УДК 51-77
DOI: http://dx.doi.org/10.21686/2500-3925-2019-2-57-64

Д.А. Петрусевич[1], К.Д. Шахардин[2]
[1] Российский технологический университет (МИРЭА), Москва, Россия
[2] Национальный исследовательский технологический университет «МИСиС», Москва, Россия

Аналитическая оценка результатов проверки выпускных квалификационных работ студентов средствами систем обнаружения текстовых заимствований

Цель исследования. *Цель представленной статьи – аналитическое сравнение результатов обработки выпускных квалификационных работ бакалавров и магистров кафедры Высшей математики Института кибернетики Российского технологического университета (МИРЭА) летом 2018 года с помощью двух систем обнаружения текстовых заимствований: Антиплагиат и Руконтекст. Исследование является актуальным в связи с развитием информационных технологий в образовании и возрастающей популярностью механизмов анализа текста на наличие заимствований путем автоматизированной проверки. Системы, разработанные с целью автоматизации обнаружения текстовых заимствований в различных видах работ, созданы с целью усовершенствования образовательного процесса, упрощения процедуры проверки студенческих работ преподавателями, соблюдения авторских прав, и ориентированы на развитие академической честности.*

Материалы и методы. *Математический анализ результатов был произведен на основе методов математической статистики, непосредственно в вычислительном эксперименте применён пакет статистической обработки данных языка R.*

Результаты. *В представленном исследовании был проведен педагогический эксперимент по статистическому анализу взаимосвязей характеристик выпускных квалификационных работ бакалавров и магистров кафедры Высшей математики Института кибернетики Российского технологического университета (РТУ МИРЭА) летом 2018 года: выявлены зависимости между параметрами, характеризующими конкретного студента, статистическими параметрами, описывающими его работу, и процентом оригинальности, полученным в системах проверки выпускных квалификационных работ на наличие текстовых заимствований Антиплагиат и Руконтекст. Произведено сравнение результатов, полученных при анализе выпускных квалификационных работ в разных системах. Формируются выводы о преимуществах каждой из рассматриваемых систем. При рассмотрении разницы между процентом оригинальности, полученным в системах Антиплагиат и Руконтекст, было выявлено, что с ростом длины текста работы (количества слов) растёт разница между результатами, полученными в этих системах.*

Заключение. *При поиске взаимосвязи между процентом оригинальности работы и статистическими параметрами, описывающими работу, а также доступными параметрами, характеризующими автора, оказалось, что тип зависимости для двух рассматриваемых систем совпадает, и масштаб коэффициентов одинаков. Различия наблюдаются в конкретном наборе параметров: зависимость оригинальности работы от характеристик студентов при использовании системы Руконтекст лучше описывается параметром пола, а в результатах системы Антиплагиат – уровнем образования. Это можно объяснить разным наполнением баз: в базах Антиплагиата больше студенческих работ. Также разные параметры лучше описывают зависимость процента оригинальности от длины текста: для Антиплагиата лучший результат получен при использовании количества символов, а для Руконтекст – числа слов. Эти зависимости, по-видимому, обьясняются различными техническими алгоритмами поиска заимствований в тексте. Также в исследовании рассмотрена статистическая зависимость между оригинальностью, полученной в каждой из систем.*

Ключевые слова: *Антиплагиат, Руконтекст, оригинальность текста, текстовое заимствование, цитирование*

Denis A. Petrusevich[1], Kirill D. Shakhardin[2]
[1] Russian technological university (MIREA), Moscow, Russia
[2] The National University of Science and Technology «MISIS», Moscow, Russia

Analysis of student's final qualification theses using text loans detection systems

In this paper there are results of the bachelor and master theses citing analysis. These students graduated from the Higher mathematics chair of the Russian Technological University in the summer of 2018. In this comparative analysis the dependencies of thesis loan percent on parameters of students, statistical values of their theses are explored. This research is actual because of the progress and development of new informational technologies used in the educational system. Popularity of the text loan detection systems increases. Automatic plagiarism detection systems are intended to make educational process better, the text drawing search easier, to support the copyright laws and academical honesty. The percentage is given by two main Russian plagiarism detection systems: Antiplagiat and Rucontext. Connections between thesis parameters are explored. Advantages of each text loan detection systems are described.

In this research there are the results of the pedagogical experiment aimed to analyze statistically the dependencies of the bachelor's and master's theses loan percentage which have been got from Antiplagiat and Rucontext systems on the author's parameters, statistical values describing thesis text. The comparison between statistical results of these systems have been made. The conclusions about their advantages have been presented in the paper.

In order to make the comparison methods of the mathematical statistics have been used. Numerical experiment has been provided by means of the packages of the R statistical language.

