На правах рукописи
УДК 004.021

Сафин Камиль Фанисович

Комбинированные методы выявления заимствований в текстовых документах

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2022

Работа прошла аппробацию на кафедре Интеллектуальных систем Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

Научный руководитель: **Чехович Юрий Викторович**
кандидат физико-математических наук, Зав. отделом Вычислительного центра, Федеральный исследовательский центр «Информатика и управление» Российской академии наук.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В.П. Иванникова Российской академии наук (ИСП РАН)

Защита состоится 23 сентября 2022 г. в 15 часов 00 минут на заседании диссертационного совета ФРКТ.05.13.17.005, созданного на базе федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» (МФТИ, Физтех)

по адресу: 141701, Московская область, г. Долгопрудный, Институтский переулок, д.9.

С диссертацией можно ознакомиться в библиотеке МФТИ, Физтех и на сайте организации https://mipt.ru.

Автореферат разослан «__» _____ 2022 г..

Ученый секретарь
диссертационного совета Сахно Сергей Владимирович

Общая характеристика работы

Актуальность темы. Поиск заимствований в текстовых документах является сложной, но в то же время востребованной задачей, особенно в академической и студенческой средах [1—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 постоянно расширять текстовую коллекцию потенциальных источников.

Поиск внутренних заимствований же, наоборот, не использует внешнюю коллекцию потенциальных источников, а анализирует текст изолированно [8]. 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 [9]. Анализируемый текст при таком подходе, как правило, разбивается на отдельные сегменты. Например, текст делится на предложения [10], или определяется некоторая ширина шага, в соответствии с которой текст разделяется на сегменты одинаковой длины [11]. 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 [12; 13], или грамматические [14] и синтаксические признаки [15]. Иногда используются векторные представления, полученные с помощью нейронных сетей [A1]. 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].

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 случая 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 выгодно 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.

Целью данной работы является разработка методов обнаружения некорректных текстовых заимствований без использования внешней коллекции потенциальных источников заимствований, а также реализация программного комплекса на основе предложенных методов. Задачей данного программного комплекса является повышение эффективности промышленной системы обнаружения текстовых заимствований за счет выбора набора методов, которыми будет осуществляться проверка. Выбор происходит таким образом, что для части документов выбираются методы с низкими требованиями к вычислительным ресурсам, а для части документов, требующих детальной проверки — методы с высокой вычислительной сложностью.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Исследовать существующие методы поиска текстовых заимствований без использования потенциальной коллекции источников.
2. Предложить метод обнаружения некорректных заимствований, использующий только информацию об исследуемом тексте, и оценить работоспособность такого метода
3. На основе предложенного алгоритма разработать способ фильтрации документов для последующего использования различных наборов методов при проверке на заимствования.
4. Протестировать и оценить качество алгоритма на реальных данных.

Научная новизна данной работы заключается в разработке набора алгоритмов по обнаружению некорректных текстовых заимствований. Предложен способ обнаружения границ смены авторского стиля письма, основанный на анализе частот употребления словесных и символьных n-грамм. На основе данного способа предложен метод фильтрации высокооригинальных текстов, которые не нуждаются в детальной проверке через систему поиска внешних заимствований.

Практическая значимость данной работы заключается в том, что предлагаемые методы предназначены для предварительного анализа документов на предмет заимствований. Документы, которые по результатам этой проверки имеют очень мало потенциальных некорректно использованных фрагментов, могут быть исключены из очереди на проверку по полноценной системе поиска заимстований, что частично снизит нагрузку на эту систему. Предложенные методы не требуют больших вычислительных мощностей, что позволяет использовать их для экономии машинного времени и ресурсов в периоды высокой нагрузки на систему поиска заимствований. Также важно упомянуть, что предлагаемые методы предназначены в том числе для работы на русском языке. Это важно ввиду того, что основные методы, предлагаемые в научном сообществе, изначально предназначены для английского языка и не адаптированы для русского.

Методология и методы исследования. Для достижения заявленных целей, используется метод, основанный на анализе частот употребления слов и символьных n-грамм [19]. Используется адаптация метода векторизации с помощью статистик tf-idf [20] применительно к задаче векторизации текстовых сегментов.

Основные положения, выносимые на защиту:

1. Предложен способ векторизации фрагментов текста, основанный на частотах встречаемости символьных и словесных n-грамм в анализируемом тексте и в каждом фрагменте по отдельности.
2. Разработан способ обнаружения заимствованных фрагментов текста, основанный на сегментировании анализируемого текста и ана-

лизе ряда статистик, построенных для каждого из полученных сегментов, на предмет наличия выбросов.

