DOI: 10.15514/ISPRAS-2022-34(1)-11

О комбинированном алгоритме обнаружения заимствований в текстовых документах

[1] *К.Ф. Сафин,* ORCID: 0000-0001-9891-5513 <*kamil.safin@phystech.edu*>
[2,3] *Ю.В. Чехович,* ORCID: 0000-0002-5204-5484 <*chehovich@ap-team.ru*>
[1] *Московский физико-технический институт,*
141701, Россия, Московская область, г. Долгопрудный, Институтский переулок, д.9
[2] *Федеральный исследовательский центр "Информатика и управление" РАН*
119333, Россия, Москва, Вавилова, д.44, кор.2
[3] *Компания "Антиплагиат"*
121205, Россия, Москва, тер. Сколково инновационного центра, Большой б-р, д. 42 стр. 1

Аннотация. Поиск заимствований в текстовом документе по отношению к обширной коллекции потенциальных источников является вычислительно тяжелой задачей. При этом существуют так называемые внутренние методы поиска заимствований, которые не используют внешний корпус, а анализируют исключительно проверяемый документ. Эти методы не отличаются точностью, но обеспечивают довольно высокую производительность. В работе предложен комбинированный подход к обнаружению текстовых заимствований, основанный на использовании внутренних методов для выявления высокооригинальных документов, проверка которых по внешней коллекции не требуется. Предлагаемый алгоритм призван разгрузить систему поиска заимствований по внешней коллекции, отфильтровывая документы с высокой степенью оригинальности. В работе предлагается алгоритм поиска внутренних заимствований, описываются результаты вычислительных экспериментов.

Ключевые слова: обработка естественного языка; обнаружение заимствований; внутренние заимствования; поиск выбросов в статистике; антиплагиат

Для цитирования: Сафин К.Ф., Чехович Ю.В. О комбинированном алгоритме обнаружения заимствований в текстовых документах. Труды ИСП РАН, том 34, вып. 1, 2022 г., стр. 151-160. DOI: 10.15514/ISPRAS–2022–34(1)–11

Combined method for plagiarism detection in text documents

[1] *K.F. Safin,* ORCID: 0000-0001-9891-5513 <*kamil.safin@phystech.edu*>
[2,3] *Y.V. Chehovich,* ORCID: 0000-0002-5204-5484 <*chehovich@ap-team.ru*>[1]
[1] *Moscow Institute of Physics and Technology,*
9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia
[2] *Federal Research Center "Computer Science and Control" RAS*
44 Vavilova St., Moscow, 119333, Russia
[3] *"Antiplagiat" Company*
42 Bol'shoy Blvd., Moscow, 121205, Russia

Abstract. There are two global approaches to the problem of searching plagiarism in the text: external and intrinsic search. The first approach implies search through an external collection of documents that could have been used for text reuse. The second approach, on the contrary, does not use any external data, but analyzes the text by itself. It is proposed to combine these two approaches to speed up the search for text plagiarism. With a large flow of documents that need to be checked, the outer corpus search system processes each document and finds plagiarised blocks in each document, if there are any. However, intrinsic search could be used to determine the fact of plagiarism. Thus, it is possible to reduce the number of documents for the expensive

procedure for searching for plagiarism by the outer corpus. Moreover, in an isolated analysis of a single document, there is no need to try to find specific blocks of plagiarism, this procedure is considered as a unique indicator of the originality of the document. If the overall originality is at a low level, then this document should be sent for a more detailed and accurate check. The proposed method allows to filter texts with a high rate of originality that do not need additional verification.

Keywords: natural language processing; plagiarism detection; intrinsic plagiarism; outliers detection; antiplagiat

For citation: Safin K.F., Chehovich Y.V. Combined method for plagiarism detection in text documents. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 1, 2022, pp. 151-160 (in Russian). DOI: 10.15514/ISPRAS–2022–34(1)–11

1. Введение

1.1 Обзор литературы

Поиск заимствований в текстовых документах является сложной, но в то же время востребованной задачей, особенно в академической и студенческой средах [1, 2, 3].