The difference between text loan percentages in the Antiplagiat and Rucontext systems has been analyzed. It has been shown that it grows when length of the text becomes larger.

The dependencies of the text loan percentage on the available parameters of the thesis author and text parameters have been presented.

The dependencies types are the same for the both systems. Scale of the coefficients in the statistical dependencies is also the same. The difference is in the very set of the parameters: the Rucontext percentage is better described statistically with the sex of the author, the Antiplagiat percentage is described with the type of the higher education (bachelor's or master's thesis). Also the dependency of the text loan percentage on the length of the thesis text differs: the Antiplagiat percentage is better described statistically with the number of words but the Rucontext percentage is described with the number of characters. It seems that these differences can be explained with different text search and analyze algorithms. The dependencies between the Rucontext percentage and the Antiplagiat text loan percentage is presented.

Keywords: *Antiplagiat, Rucontext, antiplagiarism, comparative analysis, citing, plagiarism, text drawing*

Введение

Цель представленной статьи – аналитическое сравнение результатов обработки выпускных квалификационных работ бакалавров и магистров кафедры Высшей математики Института кибернетики Российского технологического университета (МИРЭА) летом 2018 года с помощью двух систем обнаружения текстовых заимствований: Антиплагиат и Руконтекст. Выявляются статистические зависимости между параметрами, характеризующими конкретного студента, статистическими параметрами, описывающими его работу, и процентом оригинальности, полученным в системах проверки выпускных квалификационных работ на наличие текстовых заимствований Антиплагиат и Руконтекст. Произведено сравнение результатов, полученных при анализе выпускных квалификационных работ в разных системах, с помощью методов математической статистики [1–3]. Формируются выводы о преимуществах каждой из рассматриваемых систем. Математический анализ результатов был произведен с помощью пакета статистической обработки данных R.

Сегодня вопрос защиты авторских прав в системе высшего образования является достаточно актуальным. Не менее значимы проблемы формирования среди обучающихся академической этики и предупреждения некорректного использования источников информации. Решение подобных задач в современном цифровом мире практически невозможно без использования специальных технических систем, осуществляющих автоматический поиск заимствований. На рынке, предоставляющем такие услуги, в Российской Федерации и на постсоветском пространстве существует два безусловных лидера, которые признаны академическим сообществом: «Руконт» и «Антиплагиат». Первый из них позиционирует себя, в первую очередь, как межотраслевая научная библиотека, предоставляющая пользователям доступ к электронным материалам сферы образования, науки и культуры. Антиплагиат, напротив, в первую очередь предназначен именно для проверки текстов на наличие заимствований, но, кроме своей основной задачи, выполняет и ряд других функций (например, функцию электронной библиотечной системы). Оба ресурса стремительно развиваются, непрерывно модернизируют и систематически совершенствуют функциональные возможности систем. Ежегодно расширяются базы источников, разрабатываются новые технические алгоритмы поиска заимствований [4–11], ведется борьба с попытками искусственного повышения оригинальности работ, совершенствуется система поиска переводных заимствований, совершаются успешные попытки интеграции систем с LMS [12], изменяется интерфейс порталов. Вместе с тем, исследования, приведенные ниже, показывают, что обработка одного и того же текстового документа поочередно в обеих системах дает различные показатели оригинальности.

В представленном исследовании был проведен педагогический эксперимент по анализу взаимосвязей характеристик выпускных квалификационных работ бакалавров и магистров кафедры Высшей математики Института кибернетики Российского технологического университета (РТУ МИРЭА) летом 2018 года. В наборе данных присутствовала информация о выпускных работах 53 студентов, которые были проверены в системах Антиплагиат [13] и Руконт [14]. Для получения сопоставимых результатов, в обоих случаях была использована максимально доступная конфигурация модулей поиска текстовых заимствований [15], [16].

Предварительно следует дать описание анализируемого набора данных.

Описание данных, полученных при анализе выпускных квалификационных работ студентов, в системах «Антиплагиат» и «Руконтекст»

В выборке из 53 работ оказались 40 работ мужчин и 13 женщин. Все работы не имели способов технического обхода систем обнаружений заимствований.

Авторы работ характеризовались только по полу – параметр *is_male* (0 – женщина, 1 – мужчина); а также по уровню получаемого образования – параметр *is_bach* (1 – бакалавр, 0 – магистр).

Параметры выпускной работы как текста, которые учитывались в исследовании: количество символов *symbols*, слов *words*, число предложений в тексте *sent*.

