УДК 004.896 **DOI** 10.15622/ia.20.4.5

Ф.В. Краснов, И.С. Смазневич, Е.Н. Баскакова

ОПТИМИЗАЦИОННЫЙ ПОДХОД К ВЫБОРУ МЕТОДОВ ОБНАРУЖЕНИЯ АНОМАЛИЙ В ОДНОРОДНЫХ ТЕКСТОВЫХ КОЛЛЕКЦИЯХ

Краснов Ф.В., Смазневич И.С., Баскакова Е.Н. **Оптимизационный подход к выбору методов обнаружения аномалий в однородных текстовых коллекциях.**

Аннотация. Рассматривается задача обнаружения аномальных документов в текстовых коллекциях. Существующие методы выявления аномалий не универсальны и не показывают стабильный результат на разных наборах данных. Точность результатов зависит от выбора параметров на каждом из шагов алгоритма, и для разных коллекций оптимальны различные наборы параметров. Не все из существующих алгоритмов обнаружения аномалий эффективно работают с текстовыми данными, векторное представление которых характеризуется большой размерностью при сильной разреженности.

Задача поиска аномалий рассматривается в следующей постановке: требуется проверить новый документ, загружаемый в прикладную интеллектуальную информационную систему (ПИИС), на соответствие хранящейся в ней однородной коллекции документов. В ПИИС, обрабатывающих юридически значимые документы, на методы обнаружения аномалий накладываются следующие ограничения: высокая точность, вычислительная эффективность, воспроизводимость результатов, а также объяснимость решения. Исследуются методы, удовлетворяющие этим условиям.

В работе изучается возможность оценки текстовых документов по шкале аномальности путем внедрения в коллекцию заведомо инородного документа. Предложена стратегия обнаружения в документе новизны по отношению к коллекции, предполагающая обоснованный подбор методов и параметров. Показано, как на точность решения влияет выбор вариантов векторизации, принципов токенизации, методов снижения размерности и параметров алгоритмов поиска аномалий.

Эксперимент проведен на двух однородных коллекциях нормативно-технических документов: стандартов в отношении информационных технологий и в сфере железных дорог. Использовались подходы: вычисление индекса аномальности как расстояния Хеллингера между распределениями близости документов к центру коллекции и к инородному документу; оптимизация алгоритмов поиска аномалий в зависимости от методов векторизации и снижения размерности. Векторное пространство строилось с помощью преобразования TF-IDF и тематического моделирования ARTM. Тестировались алгоритмы Isolation Forest (изолирующий лес), Local Outlier Factor (локальный фактор выброса), One-Class SVM (вариант метода опорных векторов).

Эксперимент подтвердил эффективность предложенной оптимизационной стратегии для определения подходящего метода обнаружения аномалий для заданной текстовой коллекции. При поиске аномалии в рамках тематической кластеризации юридически значимых документов эффективен метод изолирующего леса. При векторизации документов по TF-IDF целесообразно подобрать оптимальные параметры словаря и использовать метод опорных векторов с соответствующей функцией преобразования признакового пространства.

Ключевые слова: выявление аномалий, выявление новизны, выявление выбросов, однородные текстовые коллекции, уменьшение размерности разреженных пространств, тематическое моделирование.

1. Введение. Алгоритм выявления аномалий в текстовых данных может выступать в качестве одного из компонентов решения во многих прикладных задачах. В полнотекстовом поиске при ранжировании поисковой выдачи можно учитывать количество новой информации, содержащейся в найденных документах. При распределении входящих обращений по отделам или экспертам полезно выявлять заявки, которые не могут быть никем обработаны - например, потому что отправлены как спам или по ошибке. В процессе тематической рубрикации массива документов необходимо помечать документы, которые не вписываются ни в одну из существующих тем. Кроме того, выявление аномалий в коллекциях важно для поддержания их однородности, влияющей на качество решения при выделении фактов, поиске противоречий и в других задачах интеллектуального анализа текста.

Для прикладных интеллектуальных информационных систем (ПИИС) необходимы точные методы обнаружения аномалий, показывающие высокую эффективность на текстовых документах, которые представляются пространством признаков высокой размерности и объединяются в коллекции больших размеров.

Аномалии в текстовых данных требуется обнаруживать на разных уровнях, в зависимости от практической задачи: на уровне тем [1], на уровне документов [2], на уровне предложений [3] и в некоторых случаях на уровне слов (например, в работе [4] как результат поиска аномалий выявляются новые значения слов). Есть и обратная задача - идентифицировать в потоке текстовых данных те элементы (документы, обращения, сообщения), которые не добавляют новой информации к уже накопленной по данной теме [5].

Рассмотрим формальное определение несоответствия нового документа документам исходной коллекции. Допустим, что коллекция обладает распределением P по признакам p_i. Тогда новый документ d может принадлежать либо P, либо другому распределению Q, и во втором случае он является аномальным (отличается новизной). При этом о степени аномальности документа говорит величина $D(P,Q)$, характеризующая расхождение между распределениями P и Q.