Можно выделить два глобальных подхода к задаче поиска заимствований в тексте: поиск внешних заимствований (external plagiarism detection) и поиск внутренних заимствований (intrinsic plagiarism detection). Поиск внешних заимствований представляет собой поиск по внешней коллекции документов, которые могли быть использованы в качестве источника заимствования. Такой подход в том или ином виде сводится к попарному сравнению исследуемого документа с каждым документом из коллекции.

Коллекция текстовых документов, по которой происходит поиск внешних заимствований, как правило, довольно большая, а значит и поиск по ней является тяжелой вычислительной задачей. Как правило, тексты представляют в виде перекрывающихся словесных n-грамм (т.н. шинглов), которые впоследствии сравнивают с n-граммами анализируемого документа [4]. Промышленные инструменты, работающие на таком принципе сравнения документов, показывают высокую точность при поиске заимствований в текстовых документах [5]. Такой метод работает только в случае дословного заимствования фрагмента текста. Однако существуют методы обфускации (маскирования) заимствованных фрагментов, например, перефразирование или перевод текстового фрагмента из документа на другом языке. Конечно, системы поиска заимствований умеют находить и перефразирования [6] и переводные заимствования [7], однако это требует дополнительных расходов. Во-первых, требуется больше времени и вычислительных ресурсов на проверку одного документа, а во-вторых, необходимо значительно расширять текстовую коллекцию потенциальных источников.

Поиск внутренних заимствований же, наоборот, не использует внешнюю коллекцию потенциальных источников, а анализирует текст изолированно. При поиске анализируются различные стилистические, синтаксические, орфографические особенности текста.

Поиск внутренних заимствований обычно рассматривается как полноценный инструмент обнаружения текстовых заимствований. То есть, в результате работы алгоритма должны быть указаны конкретные фрагменты текста, которые были заимствованы [8]. Анализируемый текст при таком подходе, как правило, разбивается на отдельные сегменты. Например, текст делится на предложения [9], или определяется некоторая ширина шага, в соответствии с которой текст разделяется на сегменты одинаковой длины [10]. Полученные сегменты сравниваются со всем текстом и делается вывод о заимствовании для каждого сегмента. Для сравнения сегментов используются различные признаки, например, частота символьных n-грамм, из которых состоит текст [11, 12], или грамматические [13] и синтаксические признаки [14]. Иногда используются векторные представления, полученные с помощью нейронных сетей [15]. Довольно часто решается более общая задача диаризации авторов, в рамках

которой нужно определить авторство для каждого фрагмента текста [16, 17]. Методы поиска внутренних заимствований, в силу ограничения на анализ только исследуемого текста, не отличаются высокими показателями точности [18].

1.2 Актуальность работы

Сравнивая эти два подхода, можно сделать вывод, что методы поиска заимствований по внешней коллекции являются точными, но ресурсоемкими, а методы поиска внутренних заимствований – гораздо менее точными, но не сильно требовательными к ресурсам. При этом, в периоды пиковой нагрузки (например, во время сессии у студентов), система поиска по внешней коллекции может перестать справляться со входящим потоком документов для проверки, что приведет либо к сильной задержке ответа, либо к отказу от проверки. Оба случаи крайне нежелательны со стороны системы проверки. Самый простой способ ускорить работу заключается в уменьшении количества проверок (например, отказ от поиска переводных заимствований) или в сокращении коллекции потенциальных источников заимствований. И то и другое сильно скажется на качестве поиска заимствований в каждом рассматриваемом документе.