Оставшиеся параметры характеризовали результат проверки в системе Антиплагиат: процент оригинальности работы a_1, процент заимствований a_{z_1} и цитирований a_{c_1}, а также число выявленных источников a_{ist}. Аналогично были измерены те же показатели с включённым модулем распознавания текста в изображениях (OCR): a_2, a_{z_2} и a_{c_2}.

Следует подчеркнуть, что в данных есть линейная зависимость (1):

$$a_i + a_{z_i} + a_{c_i} = 100,\ i = 1, 2. \quad (1)$$

Здесь i соответствует либо режиму проверки без выделения текста из изображений ($i = 1$), или с включённым модулем выделения текста из картинок ($i = 2$).

В данных, которые даёт система Руконт, присутствует три аналогичных показателя, связанных друг с другом тем же соотношением: процент оригинальности работы r, процент условно корректных заимствований (цитирований) и процент некорректных совпадений (с источниками, которых нет в списке литературы) r_{incor}, а также число выявленных источников r_{ist}. Стоит подчеркнуть, что в рамках представленного исследования было выявлено, что процент цитирований во всех проверенных работах оказался равным 0. Дело в том, что Руконт автоматически не определяет условно корректные заимствования, оставляя это на усмотрение эксперта, осуществляющего детальную работу с отчетом. Вместе с тем, на этот факт следовало бы обратить внимание сотрудникам коллектива, обслуживающего систему Руконт [17].

С учётом изложенного, так же, как и в случае системы Антиплагиат, в представленных данных есть мультиколлинеарность (2) [1–3]:

$$r + r_{incor} = 100. \quad (2)$$

Из-за этого в математических моделях, представленных ниже, будет использоваться только параметр r и не задействуется показатель r_{incor}. Из параметров, характеризующих результаты системы Антиплагиат (a_i, a_{z_i} и a_{c_i}), следует выделить только два любых параметра по причине мультиколлинеарности: результаты приведены для процента оригинальности a_i и процента некорректных заимствований a_{z_i}.

Параметры, использованные в анализе, приведены в табл. 1, 2 и 3.

Также при анализе учитывались статистические параметры текста и характеристики автора выпускной работы.

Таблица 1

Параметры, характеризующие результат анализа выпускной работы системой Антиплагиат

Переменная	Описание переменной	Формат данных
a_i	Процент оригинальности с выключенным модулем распознавания текста в изображениях (OCR)	Действительное число в отрезке [0, 100]
a_{z_1}	Процент некорректных заимствований с выключенным модулем распознавания текста в изображениях (OCR)	Действительное число в отрезке [0, 100]
a_{c_1}	Процент корректных заимствований (цитирований) с выключенным модулем распознавания текста в изображениях (OCR)	Действительное число в отрезке [0, 100]
a_2	Процент оригинальности с включенным модулем распознавания текста в изображениях (OCR)	Действительное число в отрезке [0, 100]
a_{z_2}	Процент некорректных заимствований с включенным модулем распознавания текста в изображениях (OCR)	Действительное число в отрезке [0, 100]
a_{c_2}	Процент корректных заимствований (цитирований) с включенным модулем распознавания текста в изображениях (OCR)	Действительное число в отрезке [0, 100]
a_{ist}	Количество источников, выявленных системой Антиплагиат	Целое неотрицательное число

Таблица 2

Параметры, характеризующие результат анализа выпускной работы системой Руконтекст

Переменная	Описание переменной	Формат данных
r	Процент оригинальности	Действительное число в отрезке [0, 100]
r_{incor}	Процент некорректных заимствований	Действительное число в отрезке [0, 100]
r_{ist}	Количество выявленных системой Руконтекст источников	Целое неотрицательное число

Таблица 3

Статистические параметры текста и переменные, характеризующие автора выпускной работы

Переменная	Описание переменной	Формат данных
words	Количество слов в анализируемом тексте	Целое неотрицательное число
symbols	Количество символов в тексте	Целое неотрицательное число
sent	Количество предложений в тексте	Целое неотрицательное число
is_bach	Уровень образования автора анализируемой работы	«Дамми»-переменная: 1 – бакалавр, 0 – магистр
is_male	Пол автора работы	«Дамми»-переменная: 1 – автор мужского пола, 0 – автор женского пола

Аналитическая модель оценки результатов проверки выпускных квалификационных работ на наличие текстовых заимствований.