В случае с текстами для каждого документа коллекции создается векторное представление в пространстве $p \in \mathrm{R}^{1\times|W|}$, где $|W|$ — натуральное число. Частным случаем является значение $|W|$, равное количеству уникальных слов в коллекции, однако $|W|$ может быть меньше – если в результате снижения размерности получаются плотные вектора, или

больше – когда в качестве термов рассматриваются n-граммы. Для коллекции из $|D|$ документов создается векторное пространство $R^{|D|\times|W|}$. Для добавляемого документа d также может быть получено векторное представление – в пространстве $p \in R^{1\times|\hat{W}|}$. Добавление нового документа в коллекцию может по-разному повлиять на изменение пространства коллекции в зависимости от того, как строится вектор нового элемента – обучая модель заново (полностью обновляя матрицу на основе изменившегося словаря коллекции) либо достраивая модель (формируя вектор нового документа на основе исходного словаря). В первом случае словарь дополняется: $|W| \to |W \cup \hat{W}|$; во втором остается без изменений: $|W| \to |W|$.

Рассмотрим подробнее, что подразумевается под обнаружением аномалии. Результатом работы алгоритма по обнаружению аномалий (неважно, новизны или выбросов) является вектор длины $|D|$, состоящий из рациональных чисел, характеризующих степень аномальности каждого документа. В описаниях алгоритмов эти рациональные числа называются по-разному: степенью или индексом аномальности [6], скорингом либо фактором аномальности (преимущественно в англоязычных работах) [7] или decision-функцией [8]. Аномальными считаются элементы, индекс аномальности которых превышает заданное пороговое значение (или оказывается ниже его), либо к таковым относится некоторое число элементов с наибольшим (или наименьшим) индексом аномальности.

Обнаружение аномалий можно условно разделить на две задачи: обнаружение выбросов (outlier detection) и обнаружение новизны (novelty detection). Обнаружение выбросов подразумевает анализ коллекции с целью выявления нестандартных документов среди всех имеющихся. Задача обнаружения новизны предполагает оценку нового документа и является актуальной для однородной коллекции, собранной по одному или нескольким критериям: предметная область, жанр и стиль текста, назначение информации и способ ее применения, сходная структура документов. Примером такой коллекции может быть база технических заданий, подборка нормативно-правовых актов по определенной сфере деятельности, договоры или доверенности организации. В этом случае задача обнаружения новизны состоит в следующем: если поступающий новый документ «не соответствует» данной коллекции, он должен быть идентифицирован как инородный элемент по отношению к ней. Однородность самой коллекции имеет большое значение для поиска аномально новых документов, поскольку при наличии выбросов

в коллекции значение индекса аномальности у этих элементов может оказаться выше, чем у инородного нового документа.

Данная работа посвящена задаче обнаружения новизны – выявления нового текстового документа как инородного по отношению к заданной коллекции. На рисунке 1 схематично показан каркас исследования.

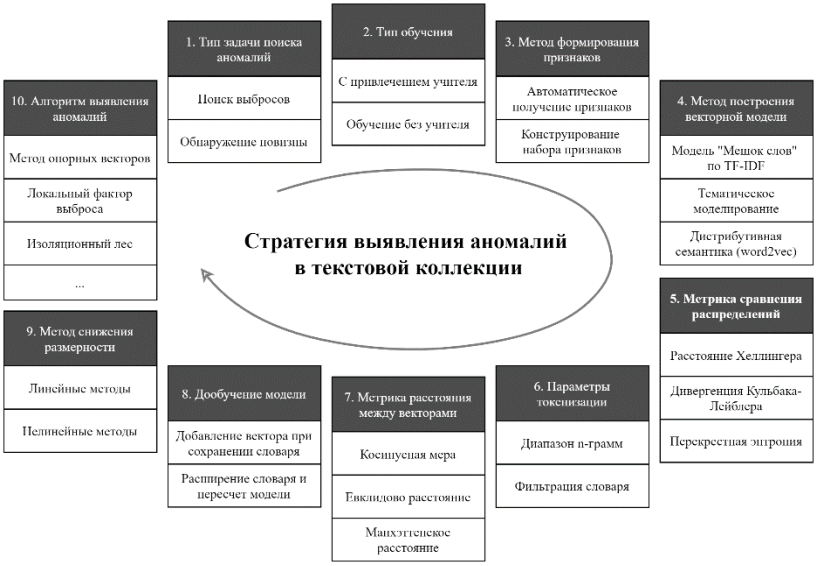

Рис. 1. Каркас исследования: этапы решения задачи обнаружения аномалий в текстовой коллекции

Решение задачи по обнаружению аномалий состоит из нескольких этапов, каждый из которых допускает несколько вариантов подходов или методов, от выбора которых зависит итоговая эффективность всего алгоритма. Нумерация этапов в общем случае соответствует последовательности шагов при поиске аномалий в текстовой коллекции. Однако при выборе определенных методов некоторые этапы могут быть пропущены, тогда как другие становятся обязательными.