В такой ситуации кажется логичным не упрощать работу точной, но ресурсоемкой системы, а каким-то образом сократить поток входящих документов. Так как основной целью работы системы является выявление документов с высоким процентом заимствований, то было бы логично сокращать поток за счет высокооригинальных (т.е. с малой долей заимствований) документов. Для этой цели предлагается использовать подход по поиску внутренних заимствований. Как было сказано, в качестве самостоятельного инструмента, такой подход имеет очень низкое качество работы. Но его можно использовать как грубый фильтр перед более точной проверкой, который будет отсеивать документы, которым не нужна детальная экспертиза.

Причем при изолированном анализе отдельно взятого документа не нужно пытаться найти конкретные блоки заимствований, эта процедура рассматривается как своеобразный показатель оригинальности документа. В случае, если общая оригинальность на низком уровне, то этот документ стоит отправить на более детальную и точную проверку. Если же документ имеет высокую степень оригинальности, его можно пропустить и не отдавать на детальную проверку.

2. Постановка задачи

Пусть D – коллекция текстовых документов:

$$D = \{d_i\}_{i=1}^{N}.$$

Каждый документ d_i этой коллекции нужно отнести к одному из двух классов:

- класс 0 – класс высокооригинальных документов,
- класс 1 – класс документов с заимствованиями.

Под высокооригинальным документом понимается документ, содержащий малое количество заимствований из любых других текстов (или не содержащий их вовсе). Соответственно, под документом с заимствованиями понимается текст, содержащий большое число вставок из других текстов.

2.1 Критерии качества

Основная цель предлагаемого алгоритма – отфильтровывать высокооригинальные документы, не пропуская при этом документы с заимствованиями. Поэтому, при оценке качества алгоритма важны, прежде всего, следующие показатели:

- полнота класса документов с заимствованиями;

- количество отфильтрованных высокооригинальных документов.

Под полнотой класса (recall) понимается следующая величина:

$$Recall = \frac{TP}{TP + FN},$$

где TP, FN – true-positive и false-negative объекты. То есть это тексты, которые верно отнесены алгоритмом к нужному классу и, наоборот, тексты, которые неверно отнесены алгоритмом к противоположному классу, соответственно. Таким образом, полнота класса – это доля верно найденных алгоритмом документов из всех документов этого класса.

Второй показатель (количество высокооригинальных документов) важен, так как можно привести пример крайнего случая, когда все документы относятся к классу документов с заимствованиями. Тогда полнота класса документов с заимствованиями будет равна 1, но ни одного документа отфильтровано не будет, и нагрузка на систему поиска внешних заимствований не уменьшится.

Количество отфильтрованных документов можно рассматривать как полноту класса высокооригинальных документов. То есть, нам важна полнота обоих классов, однако полнота класса документов с заимствованиями важнее. Можно ввести вспомогательную метрику, которую будем использовать при настройке гиперпараметров алгоритма:

$$Recall_\beta = \beta \cdot Recall_1 + Recall_0, \quad (1)$$

где $Recall_1$ и $Recall_0$ – полнота класса документов с заимствованиями и высокооригинальных документов соответственно. Для того, чтобы полнота класса 1 была в приоритете, весовой коэффициент берется β больше единицы

3. Описание алгоритма

Общую логику работы предлагаемого алгоритма можно представить в виде псевдокода:

Алгоритм определения факта заимствования в тексте

```
Require: text
     statsList ← [ ]
     text ← preprocess(text)
     segmentsList ← getSegments(text)
     for segment in segmentsList do
         segmentVector = vectorize(segment)
         stat ← calcStat(vector)
         statsList.append(stat)
     end for
     outliersCount ← 0
     for stat in statsList do
         if stat > statT hreshold then
             outliersCount+ = 1
         end if
     end for
     if outliersCount > outliersThreshold then
         return 'text is not original'
     else
         return 'text is original'
     end if
```

Алгоритм состоит из следующих основных этапов:

- предобработка текста;
- сегментация текста;
- векторизация сегментов;
- расчет статистик для сегментов;
- обнаружение выбросов в ряде статистик.