На первом этапе исследования анализировались взаимозависимости параметров, характеризующих результаты в разных системах. С помощью линейной регрессии построена зависимость $r(a_2)$ процента оригинальности в системе Руконт от оригинальности в Антиплагиате (3) [1–3]:

$$r = 38{,}09 + 0{,}57a_2. \qquad (3)$$

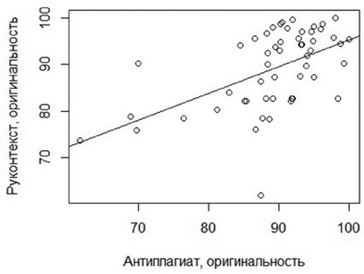

Рис. 1. Связь между оригинальностью в системах Антиплагиат и Руконт и график линейной регрессии по формуле (3)

Для большинства работ, показавших высокую оригинальность, точки группируются в правом верхнем углу диаграммы, представленной на рис. 1: по ним получены высокие проценты оригинальности как в системе Антиплагиат, так и в системе Руконт. По другим работам выявлены существенные различия в оригинальности текста. Было выявлено, что разница между показателями оригинальности в рассматриваемых системах не связана с уровнем образования и полом студента [2, 4].

В эту зависимость можно включить длину текста работы. Её адекватной характеристикой в линейной регрессии мы задали логарифм числа слов (было установлено, что при использовании других параметров, таких как число символов и предложений, модель хуже описывает данные), т.к. масштаб и единицы этого параметра сильно отличаются от других параметров в модели. В ней коэффициент при параметре, ответственном за уровень образования *is_bach*, не значим [1–3], и не отвергается гипотеза о том, что он равен нулю [1–3], поэтому для простоты мы его исключили для следующей модели и не учитывали какую-либо разницу между работами бакалавров и магистров (4):

$$r = -31{,}21 + 0{,}58a_2 + 7.58ln(words) + 4.06is\_male. \quad (4)$$

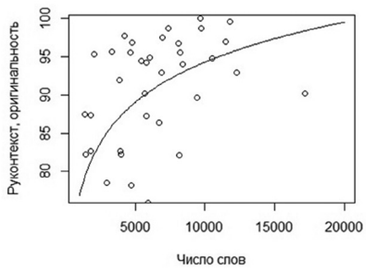

Рис. 2. График зависимости (4) процента оригинальности в системе Руконт от числа слов в проверяемом тексте для показателей is_male = 1, a_2 = 89

Все параметры в модели (4) оказались существенны [2, 3, 5, 6]. Таким образом, можно констатировать, что чем больше работа по объему, тем больше копится различий между показателями оригинальности текста, выявленными различными системами.

На следующем этапе строились зависимости между процентом оригинальности в каждой из систем и параметрами, характеризующими текст работы и студента. Базовой являлась зависимость показателя оригинальности от уровня образования и пола студента.

Для результатов в системе Руконт была получена зависимость:

$$r = 84{,}8 + 1{,}27is\_bach + 5{,}37is\_male. \qquad (5)$$

В представленной линейной регрессии (5) параметр уровня образования студента *is_bach* оказался несущественным. Гипотеза о том, что он равен 0, не отвергается [1–5].

Логично предположить, что существует связь между количеством выявленных источников при анализе работы и процентом оригинальности. В модель линейной регрессии был введён регрессор r_{ist} – число выявленных источников:

$$r = 92{,}21 - 0{,}04r_{ist} + 3{,}49is\_bach + 2{,}39is\_male. \quad (6)$$

В зависимости (6) существенную роль играет параметр r_{ist}. Коэффициент при нем мал по сравнению с коэффициентами при других параметрах, т.к. в данных r_{ist} имеет порядок десятков и сотен единиц, а остальные параметры – это «дамми»-переменные (с воз-

Таблица 4

Результаты линейной регрессии процента оригинальности в системе Руконтекст на процент оригинальности в системе Антиплагиат

	Оценка коэффициента	Стандартная ошибка	*t*-критерий Стьюдента
Свободный коэффициент	38.09	11.25	3.39
Коэффициент при параметре a_2	0.57	0.13	4.55

Таблица 5

Результаты линейной регрессии процента оригинальности в системе Руконтна характеристики автора выпускной работы

	Оценка коэффициента	Стандартная ошибка	*t*-критерий Стьюдента
Свободный коэффициент	84.80	2.03	41.77
Коэффициент при параметре *is_bach*	1.27	2.45	0.52
Коэффициент при параметре *is_male*	5.37	2.05	2.05

Таблица 6

Результаты линейной регрессии процента оригинальности в системе Руконтекст на характеристики автора и количество выявленных источников цитирования

	Оценка коэффициента	Стандартная ошибка	*t*-критерий Стьюдента
Свободный коэффициент	92.21	2.15	42.89
Коэффициент при параметре r_{ist}	-0.04	0.01	-5.31
Коэффициент при параметре *is_bach*	3.49	2.02	1.73
Коэффициент при параметре *is_male*	2.39	2.18	1.1

можными значениями 0 или 1) [1, 2, 5, 6, 7].

