weighted Word Embedding Aggregation", *Pattern Recogn. Lett.*, vol. 80, no. C, pp. 150–156, Sep. 2016, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2016.06.012.

[4] G. Kim and K. Cho, "Length-Adaptive Transformer: Train Once with Length Drop, Use Anytime with Search", *ArXiv*, vol. abs/2010.07003, 2020.

[5] O. Zafrir, G. Boudoukh, P. Izsak, and M. Wasserblat, "Q8BERT: Quantized 8Bit BERT", *ArXiv*, vol. abs/1910.06188, 2019.

[6] H. Gong, Y. Shen, D. Yu, J. Chen, and D. Yu, "Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 6751–6761. DOI: 10.18653/v1/2020.acl-main.603. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.603.

[7] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, "When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 529–535. DOI: 10.18653/v1/N18-2084. [Online]. Available: https://www.aclweb.org/anthology/N18-2084.

[8] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 440–450. DOI: 10.18653/v1/P18-1041. [Online]. Available: https://www.aclweb.org/anthology/P18-1041.

[9] A. Rücklé, S. Eger, M. Peyrard, and I. Gurevych, "Concatenated p-mean Word Embeddings as Universal Cross-Lingual Sentence Representations", *ArXiv*, vol. abs/1803.01400, 2018.

[10] P. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics", *Journal of Artificial Intelligence Research*, vol. 37, pp. 141–188, Mar. 2010. DOI: 10.1613/jair.2934.

[11] A. L. O. Shahmirzadi and K. Younge, "Text Similarity in Vector Space Models: A Comparative Study", in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 659–666. DOI: 10.1109/ICMLA.2019.00120.

[12] V. Gupta, A. Kumar, P. Nokhiz, H. Gupta, and P. Talukdar, "Improving Document Classification with Multi-Sense Embeddings", in *24th European Conference on Artificial Intelligence - ECAI 2020*, Nov. 2020, pp. 2030–2037. DOI: 10.3233/FAIA200324.

[13] V. Mekala Dheeraj and Gupta, B. Paranjape, and H. Karnick, "SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations", in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 659–669. DOI: 10.18653/v1/D17-1069. [Online]. Available: https://www.aclweb.org/anthology/D17-1069.

[14] V. Gupta, H. Karnick, A. Bansal, and P. Jhala, "Product Classification in E-Commerce using Distributional Semantics", in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 536–546. [Online]. Available: https://www.aclweb.org/anthology/C16-1052.

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

ERRATUM

Corrigendum to: V. V. Vasilchikov, “Parallel Algorithm for Solving the Graph Isomorphism Problem”, Modeling and analysis of information systems, vol. 27, no. 1, pp. 86–94, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-1-86-94

V. V. Vasilchikov[1]

DOI: 10.18255/1818-1015-2021-3-312-313

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68W10
Research article
Full text in Russian

Received August 25, 2021
After revision August 25, 2021
Accepted August 25, 2021

In the article by V. V. Vasilchikov “Parallel Algorithm for Solving the Graph Isomorphism Problem” (Modeling and analysis of information systems, vol. 27, no. 1, pp. 86–94, 2020; DOI: https://doi.org/10.18255/1818-1015-2020-1-86-94) there was a misprint in the layout. In the Table 1, in the last column of the row “Degree of graph” the value should be 3000 (instead of 300). The corrected “Table 1” is shown below. The editors apologise for the inconvenience.

Keywords: graph isomorphism problem; parallel algorithm; recursion; .NET

INFORMATION ABOUT THE AUTHORS

Vladimir Vasilyevich Vasilchikov | orcid.org/0000-0001-7882-8906. E-mail: vvv193@mail.ru
correspondence author | PhD.