3.1 Предобработка текста

Сначала текст проходит процедуру предобработки. Используются стандартные техники обработки естественного языка: удаление редких символов, удаление стоп-слов (слов, не несущих смысловую нагрузку), приведение слов к начальной форме. Конкретные техники предобработки и их параметры подбираются при настройке алгоритма на конкретном корпусе документов.

3.2 Сегментация текста

Под сегментацией текста понимается процедура разбиения текста d на сегменты s_j:

$$d = \bigcup_{j=1}^{m} s_j.$$

При этом сегменты могут быть пересекающимися и, наоборот, иметь нулевое пересечение. Для каждого сегмента затем будет рассчитываться некоторая статистика, поэтому сегментирование должно удовлетворять некоторым условиям.

Разбиение на сегменты должно быть достаточно мелким, чтобы можно было детектировать выброс в ряде статистик, и значение статистики на некотором сегменте сильно отличалось от остальных значений. С другой стороны, размер отдельного сегмента должен быть достаточно велик, чтобы можно было посчитать адекватную статистику.

Самые популярные стратегии сегментации, которые применимы в данном случае:

- разбиение по параграфам;
- разбиение окном с фиксированным шагом.

В процессе настройки гиперпараметров алгоритма, выбирается стратегия сегментации, а также ее аргументы (ширина окна и размер шага).

3.3 Векторизация сегментов

Каждый сегмент текста подвергается процедуре векторизации. Для построения векторного представления сегмента используются частоты словесных и символьных *n*-грамм.

Под символьной или словесной *n*-граммой понимается последовательность из n символов или слов в тексте. Каждой *n*-грамме ω в тексте d ставится в соответствие число:

$$freq_\omega = \frac{cnt(\omega)}{\sum_{\omega'} cnt(\omega')} \log\left(\frac{m}{seq(m)}\right), \quad (2)$$

где $cnt(\omega)$ – число вхождений ω в текст d, m – число сегментов в тексте, $seq(m)$ – число сегментов, содержащих ω.

Вектор сегмента формируется из рассчитанных величин (2) для всех уникальных n-грамм в тексте. Если *n*-грамма есть в тексте, но отсутствует в сегменте, то значение (2) равно 0.

Тип рассматриваемых *n*-грамм (символьные или словесные) выбирается при настройке гиперпараметров алгоритма на конкретном корпусе документов.

3.4 Подсчет статистик и нахождение аномалий

Для рассматриваемого текста строится ряд статистик путем подсчета некоторой статистики для каждого сегмента текста. В качестве статистики выбрано расстояние от вектора сегмента s_j до усредненного вектора всех сегментов $\overline{s}$:

$$s = \frac{1}{m}\sum_{j=1}^{m} s_j.$$

Тип расстояния между векторами также выбирается при настройке алгоритма. Выбор происходит из следующих вариантов:

- евклидово расстояние;
- косинусное расстояние;
- нормированное евклидово расстояние.

Полученный ряд статистик сглаживается скользящим средним фиксированной ширины.

В полученном ряде статистик выполняется поиск выбросов. Под выбросом подразумевается значение статистики, которое превышает некоторый заданный порог (который подбирается при настройке гиперпараметров). По количеству выбросов в тексте принимается решение об оригинальности документа.

4. Вычислительный эксперимент

Для проверки качества предложенного алгоритма было проведено два вычислительных эксперимента. Первый эксперимент был проведен на корпусе англоязычных документов, подготовленных в рамках конкурса по обнаружению текстовых заимствований PAN-2020. Для проведения второго эксперимента был использован корпус русскоязычных текстов Paraplag [19], специально составленный для проверки алгоритмов поиска текстовых заимствований.

В рамках каждого эксперимента производилась настройка гиперпараметров описанного алгоритма. Для настройки, данные разбивались на обучающую и тестовую выборки, в размерах 70% и 30% от всего корпуса соответственно. Настройка гиперпараметров производилась с помощью кроссвалидации на трех разбиениях обучающей выборки. В качестве целевой метрики использовалась предложенная метрика (1).