3. Разработан метод обнаружения и фильтрации высокооригинальных текстовых документов без внешней коллекции потенциальных источников и с использованием малых вычислительных мощностей.
4. Обоснована работоспособность предложенного алгоритма путем реализации и тестирования на подготовленных данных. Экспериментально показано, что предложенный алгоритм может отфильтровывать до 30% высокооригинальных документов, не сильно проигрывая в качестве полноценной проверке.

Апробация работы. Основные результаты работы докладывались и обсуждались на следующих научных конференциях:

1. «Определение заимствований в тексте без указания источника», Всероссийская конференция «59-ая научная конференция МФТИ с международным участием», 2016.
2. «Style Breach Detection with Neural Sentence Embeddings», Международная конференция «Conference and Labs of the Evaluation Forum», 2017
3. «Detecting a Change of Style using Text Statistics», Международная конференция «Conference and Labs of the Evaluation Forum», 2018
4. «CrossLang: The System of Cross-lingual Plagiarism Detection», Международная конференция «Workshop on Truth Discovery and Fact Checking: Theory and Practice at conference on Knowledge Discovery and Data mining», 2019
5. «CrossLang: The System of Cross-lingual Plagiarism Detection», Международная конференция «Workshop on Deep Learning for Education at conference on Knowledge Discovery and Data mining», 2019
6. «Определение факта заимствования в текстовых документах без указания источника», Всероссийская конференция «Математические методы распознавания образов (ММРО)», 2021.

Личный вклад. Все приведенные результаты, получены диссертантом лично при научном руководстве к.ф.-м.н. Ю. В. Чеховича.

Публикации. Основные результаты по теме диссертации изложены в 5 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК, 4 — в периодических научных журналах, индексируемых Web of Science и Scopus.

1. *К. Ф. Сафин.* Определение заимствований в тексте без указания источника / К. Ф. Сафин, М. П. Кузнецов, М. В. Кузнецова // Информ. и её примен. 2017. т. 11, № 3
2. *Safin, K.* Style Breach Detection with Neural Sentence Embeddings / K. Safin, R. Kuznetsova // Working Notes of CLEF 2017 - Conference

and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. Vol. 1866 / ed. by L. Cappellato [et al.]. CEUR-WS.org, 2017. (CEUR Workshop Proceedings)
3. *Safin, K.* Detecting a Change of Style using Text Statistics: Notebook for PAN at CLEF 2018 / K. Safin, A. Ogaltsov // Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. Vol. 2125 / ed. by L. Cappellato [et al.]. CEUR-WS.org, 2018. (CEUR Workshop Proceedings)
4. Near-duplicate handwritten document detection without text recognition / O. Bakhteev [et al.] // Computational Linguistics and Intellectual Technologies. 2021
5. *К. Ф. Сафин.* О комбинированном алгоритме обнаружения заимствований в текстовых документах / К. Ф. Сафин, Ю. В. Чехович // Труды Института системного программирования РАН. 2022. т. 34, № 1. с. 151—160

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы.

В **главе 1** приводится обзор литературы, посвященной проблематике поставленной задачи. Описываются и систематизируются методы, применявшиеся для решения задачи поиска текстовых заимствований. Также приводятся различные вариации поставленной задачи и их неформальные постановки. Среди родственных задач поиска внутренних заимствований можно выделить основные:

- Кластеризация по авторству: имея текстовый документ, необходимо выделить в нем сегменты и сгруппировать эти сегменты согласно авторству.
- Обнаружение факта мультиавторства: нужно сделать вывод, является ли исследуемый текст оригинальной работой одного автора или же нескольких.
- Нахождение нарушений стиля: в анализируемом тексте необходимо найти позиции, на которых происходит изменение авторского стиля.
- Определение числа авторов: для исследуемого текста необходимо установить, скольким авторам принадлежит исследуемый текст.
- Проверка авторства: имея два текста (или их фрагменты), нужно установить, принадлежат ли они одному автору или разным.

Также в главе приводится категоризация текстовых признаков, используемых в области анализа текстов на естественном языке.

В **главе 2** описывается общий предлагаемый подход по поиску некорректных текстовых заимствований без использования внешних источников.

Предлагаемый метод состоит из последовательных шагов:

1. сегментирование текста,
2. построение векторных представлений полученных сегментов,
3. расчет статистик для каждого сегмента,
4. поиск выбросов в полученном ряде статистик.