Цель настоящего исследования - определить стратегию выбора оптимальных методов и их параметров для обнаружения аномальности нового документа по отношению к существующей текстовой коллекции с учетом ограничений, накладываемых прикладными информационными системами.

С учетом каркаса исследования в качестве методики в данной работе используется оптимизационный подход: определение функционала в пространстве свободных параметров различных стратегий (цепочек алгоритмов) и поиск оптимальной стратегии.

Новизна подхода состоит в предложенной стратегии оптимизации: методы и параметры подбираются для заданной коллекции документов, и при оптимизации используется образец аномальности – заведомо инородный для коллекции документ. Параметры алгоритма считаются оптимальными, когда минимизировано число элементов коллекции, обладающих степенью аномальности выше, чем у инородного документа; в случае однородной коллекции при оптимальных параметрах алгоритма инородный документ обладает максимальной степенью аномальности среди всех элементов коллекции.

Статья состоит из введения, обзора основных алгоритмов поиска аномалий для текстовых коллекций, методики формирования стратегии выявления аномально новых документов, эксперимента по поиску оптимальных стратегий выявления аномально новых документов для двух наборов данных, описания результатов и возможностей их инженерного применения.

2. Обзор. Исследования по алгоритмам обнаружения аномалий активно ведутся с 1980-х годов [9]. Методы решения задачи поиска аномалий в текстовых коллекциях могут быть классифицированы различным образом. В работе [10] подходы к выявлению аномалий делятся на две категории: статистические и нейросетевые. В работе [11] различные методы анализируются в том числе с точки зрения применимости к той или иной прикладной задаче, определяемой типом исследуемых данных (изображения, тексты, медицинские сведения, банковские транзакции, биржевые сделки, статистика телефонных звонков и т.д.). Для анализа текстовых коллекций предлагается использовать такие методы, как моделирование на основе смеси распределений, статистическое профилирование с использованием гистограмм, метод опорных векторов, нейронные сети, кластеризацию.

В обзорной статье [12] рассматривается общая задача обнаружения новизны в коллекциях элементов без учета формата (типа данных) этих элементов, и выделяются следующие подходы к ее решению: вероятностное обнаружение аномальных элементов; методы на основе вычисления расстояний в пространстве коллекции; методы снижения размерности и последующей реконструкции; обнаружение аномалий с учетом знаний о предметной области. В [13] анализируются различные аспекты обнаружения новизны в потоке данных: автономная и онлайновая фазы алгоритма, количество классов, рассматриваемых на каждой

фазе, ансамблевые методы сравниваются с решением на основе единственного классификатора, рассматриваются подходы к обучению (с учителем и без него), а также вопросы обновления модели принятия решений и другие.

С ростом интереса к нейросетевым подходам классификация методов обнаружений аномальных элементов претерпела изменения. Так, авторы обзорной работы этого года [14] выделяют две обширные категории методов поиска аномалий в зависимости от коэффициентов нейросетевой модели (feature map): глубокие и неглубокие (deep и shallow), относя к последним и те методы, которые формально не являются нейросетевыми.

Современные методы обнаружения аномалий в текстовых данных сталкиваются с несколькими проблемами, включая сильную разреженность данных, зависимость от метрики расстояния пространства признаков, возникновение большого количества кластеров, а также неизвестные признаки аномальных элементов. Актуальность той или иной проблемы зависит от конкретной прикладной задачи, определяемой бизнес-процессом и особенностями текстового корпуса.

Полный алгоритм обнаружения аномалий в реальной прикладной информационной системе складывается из последовательности шагов, на каждом из которых решается определенная подзадача и допускается несколько вариантов методов ее решения. На рисунке 1 показана совокупность подзадач и методов, которые были рассмотрены в рамках настоящего исследования.

2.1. Тип задачи: поиск выбросов или обнаружение новизны. Большинство методов поиска аномалий вычисляют функцию оценки выброса (outlier score), характеризующую степень аномальности каждого объекта пространства, а также определяют пороговые значения этой оценки для обнаружения инородных элементов коллекции. Наиболее широко используемые методы пороговой обработки основаны на таких статистических данных, как стандартное отклонение от среднего, медианное абсолютное отклонение и межквартильный диапазон. К сожалению, при проверке новых элементов на аномальность по отношению к исходной коллекции такая статистика может оказаться необъективной, искажаясь за счет имеющихся выбросов. Минимизировать их влияние на вычисление оценки аномальности можно за счет предъявления дополнительных требований к коллекции и уточнения условий задачи: к однородной коллекции добавляется ровно один документ, который анализируется на предмет аномальности; за точку отсчета при оценке принимается центр коллекции. Такая постановка задачи рассматривается в настоящем исследовании.