4.1 Описание данных

В первой части вычислительного эксперимента используется корпус текстов, подготовленных и размеченных в рамках конкурса PAN-2020 [20]. Корпус содержит документы на английском языке. Каждый документ может содержать от 0 до 10 вставок текста другого авторства.

Корпус состоит из двух частей. Первая часть представляет из себя узкоспециализированный набор документов – все документы в ней посвящены теме технологий. Вторая часть корпуса является набором текстов различной тематики (путешествия, философия, экономика, история и т.д.). Это сделано для того, чтобы была возможность протестировать предлагаемые алгоритмы на устойчивость к смене тематики при работе с документами. Количество документов с заимствованиями примерно равно количеству высокооригинальных документов.

Для второй части эксперимента был использован русскоязычный корпус тексов, содержащий документы с заимствованиями Paraplag [19]. Корпус представляет из себя набор текстов (эссе), в которые авторы намеренно добавляли заимствования из других документов. В качестве высокооригинальных документов представлены источники этих заимствований

(статьи из энциклопедий). Доля документов с заимствованиями относительно всего корпуса около 15%.

4.2 Результаты

Для оценки итогового качества полученного алгоритма, была рассчитана полнота каждого из целевых классов на тестовой выборке. Результаты экспериментов приведены в табл. 1.

Табл..1. Результаты эксперимента
Table 1. Experiment results

Название корпуса	Язык	Доля класса 1 в корпусе	Полнота класса 0	Полнота класса 1
PAN-2020	английский	50%	10%	94%
Paraplag	русский	15%	32%	97%

Видно, что на обоих корпусах полнота класса 1 близка к 100%. Это значит, что малая часть документов, которым необходима детальная проверка с помощью системы поиска внешних заимствований, будет неправомерно отфильтрована. При этом, часть высокооригинальных документов будет правильно отсеяна, что снизит нагрузку на систему.

Примеры работы алгоритма на конкретных текстах приведены на рис. 1 и рис. 2. Синими линиями представлены значения рядов статистик для каждого текста, красной горизонтальной – верхний порог статистики, выше которого она считается выбросом.

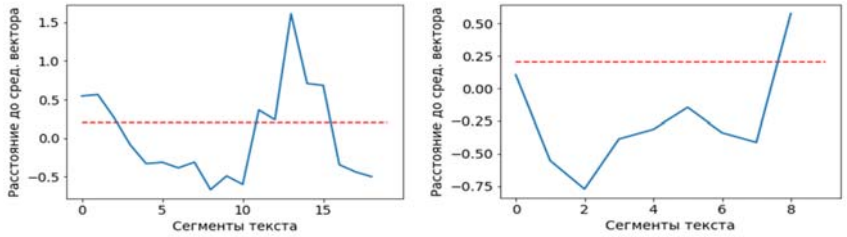

Рис. 1. Пример работы на тексте с заимствованиями (слева) и на высокооригинальном тексте (справа). Английский корпус документов
Fig. 1. Algorithm results on plagiarised text (left) and on original one (right). English corpus

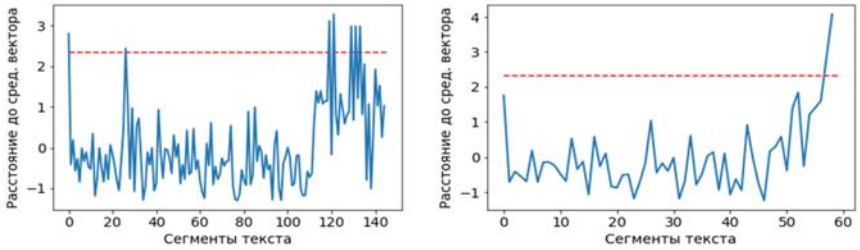

Рис. 2. Пример работы на тексте с заимствованиями (слева) и на высокооригинальном тексте (справа). Русский корпус документов
Fig. 2. Algorithm results on plagiarised text (left) and on original one (right). Russian corpus