Сегментирование. В подзадаче сегментирования необходимо разбить текст на составляющие его сегменты:

$$d = \bigcup_{i=1}^{|d|} s_i.$$

При этом сегменты должны удовлетворять следующим критериям:

- С одной стороны, сегменты должны быть достаточно малы, чтобы отдельно взятый сегмент содержал в себе либо текст оригинального автора, либо некорректно заимствованный текст.
- С другой стороны, сегменты должны быть достаточно велики, чтобы можно было собрать статистически достоверные признаки данного сегмента.

Самым подходящим подходом в таком случае является разбиение текста по предложениям. Однако к недостаткам такого метода можно отнести слишком короткие предложения, которые практически не содержат информацию об авторском стиле. В таком случае предлагается применять разбиение текста на сегменты одинаковой длины с фиксированным шагом или объединять несколько предложений в один сегмент. Выбор конкретного метода сегментации зависит от решаемой задачи и выборки документов.

Векторизация. Для векторизации сегментов текста предлагается адаптировать распространенный метод построения векторных представлений текстов с использованием tf-idf статистик. Данный подход векторизует каждый отдельно взятый текст d_i с учетом корпуса документов D, в рамках которого этот текст рассматривается.

Статистика tf-idf для слова в документе представлена в виде произведения двух независимых друг от друга статистик.

Первая статистика, tf (term frequency — частота слова), отражает то, насколько часто данное слово w_j употребляется в данном тексте d_i. Самым распространенным является способ подсчет числа вхождений слова w_j в d_i, с нормировкой на общее число слов в тексте. В таком случае, формула

для расчета статистики слова w_j в тексте d_i выглядит следующим образом:

$$\mathrm{tf}(w_j, d_i) = \frac{n_{w_j}^{d_i}}{\sum\limits_j n_{w_j}^{d_i}}, \tag{1}$$

где за $n_{w_j}^{d_i}$ обозначено число вхождений слова w_j в текст d_i

Вторая статистика, idf (inverse document frequency — обратная частота документа), является инверсией частоты, с которой рассматриваемое слово встречается в документах коллекции D:

$$\mathrm{idf}(w_j, D) = 1 + \log \frac{|D|}{|\{d_i \in D | w_j \in d_i\}|}, \tag{2}$$

где под $|D|$ понимается число документов в рассматриваемой коллекции, а под $|\{d_i \in D | w_j \in d_i\}|$ — число документов, содержащих рассматриваемое слово w_j. Добавление единицы к итоговой статистике необходимо для того, чтобы слова, которые встречаются абсолютно во всех документах, не получили нулевую величину. Логарифм берется для того, чтобы статистика не принимала очень большие значения для редких слов (для них число в знаменателе будет близко к нулю).

Итоговая величина tf-idf для слова w_j в документе d_i в рамках корпуса документов D рассчитывается как произведение описанных множителей:

$$\begin{aligned} \text{tf-idf}(w_j, d_i, D) &= \mathrm{tf}(w_j, d_i) \cdot \mathrm{idf}(w_j, D) = \\ &= \frac{n_{w_j}^{d_i}}{\sum\limits_j n_{w_j}^{d_i}} \cdot \left(1 + \log \frac{|D|}{|\{d_i \in D | w_j \in d_i\}|}\right). \end{aligned} \tag{3}$$

Векторное представление текста с учетом tf-idf статистик для слов строится следующим образом:

$$\mathbf{x}_{d_i}^{tf-idf} = \begin{bmatrix} \text{tf-idf}(w_1, d_i, D) \\ \vdots \\ \text{tf-idf}(w_{V_d}, d_i, D) \end{bmatrix}. \tag{4}$$

Предлагается адаптировать данный подход для подзадаче векторизации сегментов. Изменение будет заключаться в том, что в роли корпуса документов теперь будет выступать набор сегментов текста, а в роли векторизуемого текста – рассматриваемый сегмент. Тогда формула для расчета tf статистики (1) для слова w_j в сегменте s_i будет выглядеть следующим образом:

$$\mathrm{tf}_{seg}(w_j, s_i) = \frac{n_{w_j}^{s_i}}{\sum\limits_j n_{w_j}^{s_i}}, \tag{5}$$

где за $n_{w_j}^{s_i}$ обозначено число вхождений слова w_j в сегмент s_i. А статистика idf для слова w_j в документе d (2) видоизменится следующим образом:

$$\text{idf}_{seg}(w_j, d) = 1 + \log \frac{|d|}{|\{s_i \in d | w_j \in s_i\}|}, \tag{6}$$