2.2. Тип обучения: с привлечением учителя и без него. Для обнаружения аномалий в текстовой коллекции могут применяться разные подходы к обучению. Для решения этой задачи могут использоваться методы с учителем (supervised) либо с частичным привлечением учителя (semi-supervised) [1, 15]. Выявление аномалий в режиме без учителя также применяется [16], будучи востребованным в ситуациях, когда невозможно сообщить алгоритму предполагаемые характеристики аномальных документов или предоставить их образцы. Кроме того, такие методы не требуют трудоемкой разметки большого количества данных на основании экспертных знаний. Однако для задачи поиска несоответствующих элементов в коллекции объемных текстовых документов, которая рассматривается в рамках данного исследования, наиболее подходят методы обучения с частичным привлечением учителя, поскольку в реальных прикладных системах известны общие характеристики «корректных» документов, и у пользователей есть представление о том, какие документы точно не должны попасть в коллекцию.

2.3. Принцип формирования набора признаков. При поиске аномалий документы рассматриваются как объекты с некоторым набором признаков. Векторы признаков могут формироваться автоматически, при построении модели коллекции на основе статистических данных словаря, либо в результате процесса конструирования признаков (feature engineering) [17], когда для каждого из документов определяется набор характеристик, которые с точки зрения эксперта являются значимыми для разделения ядра коллекции и ее аномалий. Например, такими признаками могут быть именованные сущности, которые встречаются в тексте, или части речи каждого из слов [18]. При этом числовые значения признаков также могут быть вычислены в автоматическом режиме, в том числе на основе статистики словаря или характеристик графового представления текстов [19].

Плюс первого метода состоит в его универсальности и независимости от предметной области, что на практике означает широкую сферу применения алгоритма и отсутствие необходимости существенной перенастройки прикладной информационной системы для каждого заказчика. К преимуществам второго способа формирования признаков можно отнести умеренную вычислительную сложность и возможность максимальной адаптации под конкретную прикладную задачу.

Следует отметить, что автоматическое конструирование признаков без учета специфики обрабатываемых данных применимо не всегда. Например, метод COPOD (Copula-Based Outlier Detection) [20], демонстрирующий результативность на разнообразных наборах данных, опирается на набор признаков, сформированный на основании копулы -

функции от многомерного распределения, позволяющей отделить частные распределения от структуры зависимостей данного многомерного распределения. Существенным преимуществом этого метода является отсутствие параметров (в отличие от других методов, где выбор параметров способен существенно повлиять на результат) и высокая производительность, в том числе на пространствах большой размерности. Однако заявленная в статье эффективность метода не была экспериментально показана его авторами на текстовых корпусах и не подтвердилась в рамках настоящего исследования, поэтому метод был исключен из рассмотрения.

2.4. Метод построения векторной модели. Базовым вариантом автоматического конструирования набора признаков для коллекции текстовых документов является построение их векторного представления на основе статистических характеристик словаря коллекции. В основе такого представления лежит модель Bag Of Words («мешок слов», BoW), в которой порядок слов в исходных текстах не учитывается.

При отсутствии или неявной выраженности содержательной специфики коллекции документов пространство признаков может быть построено с помощью методов дистрибутивной семантики на основе уже известных данных о распределении слов в универсальных языковых корпусах (word2vec [21]).

Однако, если учитывать особенности лексики документов в рамках определенной прикладной системы, точность решения может оказаться существенно выше. Для реализации этого преимущества строится матрица «терм-документ», содержащая веса, вычисляемые как значения функции TF-IDF.

Еще одним способом построения векторной модели текстовой коллекции является тематическое моделирование. В этом случае для представления коллекции строятся две модели: «терм-тема» и «тема-документ», и вектор каждого документа содержит веса тем. Плюсом такого представления текстовой коллекции для обнаружения аномалий является гораздо меньшая, по сравнению с предыдущими двумя способами, размерность пространства признаков, а также более высокая объяснимость решения, что важно при разработке прикладных информационных систем. Однако процесс построения тематической модели требует больше вычислительных ресурсов и более чувствителен к выбору параметров алгоритма.

2.5. Метрика сравнения распределений. Необходимым этапом при решении задачи поиска аномалий в коллекции является выбор метрики аномальности нового документа (скоринговой функции, скоринга аномальности).

В решаемой задаче вычисление скоринга аномальности нового документа может быть сделано на основе сравнения расстояния от каждого документа исходной коллекции до двух объектов: до ее центрального элемента и до нового документа. Полученные два набора расстояний между документами сравниваются между собой с помощью метрики различия между распределениями.

В качестве такой метрики могут быть рассмотрены, например, расстояние Хеллингера, дивергенция Кульбака-Лейблера, перекрестная энтропия и другие функции. На этом этапе решения задачи необходимо подобрать наиболее выразительную метрику сравнения распределений для данной коллекции текстовых документов, которая и станет результирующей скоринговой функцией для рассматриваемых входных данных: коллекции документов и образца инородного документа.