5. Заключение

В статье предложен алгоритм по обнаружению факта заимствований в тексте. При этом алгоритм анализирует текст изолированно, не используя внешнюю коллекцию возможных источников заимствований. Для установления факта заимствования, текст сегментируется, и для каждого сегмента рассчитывается статистика, основанная на частотах распределения n-грамм. В полученном ряде статистик происходит поиск выбросов, и по их количеству делается вывод о степени оригинальности текста.

Предложенный алгоритм был настроен и протестирован на корпусах английских и русских текстов. В обоих случаях алгоритм корректно отбирает высокооригинальные документы, оставляя при этом документы с заимствованиями для дальнейшей проверки. Таким образом, алгоритм удовлетворяет выдвинутым к нему требованиям по качеству и может быть использован в качестве первичного отбора документов при использовании высоконагруженной системы поиска заимствований по внешнему корпусу.

Как было сказано, основная цель предложенного алгоритма заключается в фильтрации высокооригинальных документов, для которых не требуется детальная проверка. Конечно, в идеальном случае все поступающие на проверку документы должны проходить полноценную проверку. Однако реальность такова, что при высокой нагрузке, система поиска внешних заимствований может сильно задерживать ответ или пропускать документы. В такой постановке кажется логичным пожертвовать малым (частью документов, с низкой долей заимствований), чтобы сохранить работоспособность системы.

Предложенный алгоритм как раз и осуществляет такую логику. При сравнительно небольшом количестве пропущенных документов с заимствованиями (около 3% для русского корпуса) удалось сократить поток документов для обработки почти на треть.

Список литературы / References

[1] Никитов А.В., Орчаков О.А., Чехович Ю.В. Плагиат в работах студентов и аспирантов: проблема и методы противодействия. Университетское управление: практика и анализ, no. 5, 2012 г., стр. 61-68 / Nikitov A.V., Orchakov O.A., Chehovich Yu.V. Plagiarism in works of undergraduate and graduate students: problem and methods of counteraction. University Management: Practice and Analysis, no. 5, 2012, pp. 61-68 (in Russian).

[2] Stein B., Koppel M., Stamatatos E. Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07. SIGIR Forum, vol. 41, no. 2, 2007, pp. 68–71.

[3] Chekhovich Y.V., Khazov A.V. Analysis of duplicated publications in Russian journals. Journal of Informetrics, vol.16, issue 1, 2022, article no. 101246.

[4] Зеленков И.В., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов. Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007), 2007 г., стр. 166-174 / Zelenkov I.V., Segalovich I.V. Comparative analysis of methods for determining fuzzy duplicates for Web documents. In Proc. of the 9th All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» (RCDL'2007), 2007, pp. 166-174 (in Russian).

[5] Журавлев Ю.И., Рудаков К.В. и др. Система распознавания интеллектуальных заимствований «Антиплагиат». Математические методы распознавания образов, том 12, no. 1, 2005 г., стр. 329-332 / Zhuravlev Yu.I., Rudakov K.V. et al. The system of recognition of intellectual borrowings «Anti-plagiarism». Mathematical methods of pattern recognition, vol. 12, no. 1, 2005, pp. 329-332 (in Russian).

[6] Socher R., Huang E.H.-C. et al. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In Proc. of the 24th International Conference on Neural Information Processing Systems, 2011, pp. 801-809.

[7] Кузнецова Р.В., Бахтеев О.Ю., Чехович Ю.В. Методы обнаружения переводных заимствований в больших текстовых коллекциях. Информатика и её применения, том 15, no. 1, 2021 г., стр. 30–41 / Kuznetsova R.V., Bakhteev O.Yu., Chekhovich Yu.V. Methods of cross-lingual text reuse detection in large textual collections. Informatics and Applications, vol. 15, no. 1, 2021, pp. 30-41 (in Russian).