где через $|d|$ обозначено количество сегментов текста. Аналогично, итоговая статистика tf-idf для слова w_j в документе d получается путем перемножения двух величин (3):

$$\begin{aligned} \text{tf-idf}_{seg}(w_j, s_i, d) &= \text{tf}_{seg}(w_j, s_i) \cdot \text{idf}_{seg}(w_j, d) = \\ &= \frac{n_{w_j}^{s_i}}{\sum\limits_j n_{w_j}^{s_i}} \cdot \left(1 + \log \frac{|d|}{|\{s_i \in d | w_j \in s_i\}|}\right). \end{aligned} \tag{7}$$

Так же, как и в случае векторизации текстов, построенная статистика может использоваться для векторизации сегментов текста:

$$\mathbf{x}_{s_i}^{tf-idf} = \begin{bmatrix} \text{tf-idf}_{seg}(w_1, s_i, d) \\ \vdots \\ \text{tf-idf}_{seg}(w_{S_d}, s_i, d) \end{bmatrix}. \tag{8}$$

При этом также слова будут иметь веса соответственно их важности. Распространенные слова, которые встречаются во всем тексте, получат меньший коэффициент, а более редкие слова, которые встречаются в малом количестве сегментов, получат больший коэффициент.

Расчет статистик. Для выявления некорректно заимствованных сегментов текста, необходимо построить для каждого из них некоторую статистику, которая будет отображать степень принадлежности сегмента стилю письма основного автора текста.

Так как векторы сегментов находятся в пространстве единой размерности, то надо оценить, насколько вектора удалены друг от друга. Предлагается рассчитывать попарные расстояния между векторами в качестве такой оценки. Чаще всего для этих целей используют косинусную меру близости, так как в высокоразмерных пространствах удобнее использовать ограниченную функцию, однако также используются и другие метрики, например, L1.

Поиск выбросов. Как уже было сказано, сегменты, отличающиеся от остального текста, потенциально могут быть некорректно заимствованными. С учетом описанных ранее процедур векторизации и расчета статистик для каждого сегмента, это значит, что статистика рассматриваемого сегмента должна сильно отличаться от остальных. Для обнаружения выбросов используется подход фильтрации по некоторому пороговому значению.

Статистики, превышающие некоторое пороговое значение считаются выбросами, а соответствующие сегменты — некорректно заимствованными.

Также в главе приводится базовый эксперимент, демонстрирующий работоспособность предложенного метода. Эксперимент был проведен на подвыборке корпуса текстов, подготовленных в рамках конкурса по поиску текстовых заимствований PAN CLEF.

В **главе 3** описывается система поиска некорректных текстовых заимстований с явным указанием подозрительных участков текста. Используется метод, описанный в главе 2, адаптированный под данную постановку задачи.

Стратегия сегментирования заключается в следующем. Текст разбивается на предложения. Затем формируется разбиение текста на сегменты s_i: если длина очередного предложения меньше минимальной длины сегмента l_{segm}, к этому предложению добавляется следующее за ним — процесс повторяется, пока длина сегмента s_i не превысит заданную минимальную длину. Минимальная длина сегмента l_{segm} является настраиваемым параметром алгоритма.

Метод векторизации сегментов используется в виде, описанном в главе 2. Ряд статистик, полученный путем расчета отклонений векторов сегментов $\mathbf{t}_i$ от среднего вектора, сглаживается скользящим средним: новые значения статистики $stat'(\mathbf{t}_i)$ вычисляются по формуле

$$stat'(\mathbf{s}_i) = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} stat(\mathbf{s}_i), \tag{9}$$

где n — ширина сглаживания, которая также является настраиваемым параметром. Значения в крайних точках вычисляются по формуле (N — число сегментов)

$$\begin{aligned} stat'(\mathbf{t}_i) &= \frac{1}{i+n+1} \sum_{k=0}^{i+n} stat(\mathbf{t}_i), \\ stat'(\mathbf{t}_i) &= \frac{1}{i+n+1} \sum_{k=i-n}^{N} stat(\mathbf{t}_i). \end{aligned} \tag{10}$$

В полученном ряде статистик происходит поиск выбросов, которые помечаются как некорректно заимствованные.

Также в главе описывается эксперимент, в рамках которого показывается работоспособность метода и производится сравнение с другими методами решения данной задачи.