2.6. Параметры токенизации текста. При построении векторной модели коллекции на основе статистики словаря возможны различные варианты токенизации текста: в качестве термов могут рассматриваться униграммы, *n*-граммы, субсловарные единицы (части слов) разной длины либо их комбинация. Выбор принципа токенизации влияет на вычислительную сложность алгоритма и одновременно на его точность. При этом для некоторых прикладных задач допустимо не приводить слова к начальной форме. Удаление из словаря наиболее часто и наиболее редко встречающихся слов также опционально, как и любая другая его фильтрация.

2.7. Метрика расстояния в векторном пространстве. В построенном векторном пространстве документов коллекции необходимо задать метрику расстояния между элементами. Близость между документами в таком представлении может быть измерена различными способами, наиболее объяснимыми среди которых являются Манхэттенское расстояние, Евклидово расстояние и косинусная мера. При этом оптимальной метрикой пространства при работе с векторным представлением текстовых данных является косинусная мера, что, в частности, продемонстрировано в работе [22].

2.8. Способ дообучения модели. При добавлении в коллекцию нового документа, который должен быть оценен на предмет аномальности по отношению к данной коллекции, для него необходимо построить вектор в имеющемся семантическом пространстве. В случае, если используются методы дистрибутивной семантики, вектор создается по тем же принципам, что и вся модель. Если же используется метод построения матрицы коллекции по TF-IDF или строится тематическая модель, то существуют два варианта создания вектора для нового документа.

Первый способ - построение вектора нового документа в имеющемся пространстве без увеличения его размерности. В этом случае словарь коллекции не меняется, поскольку новый документ не обогащает его. Второй способ - пересчет всей модели с учетом добавления в коллекцию нового документа. В этом случае словарь расширяется за счет слов нового документа.

При решении реальных задач в рамках прикладных информационных систем выбор варианта построения вектора нового документа определяется имеющимися вычислительными ресурсами и необходимой скоростью, а также предполагаемым сценарием работы с новым документом. При последующем сохранении нового документа в системе полный пересчет модели является оправданным решением, а если аномальный документ удаляется сразу после анализа, то включать его слова в словарь модели нецелесообразно, и вектор должен рассчитываться по существующему словарю.

2.9. Методы снижения размерности векторов. При формировании векторных представлений текстовых коллекций на основе статистики их словаря строятся пространства большой размерности, при этом векторы документов получаются сильно разреженными. Это затрудняет работу многих алгоритмов определения аномалий, делая их неэффективными для анализа текстовых данных. Такие алгоритмы требуют предварительной процедуры понижения размерности матриц.

Уменьшение размерности с помощью линейных алгоритмов, таких как метод главных компонент (PCA) и разложение по сингулярным числам (TruncatedSVD), не приносит большой пользы из-за того, что эти алгоритмы линейные. Даже сохраняя в модели большую часть разнообразия данных, при сворачивании измерений они не способны к нелинейным преобразованиям, которые могут учитывать информацию о совстречаемости слов.

В то же время алгоритмы, основанные на нелинейных преобразованиях (такие как SNE и его модификации [23]) не сохраняют отношения плотности и расстояний. Однако в случае разреженных данных сохранение локальных областей с высокой плотностью приобретает особую важность. Это способны делать некоторые нелинейные алгоритмы, например, UMAP [24]. Алгоритмы denSNE и densMAP [25] могут сохранять информацию о плотности при правильном выборе параметров: они вычисляют оценки локальной плотности и используют эти оценки в качестве регуляризатора при оптимизации низкоразмерного представления.

2.10. Алгоритмы выявления аномалий. После построения векторной модели коллекции и установления необходимых метрик следующим этапом в решении задачи является выбор метода поиска аномальных элементов.

В рамках данного исследования изучается применимость алгоритмов, относящихся, согласно классификации [14], к категории неглубоких. В качестве базового требования к алгоритмам, исходящего из реального опыта разработки прикладных систем, авторами была установлена достаточно хорошая объяснимость получаемых решений. По этой причине не рассматривались нейросетевые типы алгоритмов, относящиеся в указанном обзоре к категории deep. Кроме того, исключались из рассмотрения вероятностные алгоритмы, что обусловлено требованиями воспроизводимости результатов в реальных прикладных информационных системах.

В таблице 1 для различных групп методов машинного обучения показано сопоставление их объяснимости и точности, а также указана их вычислительная сложность.

Таблица 1. Общие характеристики методов обнаружения аномалий

Метод	**Точность**	**Объяснимость**	**Сложность вычислительная**
Логистическая регрессия	Низкая	Высокая	$O(n \cdot d)$, где n - количество элементов, d - размерность; используется для данных низкой размерности.
Дерево решений	Низкая	Высокая	$O(n \cdot \log(n) \cdot d)$ - вычислительная сложность обучения модели; $O(n)$ - сложность во время выполнения; используется для данных низкой размерности.
Метод ближайших соседей (локальный фактор выброса)	Средняя	Средняя	$O(k \cdot n \cdot d)$, где n - количество элементов, d - размерность, k - количество соседей.
Кластеризация	Средняя	Средняя	Зависит от алгоритма кластеризации: $O(n \cdot k \cdot l)$, где k – число кластеров, l – число итераций для k-means.