На последнем этапе в модель были введены параметры числа слов, предложений и символов в тексте. Было выявлено, что достаточно введения одного из этих параметров. Лучшие показатели демонстрирует модель, в которую введён параметр числа слов *words*, но не линейно, а логарифмически (в связи с тем, что природа коэффициентов, входящих в модель, существенно различается) [4–9]:

$$r = 47{,}8 - 0{,}04r_{ist} + \\ + 3{,}62is\_bach + 3{,}53is\_male + \\ + 4{,}92\ln(words). \quad (7)$$

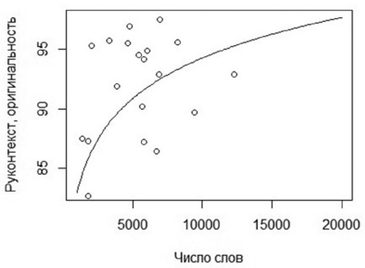

Рисунок 3. График зависимости (7) процента оригинальности в системе Руконт от числа слов в проверяемом тексте для показателей *is_male* = 1, *is_bach* = 1, r_{ist} = 150.

Существенными в этой модели являются свободный коэффициент, коэффициенты при r_{ist} и *ln(words)*. Наиболее сильная связь наблюдается между процентом оригинальности работы и параметрами числа источников, выявленных в работе, и длиной текста (количеством слов). Априори было понятно, что число выявленных источников должно снижать оригинальность текста, что мы и увидели в представленной модели [18 – 21]. При этом, интересна выявленная зависимость между оригинальностью и длиной текста: чем больше объем работы, тем, в среднем, выше оригинальность. Можно предположить, что с увеличением объема работы относительная доля заимствованного текста падает, и наоборот.

На следующем этапе анализа аналогичные зависимости были построены для процента оригинальности в системе Антиплагиат с включенным модулем распознавания текста в изображениях (OCR). Базовая модель линейной регрессии, включающая пол и уровень образования студента:

$$a_2 + 84{,}24 + 5{,}25is\_bach + \\ + 3{,}27is\_male. \quad (8)$$

В отличие от соответствующей модели для Руконта, в (8) существенным параметром, кроме свободного коэффициента, оказался коэффициент при показателе уровня образования *is_bach*. При проверке в системе Антиплагиат уровень заимствований в работах бакалавров и магистров отличался. Гипотеза о том, что коэффициент при регрессоре, описывающем пол *is_male*, равен 0, не отвергается [1 – 3].

При вводе в модель параметра количества выявленных источников a_{ist} остаётся существенным показатель уровня образования *is_bach*. Отметим, что для большинства исследуемых работ студентов число выявленных источников оказалось значительно ниже аналогичного показателя для Руконтекста r_{ist}.

$$a_2 = 89{,}3 - 0{,}09a_{ist} + \\ 4{,}76ist\_bach + 1{,}76is\_male. \quad (9)$$

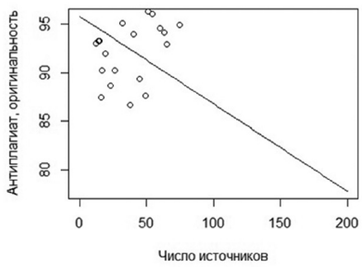

Рис. 4. График зависимости (9) процента оригинальности в системе Антиплагиат от числа источников, выявленных в проверяемом тексте, для показателей *is_male* = 1, *is_bach* = 1, a_{ist} = 41.

Так же, как и для Руконт, коэффициент при числе источников имеет малое отрицательное значение.

Судя по результатам, параметр, ответственный за пол студента, не существенен для представленной модели, поэтому для простоты мы его уберём из дальнейшего анализа. При анализе результатов, полученных в системе Антиплагиат, можно ввести параметр процента цитирований a_{z_2}, который оказывается существенным вместе со свободным коэффициентом в модели [1–3]. Не останавливаясь на этих результатах, добавим также в систему характеристику длины текста.