В **главе 4** рассматривается родственная задача по поиску границ смены авторского стиля. Также как и в главе 3 описывается система поиска заимствований с указанием подозрительных участков текста. Ее отличие

заключается в том, что она использует вспомогательную модель векторизации предложений. В качестве вспомогательной модели векторизации была выбрана модель Skip-Thought Vectors. Данная модель представляет из себя нейросеть, параметры которой настроены для того, чтобы возвращать репрезентативный вектор входного предложения на естественном языке без использования какой-либо дополнительной информации. Причем вектора строятся с учетом предыдуще го и последующего предложений, т.е. модель учитывает контекст и порядок следования предложений в тексте.

В качестве процедуры сегментирования в данной задаче выбран процесс разбиения на предложения. Статистика для сегмента текста строится путем рассчета средней косинусной меры до всех остальных сегментов:

$$stat(s_i) = \frac{1}{|d|} \sum_{j \neq i} \cos(\mathbf{s}_i, \mathbf{s}_j), \tag{11}$$

Приводится описание вычислительного эксперимента, в рамках которого предложенный метод сравнивается с моделями, которые не используют вспомогательные модели векторизации текстов. Показывается, что метод хорошо справляется с отбором высокооригинальных текстов (т.е. тех, в которых нет некорректных заимствований).

В **главе 5** приводится описание алгоритма отбора документов, не содержащих некорректных текстовых заимствований. Так как постановка задачи не подразумевает указания конкретных сегментов текста, задача сводится к задаче бинарной классификации:

– 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.

Общую логику работы предлагаемого алгоритма можно представить в виде псевдокода (1).

Стратегия сегментации выбирается исходя из конкретного корпуса документов, на котором настраиваются параметры алгоритма. Рассматривается разбиение по параграфам и разбиение окном с фиксированным шагом.

Рассчет частотной статистики для символьной (или словесной) n-граммы в тексте d происходит по следующей формуле:

$$freq_w = \frac{cnt(w)}{\sum\limits_{w' \in d} cnt(w')} \cdot \log\left(\frac{m}{seg(w)}\right), \tag{12}$$

где $cnt(w)$ – число вхождений w в текст d, m – число сегментов в тексте, $seg(w)$ – число сегментов, содержащих w. Из рассчитанных значений формируется вектор сегмента, как описано в главе 2.

Algorithm 1: Алгоритм определения факта заимствования в тексте

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

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

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

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

- евклидово расстояние,
- косинусное расстояние.

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.

Предложенный алгоритм настраивается и тестируется на корпусах английских и русских текстов. Показывается, что в обоих случаях алгоритм корректно отбирает высокооригинальные документы, оставляя при этом документы с заимствованиями для дальнейшей проверки.

Также в главе описана принципиальная схема разработанного программного комплекса, который внедряется в промышленную систему поиска текстовых заимствований.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Предложен способ векторизации фрагментов текста, основанный на частотах встречаемости символьных и словесных n-грамм в анализируемом тексте и в каждом фрагменте по отдельности.
2. Разработан способ обнаружения заимствованных фрагментов текста, основанный на сегментировании анализируемого текста и анализе ряда статистик, построенных для каждого из полученных сегментов, на предмет наличия выбросов.
3. Разработан метод обнаружения и фильтрации высокооригинальных текстовых документов без использования внешней коллекции потенциальных источников
4. Обоснована работоспособность предложенного алгоритма путем реализации и тестирования на подготовленных данных. Экспериментально показано, что предложенный алгоритм может отфильтровывать до 30% высокооригинальных документов, не сильно проигрывая в качестве полноценной проверке.
5. Разработана и внедрена программная реализация предложенного метода.

Публикации автора по теме диссертации

A1. *Safin, K.* Style Breach Detection with Neural Sentence Embeddings / K. Safin, R. Kuznetsova // Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. Vol. 1866 / ed. by L. Cappellato [et al.]. — CEUR-WS.org, 2017. — (CEUR Workshop Proceedings).

A2. *К. Ф. Сафин.* Определение заимствований в тексте без указания источника / К. Ф. Сафин, М. П. Кузнецов, М. В. Кузнецова // Информ. и её примен. — 2017. — т. 11, № 3.

A3. *Safin, K.* Detecting a Change of Style using Text Statistics: Notebook for PAN at CLEF 2018 / K. Safin, A. Ogaltsov // Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. Vol. 2125 / ed. by L. Cappellato [et al.]. — CEUR-WS.org, 2018. — (CEUR Workshop Proceedings).

A4. Near-duplicate handwritten document detection without text recognition / O. Bakhteev [et al.] // Computational Linguistics and Intellectual Technologies. — 2021.