For citation: V. V. Vasilchikov, “Corrigendum to: V. V. Vasilchikov, “Parallel Algorithm for Solving the Graph Isomorphism Problem”, Modeling and analysis of information systems, vol. 27, no. 1, pp. 86–94, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-1-86-94”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 312-313, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

ERRATUM

Исправление к статье: В. В. Васильчиков, «Параллельный алгоритм решения задачи об изоморфизме графов», Моделирование и анализ информационных систем, Том 27, №1, с. 86–94, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-1-86-94

В. В. Васильчиков[1]

DOI: 10.18255/1818-1015-2021-3-312-313

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 519.688: 519.85
Научная статья
Полный текст на русском языке

Получена 25 августа 2021 г.
После доработки 25 августа 2021 г.
Принята к публикации 25 августа 2021 г.

В статье В.В. Васильчикова «Параллельный алгоритм решения задачи об изоморфизме графов» (Моделирование и анализ информационных систем, том 27, №1, с. 86–94, 2020; https://doi.org/10.18255/1818-1015-2020-1-86-94) при вёрстке была допущена опечатка. В таблице 1 в последнем столбце строки «Степень графа» должно быть значение 3000 (вместо 300). Исправленная «Таблица 1» приводится ниже. Редакция приносит извинения за доставленные неудобства.

Ключевые слова: изоморфизм графов; параллельный алгоритм; рекурсия; .NET

Table 1. Execution time of sequential algorithm for regular graphs on 7000 vertices

Таблица 1. Длительность работы последовательного алгоритма для регулярных графов на 7000 вершин

Степень графа	**2**	**3**	**5**	**20**	**200**	**3000**
Диаметр графа	∞	16	9	4	3	2
Все вычисления (с)	*941.11*	*3149.81*	*3240.97*	*2381.29*	*2794.24*	*2243.00*
Выч. инвариантов	932.41	3149.22	3240.36	2380.51	2792.84	2241.73
Инварианты, %	*99.08*	*99.98*	*99.98*	*99.97*	*99.95*	*99.94*

ИНФОРМАЦИЯ ОБ АВТОРАХ

Владимир Васильевич Васильчиков
автор для корреспонденции
orcid.org/0000-0001-7882-8906. E-mail: vvv193@mail.ru
канд. техн. наук, зав. кафедрой вычислительных и программных систем.

Для цитирования: V. V. Vasilchikov, “Corrigendum to: V. V. Vasilchikov, “Parallel Algorithm for Solving the Graph Isomorphism Problem”, Modeling and analysis of information systems, vol. 27, no. 1, pp. 86–94, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-1-86-94”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 312-313, 2021.

MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

ERRATUM

Corrigendum to: Y. V. Kosolapov, "On the Detection of Exploitation of Vulnerabilities Leading to the Execution of a Malicious Code", Modeling and analysis of information systems, vol. 27, no. 2, pp. 138–151, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-2-138-151

Y. V. Kosolapov[1]

DOI: 10.18255/1818-1015-2021-3-314-316

[1]Southern Federal University, 8a Milchakova str., Rostov-on-Don 344090, Russia.

MSC2020: 68M25
Research article
Full text in Russian

Received August 25, 2021
After revision August 25, 2021
Accepted August 25, 2021

In the article by Y. V. Kosolapov "On the Detection of Exploitation of Vulnerabilities Leading to the Execution of a Malicious Code" (Modeling and analysis of information systems, vol. 27, no. 2, pp. 138–151, 2020; https://doi.org/10.18255/1818-1015-2020-2-138-151) an inaccurate description of the algorithm *CheckTrace* is committed. The correct description of the algorithm *CheckTrace* is given below. The author apologises for the inconvenience.

Keywords: system calls; library calls; software vulnerability

INFORMATION ABOUT THE AUTHORS

Yury V. Kosolapov | orcid.org/0000-0002-1491-524X. E-mail: itaim@mail.ru
correspondence author | PhD.