При подключении параметров, ответственных за длину

Таблица 7

Результаты линейной регрессии процента оригинальности в системе Антиплагиат на характеристики автора выпускной работы

	Оценка коэффициента	Стандартная ошибка	*t*-критерий Стьюдента
Свободный коэффициент	84.24	1.81	46.53
Коэффициент при параметре *is_bach*	5.25	2.19	2.4
Коэффициент при параметре *is_male*	3.27	2.33	1.4

Таблица 8

Результаты линейной регрессии процента оригинальности в системе Антиплагиат на характеристики автора, количество выявленных источников цитирования и статистические характеристики текста выпускной работы

	Оценка коэффициента	Стандартная ошибка	*t*-критерий Стьюдента
Свободный коэффициент	104.11	5.21	19.97
Коэффициент при параметре a_{ist}	0.01	0.01	0.59
Коэффициент при параметре *is_bach*	-0.26	0.55	-0.47
Коэффициент при параметре a_{z_2}	-1.01	0.04	-26.36
Коэффициент при параметре *ln(symbols)*	-0.51	0.5	-1.04

текста, наиболее целесообразным для описания результатов системы Антиплагиат оказалось ввести в модель логарифм числа символов *symbols* в тексте.

$$a_2 = 104{,}11 + 0{,}01a_{ist} - 0{,}26ist\_bach - 1{,}01a_{z_2} - \underset{0{,}51\ln(symbols)}{} \quad (10)$$

Из представленного набора регрессоров существенными являются свободный коэффициент и коэффициент при проценте корректных цитирований a_{z_2}. Для остальных коэффициентов не отвергается гипотеза о равенстве коэффициента 0 [1–5].

Заключение

Подведём итоги проведённого исследования. При рассмотрении разницы между процентом оригинальности, полученным в системах Антиплагиат и Руконт, было выявлено, что из доступных параметров наилучшим образом её описывают: длина текста (число слов) и «дамми»-переменная пол студента. Вероятно, при увеличении выборки «дамми»-переменная пол студента *is_bach* перестанет оказывать значительное воздействие на модель.

При поиске взаимосвязи между процентом оригинальности работы и статистическими параметрами, описывающими работу, а также доступными параметрами, характеризующими автора, оказалось, что тип зависимости для двух рассматриваемых систем совпадает, и масштаб коэффициентов одинаков. Различия наблюдаются в конкретном наборе параметров: зависимость в системе Руконт лучше описывается параметром пола, а в результатах системы Антиплагиат – уровнем образования. Это можно объяснить разным наполнением баз: в базах Антиплагиата больше студенческих работ. Также разные параметры лучше описывают зависимость процента оригинальности от длины текста: для Антиплагиата лучший результат получен при использовании количества символов, а для Руконт – числа слов. Эти зависимости, по-видимому, объясняются различными техническими алгоритмами поиска заимствований в тексте.

Сравнение рассматриваемых систем на предмет функциональной составляющей и визуальных характеристик, имеющих значение для конечного пользователя [15 – 17]: преподавателя, эксперта, администратора, показало разницу в интерфейсе систем, формировании истории проверок, оформлении страниц личного кабинета, отсутствия стандартизации отчетных документов.

Вместе с этим, следует отметить, что область анализа текста испытывает динамическое развитие в связи с применением методов искусственного интеллекта: в первую очередь, речь идёт об автоматическом выявлении смысла текста, измерении близости текстов, автоматическом выделении ключевых слов и фраз, применении тематического моделирования [8, 10, 18 – 21]. Эти методы преобразуют анализ текста из автоматического сравнения с учётом синонимов в понимание текста, в котором выделена основная тематика, а также в сравнительный анализ с источниками, отнесёнными к конкретной предметной и тематической области. Поэтому в ближайшее время можно ожидать улучшение качества систем анализа текста, а вместе с ними и систем, направленных на борьбу с некорректными заимствованиями.

Литература

1. Айвазян С.А. Прикладная статистика. Основы эконометрики. Том 2. М.: Юнити-Дана, 2001. 432 с.

2. Stock J.H., Watson M.W. Introduction to Econometrics. 3rd Edition. Pearson, Cloth, 2015. 840 p. ISBN 13: 9780133486872

3. Кремер Н.Ш., Путко Б.А. Эконометрика. 3-е изд., перераб. и доп. М.: Юнити-Дана, 2010. 328 с.

4. Stein R.A., Jaques P.A., Valiati J.F. An analysis of hierarchical text classification using word embeddings // Information Sciences. 2019. Vol. 471. P. 216–232.

5. Ke X., Zeng Y., Ma Q., Zhu L. Complex dynamics of text analysis // Physica A: Statistical Mechanics and its Applications. 2014. Vol. 415. P. 307–314.

6. Jones-Diette J.S., Dean R.S., Cobb M., Brennan M.L. Validation of text-mining and content analysis techniques using data collected from veterinary practice management software systems in the UK // Preventive Veterinary Medicine. 2019. Vol. 167. P. 61–67.

7. Hu N., Zhang T., Gao B., Bose I. What do hotel customers complain about? Text analysis using structural topic model // Tourism Management 2019. Vol. 72. P. 417–426.