Продолжение таблицы 1.

Метод	**Точность**	**Объяснимость**	**Сложность вычислительная**
Метод опорных векторов	Средняя	Средняя	$O(n^2)$ или $O(n^3)$ - вычислительная сложность обучения модели в зависимости от используемого ядра; $O(k \cdot d)$ - сложность во время выполнения, где k - количество опорных векторов, d - размерность данных.
Ансамблевые методы (случайный лес)	Высокая	Низкая	$O(n \cdot \log(n) \cdot d \cdot k)$ - вычислительная сложность обучения модели, где k - количество деревьев; $O(n \cdot k)$ - сложность во время выполнения.
Нейронные сети	Высокая	Низкая	$O(n^4)$ – прямое распространение; $O(n^5)$ – обратное распространение ошибки.

Исходя из требований реальной задачи – точности решений при достаточно высоком уровне их объяснимости и воспроизводимости (что на практике означает надежность рекомендаций информационной системы) – в рамках данной работы для построения оптимальной стратегии выявления аномалий в коллекции документов рассматриваются следующие методы: локальный уровень выброса, мягкая кластеризация, метод опорных векторов, а также один из ансамблевых методов – изолирующий лес.

Метод ближайших соседей характеризуется наиболее высокой объяснимостью, поскольку интуитивно понятно, что в векторном пространстве у аномалии мало близких соседей, а у типичной точки (относящейся к т. н. ядру коллекции) их много. Поэтому величина «расстояние до *k*-го соседа» может служить мерой аномальности документа. Метод хорошо работает на однородных коллекциях, позволяющих задать единый порог аномальности для всех элементов коллекции, либо на коллекциях, состоящих из нескольких кластеров, плотность которых примерно равна.

В то же время для менее однородных коллекций значимым может оказаться обнаружение локальных аномалий, то есть выбросов, в том

числе для задачи приведения коллекции к однородному состоянию. Метод локального фактора выброса (Local Outlier Factor, LOF) позволяет присвоить каждому объекту степень локальной аномальности (локальный фактор выброса). Она показывает, насколько объект изолирован от других в окружающей его окрестности, то есть от совокупности его ближайших соседей. Преимущество LOF перед другими алгоритмами обнаружения аномалий проявляется в способности обрабатывать наборы данных с кластерами разной плотности.

Результаты работы алгоритмов зависят не только от выбранных вариаций и параметров, но и от свойств набора данных, на которых этот алгоритм определения аномалий применяется. В работе [26] экспериментально показано, что алгоритм ближайших соседей по-разному работает на различных дата-сетах, хотя все они репрезентативны и созданы специально для исследования задачи выявления аномалий.

Методы кластеризации определяют аномалии как элементы, которые не могут быть включены ни в один из обнаруженных кластеров. Кластеризация может проводиться по различным критериям: по тематической принадлежности элементов, по плотности их распределения, на основании дополнительных признаков.

Кластеризация по плотности распределения [27] для текстовых документов неэффективна из-за большой размерности пространства в совокупности с сильной разреженностью векторных представлений. В матрицах модели BoW доля ненулевых значений составляет около 1% и даже менее, что затрудняет формирование кластеров. Поэтому такие алгоритмы кластеризации, как HDBSCAN [28], OPTICS [29], неприменимы для полноразмерных векторных представлений текста.

Алгоритм кластеризации CHAMELEON [30], основанный на сегментации графа методами семейства METIS [31], также подвержен «проклятию размерности». Алгоритм работает со связными графами, тогда как по разреженному матричному представлению формируются многочисленные связные компоненты малого размера, и анализ такого графа становится неэффективным.

В работе [32] для определения аномалий используется алгоритм кластеризации с помощью модификации неотрицательного матричного разложения (Non-negative Matrix Factorization, NMF), и аномальными считаются документы, которые не соответствуют построенной тематической модели.

Кроме того, кластеры сами могут быть основой для формирования набора признаков: например, в работе [33] описана вероятностная кластеризация, и характеристики кластеров передаются классификатору в качестве дополнительных признаков элементов коллекции.

Метод опорных векторов (Support Vector Machine, SVM) может рассматриваться в задаче обнаружения аномалий как метод кластеризации с единственным кластером. Этот подход реализован в алгоритме One-Class SVM [34]. Будучи чувствительным к выбросам, он лучше подходит для обнаружения новизны, когда обучающая выборка достаточно однородна. Тем не менее, обнаружение выбросов в пространстве большой размерности, а также без каких-либо предположений о распределении входящих данных является очень сложной задачей, и One-Class SVM может оказаться полезным, если выбрать подходящие значения его гиперпараметров.