For citation: Y. V. Kosolapov, "Corrigendum to: Y. V. Kosolapov, "On the Detection of Exploitation of Vulnerabilities Leading to the Execution of a Malicious Code", Modeling and analysis of information systems, vol. 27, no. 2, pp. 138–151, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-2-138-151 ", *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 314-316, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

ERRATUM

Исправление к статье: Ю. В. Косолапов, «Об обнаружении эксплуатации уязвимостей, приводящей к запуску вредоносного кода», Моделирование и анализ информационных систем, Том 27, №2, с. 138–151, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-2-138-151

Ю. В. Косолапов[1]

DOI: 10.18255/1818-1015-2021-3-314-316

[1]Южный Федеральный Университет, ул. Мильчакова, д. 8а, г. Ростов-на-Дону, 344090 Россия.

УДК 517.9
Научная статья
Полный текст на русском языке

Получена 25 августа 2021 г.
После доработки 25 августа 2021 г.
Принята к публикации 25 августа 2021 г.

В статье Ю. В. Косолапова «Об обнаружении эксплуатации уязвимостей, приводящей к запуску вредоносного кода» (Моделирование и анализ информационных систем, том 27, №2, с. 138–151, 2020; https://doi.org/10.18255/1818-1015-2020-2-138-151) допущена неточность в описании алгоритма *CheckTrace*. Корректное описание алгоритма *CheckTrace* приведено ниже. Автор приносит извинения за причинённые неудобства.

Ключевые слова: системные вызовы; вызовы библиотек; уязвимости программного обеспечения

ИНФОРМАЦИЯ ОБ АВТОРАХ

Юрий Владимирович Косолапов
автор для корреспонденции
orcid.org/0000-0002-1491-524X. E-mail: itaim@mail.ru
канд. техн. наук.

Для цитирования: Y. V. Kosolapov, “Corrigendum to: Y. V. Kosolapov, “On the Detection of Exploitation of Vulnerabilities Leading to the Execution of a Malicious Code”, Modeling and analysis of information systems, vol. 27, no. 2, pp. 138–151, 2020. DOI: https://doi.org/10.18255/1818-1015-2020-2-138-151 ”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 314-316, 2021.

Корректное описание алгоритма *CheckTrace* приведено ниже.

Алгоритм 1: *CheckTrace*

Исходные параметры: 1) Последовательность $\mathrm{Path}_{t_2}(P(I))$ вида (1) длины n_I ($I \notin \boldsymbol{\mathcal{I}}(P)$),
2) профиль $\boldsymbol{\mathcal{D}}_{t_1}(P)$ вида (5) и профиль $\boldsymbol{\mathcal{C}}_{t_1}(P, l)$ вида (6),
3) порог T обнаружения нетипичного выполнения

Результат: Сообщение о нетипичной (`not typical`) или типичной (`typical`) последовательности API-вызовов

1 $result$ = `typical`, $counter$ = 0
2 **цикл** $k = 1, ..., n_I$ **выполнять**
3 **если** $k \geqslant l$ u $(n_{k-l+1}^{t_2,I}, ..., n_k^{t_2,I}) \notin \mathcal{C}_{t_1}(P, l)$ **тогда**
4 $result$ = `not typical`
5 Выйти из цикла
6 **конец условия**
7 **если** $k \leqslant n_I - 1$ **тогда**
8 $d = d_{k,k+1}^{t_2,I} - \Delta_{k+1}^{t_1,t_2,I} + \Delta_k^{t_1,t_2,I}$
9 **если** $d \in [d_{min}^{t_1}(P) : d_{max}^{t_1}(P)]$ **тогда**
10 **если** $d \notin D_{f_k^{t_2,I}, f_{k+1}^{t_2,I}}$ **тогда**
11 $counter = counter + 1$
12 **если** $counter \geqslant T + 1$ **тогда**
13 $result$ = `not typical`
14 Выйти из цикла
15 **конец условия**
16 **конец условия**
17 **конец условия**
18 **иначе**
19 $result$ = `not typical`
20 Выйти из цикла
21 **конец условия**
22 **конец условия**
23 **конец цикла**
24 **возвратить** $result$