8. Parinov S. CRIS with in-text citations as interactive entities // Procedia Computer Science. 2019. Vol. 146. P. 20–28.

9. Chen Y.-T., Chen M.C. Using chi-square statistics to measure similarities for text categorization // Expert Systems with Applications. 2011 Vol. 38(4). P. 3085–3090.

10. Петрусевич Д.А. Некоторые проблемы поиска и использования тематического моделирования при обнаружении заимствований // Сборник научных трудов Международной научно-практической конференции «Электронные системы обнаружения заимствований в оказании услуг для различных сегментов рынка». Липецк: Институт развития образования, 2016. С. 133–136.

11. Золкина А.В., Ломоносова Н.В. Опыт экспертизы выпускных квалификационных работ студентов НИТУ «МИСиС» путем обнаружения текстовых заимствований // Педагогическая информатика. 2018. № 2. С. 45–50.

12. Золкина А.В., Ломоносова Н.В. Административные особенности проверки научно-исследовательских работ в вузе на наличие текстовых заимствований // Сборник научных трудов Международной научно-практической конференции «Электронные системы обнаружения заимствований в оказании услуг для различных сегментов рынка», 27–28 октября 2016 г. Липецк: Институт развития образования, 2016. С. 87–89.

13. Чехович Ю.В., Беленькая О.С. О практике обнаружения заимствований в российских вузах // Университетская книга. 2017. №4. С. 74 – 75.

14. Воробьев М.В. Процедура выявления содержательных заимствований: противоречия гражданского права и административного права // История, теория, практика российского права. 2018. №11. С. 6 – 13.

15. Скаковская Л.Н., Медведева О.Н., Мидоренко Д.А. Использование информационных систем при оценке качества квалификационных работ // Высшее образование в России. 2015. № 5. С. 110 – 114.

16. Авдеева Н.В., Сусь И.В. Роль эксперта в оценке качества научных документов с помощью программных систем // Информационные ресурсы России. 2016. № 6 (154). С. 2–5.

17. Золкина А.В., Шахардин К.Д. Критический взгляд на использование систем автоматизированной проверки текста на заимствования // Сборник научных трудов Международной научно-практической конференции «Электронные системы обнаружения заимствований в оказании услуг для различных сегментов рынка». Липецк: Институт развития образования, 2016. С. 24–27.

18. Chatterjee A., Gupta U., Chinnakotla M.K., Srikanth R., Galley M., Argawal P. Understanding Emotions in Text Using Deep Learning and Big Data // Computers in Human Behavior. 2019. Vol. 93. P. 309–317.

19. Li X., Wang Y., Zhang A., Li C., Chi J., Ouyang J. Filtering out the noise in short text topic modeling // Information Sciences. 2018. Vol. 456. P. 83–96.

20. Chen Y., Znahg H., Liu R., Ye Z., Lin J. Experimental explorations on short text topic mining between LDA and NMF based Schemes // Knowledge-Based Systems. 2019. Vol. 163. P. 1–13.

21. Chi J., Ouyang J., Li C., Dong X., Li X., Wang X. Topic representation: Finding more representative words in topic models // Pattern Recognition Letters. 2019. Vol. 123. P. 53–60.

References

1. Ayvazyan S.A. Prikladnaya statistika. Osnovy ekonometriki. Tom 2. = Applied statistics. Basics of Econometrics. Volume 2. Moscow: Unity-Dana; 2001. 432 p. (In Russ.)

2. Stock J.H., Watson M.W. Introduction to Econometrics. 3rd Edition. Pearson, Cloth; 2015. 840 p. ISBN-13: 9780133486872

3. Kremer N.SH., Putko B.A. Ekonometrika. 3-e izd., pererab. i dop. = Econometrics. 3rd ed. Moscow: YUniti-Dana; 2010. 328 p. (In Russ.)

4. Stein R.A., Jaques P.A., Valiati J.F. An analysis of hierarchical text classification using word embeddings. Information Sciences. 2019; 471: 216–232.

5. Ke X., Zeng Y., Ma Q., Zhu L. Complex dynamics of text analysis. Physica A: Statistical Mechanics and its Applications. 2014; 415: 307–314.

6. Jones-Diette J.S., Dean R.S., Cobb M., Brennan M.L. Validation of text-mining and content analysis techniques using data collected from veterinary practice management software systems in the UK. Preventive Veterinary Medicine. 2019; 167: 61–67.

7. Hu N., Zhang T., Gao B., Bose I. What do hotel customers complain about? Text analysis using structural topic model. Tourism Management 2019; 72: 417–426.