Ансамблевые методы поиска аномалий как правило предполагают обучение с привлечением учителя, хотя ансамблям без учителя посвящены некоторые исследования [35]. Чтобы автоматизировать оценку эффективности различных ансамблевых методов, в работе [36] предлагается использовать скоринговые функции.

К ансамблевым методам относится метод изолирующего леса (Isolation Forest, iForest [37]), реализованный на основе ансамбля Extra Tree Regressor. В авторском варианте максимальная глубина каждого дерева устанавливается равной количеству сэмплов, использованных для построения дерева, что увеличивает скорость работы метода. Кроме того, iForest допускает распараллеливание при вычислении, что также ускоряет их. К плюсам этого метода также относится его способность эффективно работать с мультимодальными наборами данных; то же характеризует и метод LOF.

Обзор большого числа научных работ с описанием методов и модификаций, демонстрирующих хорошую точность и высокую эффективность, позволяет сделать вывод, что результаты достоверны преимущественно лишь на описываемых экспериментальных данных либо на аналогичных им. В частности, не все методы тестируются на текстовых корпусах, в том числе на русскоязычных - большинство экспериментов проводится с данными на английском языке. Поэтому, несмотря на результативность представленных во многих работах алгоритмов, они не могут быть непосредственно применены для анализа русскоязычных документов [38], для этого требуется их адаптация с учетом специфики русского языка. Кроме того, очевидно, что такие характеристики корпуса, как длина текстов, широта тематики, структурированность документов и однородность коллекции имеют большое значение при выборе алгоритма поиска аномалий.

3. Методы. В рамках данной работы изучалась возможность выбора оптимального алгоритма для обнаружения аномальности нового

документа по отношению к однородной коллекции русскоязычных текстовых документов средней и большой длины, обладающих юридической значимостью.

Для решения задачи была выбрана следующая стратегия: среди различных алгоритмов обнаружения аномалий подобрать оптимальный для конкретной коллекции документов, с учетом циклического перебора гиперпараметров на каждом шаге алгоритма и наличия образца аномального документа.

Рассматривались методы поиска аномалий из числа достаточно объяснимых и способных обеспечивать приемлемую точность. Первоначальный выбор методов был обусловлен накопленными у авторов сведениями об их объяснимости, характеристиками вычислительной сложности, а также известной точностью результатов. Исходя из этих требований нейросетевые методы не рассматривались. Таким образом, для исследования были выбраны следующие методы: локальный фактор выброса (LOF), изоляционный лес (iForest), метод опорных векторов для одного класса (One-Class SVM).

Для целей исследования была проведена серия экспериментов. В качестве экспериментальных данных были выбраны две однородные коллекции нормативных документов, представляющих собой структурированные тексты большой и средней длины, схожие по тематике. Векторное представление коллекции документов формировалось двумя способами: на основе функции TF-IDF и методом тематического моделирования.

Для выбора метрики был использован следующий подход в рамках обучения с привлечением учителя: к исходной коллекции добавляется априори инородный документ и среди различных метрик расстояния между распределениями выбирается такая, которая показывает различие между ядром коллекции и инородным элементом максимально «контрастно». На основе этой метрики, оптимальной для данной коллекции, может быть построена метрика аномальности.

В ходе предварительного анализа в качестве метрики расстояния между распределениями было выбрано расстояние Хеллингера, которое рассчитывается по формуле (1).

$$H(P,Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}\ , \qquad (1)$$

где $P = (p_1, ..., p_k)$ – нормированный вектор косинусных расстояний от всех документов коллекции до ее центра, $Q = (q_1, ..., q_k)$ – нормированный вектор косинусных расстояний от всех документов коллекции до инородного документа.

Для обеих коллекций был выбран набор образцов аномальности – инородных документов разного типа. Функция оценки аномальности устанавливалась в зависимости от алгоритма, однако общий принцип ее вычисления базировался на сравнении расстояний от нового элемента до двух точек: до центра коллекции и до одного из инородных документов. Центром коллекции считается документ, сумма расстояний от которого до всех остальных документов минимальна (по косинусной метрике).

Для каждого из рассматриваемых алгоритмов был проведен процесс подбора оптимальных параметров. В качестве критериев оптимизации выступали следующие величины: количество документов коллекции, оценка аномальности которых выше, чем у инородного документа; расстояние от центра коллекции до инородного документа. Таким образом, целью оптимизации было получение таких параметров, при которых инородные документы оказались бы элементами с наиболее высокой оценкой аномальности (с учетом условия однородности коллекции), и при этом картина распределения оценок аномальности была бы наиболее выразительная.