А5. *К. Ф. Сафин.* О комбинированном алгоритме обнаружения заимствований в текстовых документах / К. Ф. Сафин, Ю. В. Чехович // Труды Института системного программирования РАН. — 2022. — т. 34, № 1. — с. 151—160.

Список литературы

1. *Никитов, А. В.* Плагиат в работах студентов и аспирантов: проблема и методы противодействия / А. В. Никитов, О. А. Орчаков, Ю. В. Чехович //. — 2012.
2. *Stein, B.* Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07 / B. Stein, M. Koppel, E. Stamatatos // SIGIR Forum. — 2007. — Vol. 41, no. 2. — P. 68—71. — URL: https://doi.org/10.1145/1328964.1328976.
3. *Chekhovich, Y. V.* Analysis of duplicated publications in Russian journals / Y. V. Chekhovich, A. V. Khazov // Journal of Informetrics. — 2022. — Vol. 16, no. 1. — P. 101246. — URL: https://www.sciencedirect.com/science/article/pii/S1751157721001176.
4. *Зеленков, И. В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов / И. В. Зеленков, И. В. Сегалович // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. 9-й Всеросс. научн. конф. RCDL. — Переславль-Залесский: Университет г. Переславля. — 2007.
5. Система распознавания интеллектуальных заимствований «Антиплагиат» / Ю. Журавлев [и др.] // Математические методы распознавания образов: 12-я Всероссийская конференция: Сборник докладов. — 2005.
6. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection / R. Socher [et al.] // NIPS. — 2011.
7. *Кузнецова, Р. В.* Методы обнаружения переводных заимствований в больших текстовых коллекциях / Р. В. Кузнецова, О. Ю. Бахтеев, Ю. В. Чехович // Информатика и её применения. — 2021. — т. 15, № 1. — с. 30—41.
8. *Е. М. Ешилбашян.* Поиск заимствований в армянских текстах путем внутреннего стилометрического анализа / Е. М. Ешилбашян, А. А. Асатрян, Ц. Г. Гукасян // Труды ИСП РАН. — 2021. — т. 33, № 1. — с. 209—224.
9. *Eissen, S. M. z.* Intrinsic Plagiarism Detection / S. M. z. Eissen, B. Stein // Advances in Information Retrieval. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2006. — P. 565—569.

10. *Muhr, M.* External and Intrinsic Plagiarism Detection Using Vector Space Models / M. Muhr, M. Zechner, R. Kern // CEUR Workshop Proceedings. — 2009. — Jan. — Vol. 502.

11. Outlier-Based Approaches for Intrinsic and External Plagiarism Detection / G. Oberreuter [et al.] // KES. — 2011.

12. *Stamatatos, E.* Intrinsic Plagiarism Detection Using Character n-gram Profiles / E. Stamatatos //. — 2009.

13. *Bensalem, I.* Intrinsic Plagiarism Detection using N-gram Classes / I. Bensalem, P. Rosso, S. Chikhi //. — 01/2014.

14. *Tschuggnall, M.* Countering Plagiarism by Exposing Irregularities in Authors' Grammar / M. Tschuggnall, G. Specht // Proceedings - 2013 European Intelligence and Security Informatics Conference, EISIC 2013. — 2013. — Aug. — P. 15–22.

15. *Романов, А. С.* Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции / А. С. Романов, Р. В. Мещеряков, З. И. Резанова // Доклады Томского государственного университета систем управления и радиоэлектроники. — 2014.

16. Methods for Intrinsic Plagiarism Detection and Author Diarization / M. P. Kuznetsov [et al.] // Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. Vol. 1609 / ed. by K. Balog [et al.]. — CEUR-WS.org, 2016. — P. 912–919. — (CEUR Workshop Proceedings). — URL: http://ceur-ws.org/Vol-1609/16090912.pdf.

17. *Gillam, L.* Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification / L. Gillam, A. Vartapetiance. — 2012.

18. Overview of the 3rd International Competition on Plagiarism Detection. / M. Potthast [et al.] //. — 01/2011.

19. *Stamatatos, E.* A survey of modern authorship attribution methods / E. Stamatatos // J. Assoc. Inf. Sci. Technol. — 2009. — Vol. 60, no. 3. — P. 538–556. — URL: https://doi.org/10.1002/asi.21001.

20. *Jones, K. S.* A statistical interpretation of term specificity and its application in retrieval / K. S. Jones // Journal of Documentation. — 1972. — Vol. 28. — P. 11–21.