8. Parinov S. CRIS with in-text citations as interactive entities. Procedia Computer Science. 2019; 146: 20–28.

9. Chen Y.-T., Chen M.C. Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications. 2011; 38(4): 3085–3090.

10. Petrusevich D.A. Some problems of search and use of thematic modeling in the detection of borrowing. Sbornik nauchnykh trudov Mezhdunarodnoy nauchno-prakticheskoy konferentsii «Elektronnyye sistemy obnaruzheniya

zaimstvovaniy v okazanii uslug dlya razlichnykh segmentov rynka» = Collection of scientific papers of the International Scientific and Practical Conference «Electronic systems for detection of borrowing in the provision of services for different market segments.» Lipetsk: Institute for the Development of Education; 2016: 133–136. (In Russ.)

11. Zolkina A.V., Lomonosova N.V. O The experience of examination of final qualifying works of students of NITU «MISiS» by detecting text borrowings. Pedagogicheskaya informatika = Pedagogical computer science. 2018; 2: 45–50. (In Russ.)

12. Zolkina A.V., Lomonosova N.V. Administrative features of testing scientific research at high schools for text borrowing. Sbornik nauchnykh trudov Mezhdunarodnoy nauchno-prakticheskoy konferentsii «Elektronnyye sistemy obnaruzheniya zaimstvovaniy v okazanii uslug dlya razlichnykh segmentov rynka» = Proceedings of the International Scientific and Practical Conference «Electronic systems for detecting loans in providing services for various market segments», October 27–28; 2016. Lipetsk: Institute for Educational Development; 2016: 87–89. (In Russ.)

13. CHekhovich YU.V., Belen'kaya O.S. On the practice of borrowing detection in Russian universities. Universitetskaya kniga = University Book. 2017; 4: 74–75. (In Russ.)

14. Vorob'yev M.V. The procedure for identifying meaningful borrowing: the contradictions of civil law and administrative law. Istoriya, teoriya, praktika rossiyskogo prava = History, theory, practice of Russian law. 2018; 11: 6 – 13. (In Russ.)

15. Skakovskaya L.N., Medvedeva O.N., Midorenko D.A. The use of information systems in assessing the quality of qualification works. Vyssheye obrazovaniye v Rossii = Higher education in Russia. 2015; 5: 110–114. (In Russ.)

16. Avdeyeva N.V., Sus' I.V. The role of an expert in assessing the quality of scientific documents using software systems. Informatsionnyye resursy Rossii = Information Resources of Russia. 2016; 6 (154): 2–5. (In Russ.)

17. Zolkina A.V., SHakhardin K.D. A critical look at the use of automated text verification systems for borrowing. Sbornik nauchnykh trudov Mezhdunarodnoy nauchno-prakticheskoy konferentsii «Elektronnyye sistemy obnaruzheniya zaimstvovaniy v okazanii uslug dlya razlichnykh segmentov rynka» = Collection of scientific papers of the International Scientific and Practical Conference «Electronic systems for detecting borrowing in the provision of services for various market segments.» Lipetsk: Institute for Education Development; 2016: 24–27. (In Russ.)

18. Chatterjee A., Gupta U., Chinnakotla M.K., Srikanth R., Galley M., Argawal P. Understanding Emotions in Text Using Deep Learning and Big Data. Computers in Human Behavior. 2019; 93: 309–317.

19. Li X., Wang Y., Zhang A., Li C., Chi J., Ouyang J. Filtering out the noise in short text topic modeling. Information Sciences. 2018; 456: 83–96.

20. Chen Y., Znahg H., Liu R., Ye Z., Lin J. Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems. 2019; 163: 1–13.

21. Chi J., Ouyang J., Li C., Dong X., Li X., Wang X. Topic representation: Finding more representative words in topic models. Pattern Recognition Letters. 2019; 123: 53–60.

Сведения об авторах

Денис Андреевич Петрусевич
К.ф.-м.н., доцент кафедры
Высшей математики Института кибернетики
Российский технологический университет (МИРЭА),
Москва, Россия
Эл. почта: petrdenis@mail.ru

Кирилл Денисович Шахардин
Ведущий инженер-программист отдела образовательных информационных технологий
Национальный исследовательский технологический университет «МИСиС», Москва, Россия
Эл. почта: dissertationvideo@gmail.com

Information about the authors

Denis A. Petrusevich
Cand. Sci. (Physics and Mathematics), Associate Professor, Department of Higher Mathematics of Cybernetics Institute
Russian Technological University (MIREA),
Moscow, Russia
E-mail: petrdenis@mail.ru

Kirill D. Shakhardin
Leading Software Engineer of Educational Information Technology Department
The National University of Science and Technology «MISIS», Moscow, Russia
E-mail: dissertationvideo@gmail.com