[8] Meier zu Eissen S., Stein B. Intrinsic Plagiarism Detection. Lecture Notes in Computer Science, vol. 3936, 2006, pp. 565-569.

[9] Zechner M., Muhr M. et al. External and Intrinsic Plagiarism Detection Using Vector Space Models. In Proc. of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, CEUR Workshop Proceedings, vol. 502, 2009, pp. 47-55.
[10] Oberreuter G., L'Huillier G. et al. Outlier-Based Approaches for Intrinsic and External Plagiarism Detection. Lecture Notes in Computer Science, vol. 6882, 2011, pp. 11-20.
[11] Stamatatos E. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In Proc. of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, CEUR Workshop Proceedings, vol. 502, 2009, pp. 38–46.
[12] Bensalem I., Rosso P., Chikhi S. Intrinsic Plagiarism Detection using Ngram Classes. In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1459-1464.
[13] Tschuggnall M., Specht G. Countering plagiarism by exposing irregularities in authors grammars. In Proc. of the European Intelligence and Security Informatics Conference, 2013, pp. 15-22.
[14] Романов А.С., Мещеряков Р.В., Резанова З.И. Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции. Доклады Томского государственного университета систем управления и радиоэлектроники, no. 2(32), 2014 г., стр. 264-269 / Romanov A.S., Mescheryakov R.V., Rezanova Z.I. Plagiarism detection and text homogeneity checking technique based on one-class support machine and fast correlation-based filter. Proceedings of TUSUR University, no. 2(32), 2014, pp. 264-269.
[15] Safin K., Kuznetsova R. Style Breach Detection with Neural Sentence Embeddings. In Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 1866, 2017, 7 p.
[16] Kuznetsov M., Motrenko A., Kuznetsova R., Strijov V Methods for intrinsic plagiarism detection and author diarization. In Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, CEUR Workshop Proceedings, vol. 1609, 2016, 8 p.
[17] Gillam L., Vartapetiance A. Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification. In Working Notes for CLEF 2012 Conference, CEUR Workshop Proceedings, vol. 1178, 2012, 12 p.
[18] Potthast M., Eiselt A. et al. Overview of the 3rd International Competition on Plagiarism Detection. Working Notes for CLEF 2011 Conference, CEUR Workshop Proceedings, vol. 1171, 2011, 10 p.
[19] Sochenkov I.V., Zubarev D.V., Smirnov I.V. The ParaPlag:: Russian dataset for paraphrased plagiarism detection. In Proc. of the International Conference "Dialogue 2017", 2017, 13 p.
[20] Zangerle E., Mayerl, M. et al. PAN20 Authorship Analysis: Style Change Detection. Available at: https://doi.org/10.5281/zenodo.3660984.

Информация об авторах / Information about authors

Камиль Фанисович САФИН – аспирант. Сфера научных интересов: прикладной анализ данных, обработка естественного языка, поиск текстовых заимствований, методы оптимизации.

Kamil Fanisovich SAFIN – PhD student. Research interests: applied data analysis, natural language processing, text plagiarism detection, optimization methods.

Юрий Викторович ЧЕХОВИЧ – к.ф.-м.н., заведующий отделом Федерального исследовательского центра "Информатика и управление" РАН, исполнительный директор компании "Антиплагиат". Сфера научных интересов: прикладные задачи анализа данных, алгоритмы и системы обнаружения заимствований в текстовых документах, моделирование транспортных систем.

Yury Victorovich CHEHOVICH – Head of Department, Federal Research Center for Informatics and Control, Russian Academy of Sciences, CEO at Antiplagiat Company. Research interests: applied problems of data analysis, algorithms and systems for text plagiarism detection, transport systems modeling.