4. Эксперимент. Цель эксперимента состояла в проверке стратегии обнаружения аномалий, приведенной в разделе «Методы», на экспериментальных данных. Для достижения поставленной цели были выделены этапы:

1. Подготовка набора данных для проведения эксперимента.
2. Выбор метода построения векторной модели.
3. Выбор параметров построения словаря векторной модели, которая обеспечивает наибольшее расстояние от центра коллекции до инородного документа.
4. Подбор методов обнаружения новизны/выбросов для векторной модели документов, полученной с помощью преобразования TF-IDF.
5. Подбор методов обнаружения новизны/выбросов для векторной модели документов, полученной с помощью тематического моделирования.

Для проведения эксперимента было выбрано 2 коллекции юридически значимых документов ГОСТ:

- относящихся к тематике информационных технологий (далее - ГОСТ ИТ) – 1198 документов[1];
- относящихся к тематике железных дорог (далее - ГОСТ ЖД) – 458 документов[2].

В качестве образцов аномальности для вышеуказанных коллекций было выбрано 5 инородных документов (далее - ИД), которые являются аномальными для данных коллекций согласно экспертной оценке, список документов и фрагменты текстов (для примера) приведены в таблицах 2 и 3.

Таблица 2. Образцы аномальности - инородные документы[3]

Обозначение в эксперименте	**Категория**	**Название ИД**	**Длина документа**
ИД0	Научная статья	Цифровой двойник сортировочной горки	1505 слов
ИД1	Финансовая отчетность	Заключение по результатам обзорной проверки промежуточной финансовой информации ОАО «РЖД» и его дочерних компаний	14842 слова
ИД2	Статья в СМИ	Сколько строить и в чем возить: транспортное машиностроение на пространстве	1712 слов
ИД3	Шаблон договора	Договор с ОАО «РЖД» на оказание услуг, связанных с перевозкой грузов	4327 слов
ИД4	Художественная литература	Отрывок из романа Л. Н. Толстого «Война и Мир» (т.1, ч.1, гл. I-III)	4815 слов

[1] Документы из раздела 35. Информационные технологии. Машины конторские в Общероссийском классификаторе стандартов (ОКС)

[2] Документы из раздела 45. Железнодорожная техника в ОКС

[3] Все перечисленные документы размещены в открытом доступе в Интернете.

Таблица 3. Фрагменты инородных документов

Обозначение	**Название документа**
Фрагмент текста документа	
ИД0	Цифровой двойник сортировочной горки
В целях моделирования наиболее оптимальных режимов работы объекта автоматизации (в зависимости от конкретной ситуации и выбранного критерия оптимизации и анализа возможности оптимизации численности персонала, обслуживающего и управляющего объектом) ведется разработка цифровых двойников как элементов станционной инфраструктуры, так и сортировочной горки в целом – как наиболее сложной в плане автоматизации процессов и обеспечения безопасности части сортировочной станции. Создание цифровых двойников элементов железнодорожной инфраструктуры и подвижного состава является логичным развитием концепции цифровой станции и технологии промышленного интернета вещей. Цифровой двойник (digital twin) – это перенесенный в цифровую среду двойник физического устройства, процесса или системы. Цифровой двойник – это математическая модель высокого уровня адекватности, которая позволяет с большой точностью описывать поведение объекта во всех ситуациях, на всех этапах жизненного цикла, включая аварийные. Применение цифровых двойников позволяет быстро смоделировать развитие событий в зависимости от тех или иных факторов, определить потенциальные риски, найти наиболее эффективные режимы работы, выстроить шаги по обеспечению безопасности.	
ИД1	Заключение по результатам обзорной проверки промежуточной финансовой информации ОАО «РЖД» и его дочерних компаний
В состав сумм в таблице выше по состоянию на 30 июня 2020 г. включены суммы по договорам, заключенным со связанными сторонами – совместными предприятиями Компании и компаниями под контролем Российской Федерации, в размере 36 503 миллиона рублей и 68 328 миллионов рублей, соответственно. По состоянию на 30 июня 2020 г. обязательства по данным договорам со связанными сторонами составили 7 496 миллионов рублей и 36 424 миллиона рублей, соответственно. Договоры на приобретение локомотивов, заключенные в 2018–2019 годах, также предусматривают обязательства по обеспечению сервисного обслуживания в период жизненного цикла на срок до 28 лет, не включенные в суммы в таблице выше. Кроме того, в 2017–2020 годах Группа заключила договоры на выполнение работ по капитальному ремонту и модернизации подвижного состава, капитальному ремонту объектов электрификации и электроснабжения на общую сумму 170 950 миллионов рублей (в том числе 123 397 миллионов рублей по договорам, заключенным со связанными сторонами – совместными предприятиями Компании по состоянию на 30 июня 2020 г.). Стоимость работ, планируемых к выполнению по данным договорам после 30 июня 2020 г., составляет 105 873 миллиона рублей (в том числе 66 315 миллионов рублей по договорам, заключенным со связанными сторонами по состоянию на 30 июня 2020 г.).	

Продолжение таблицы 3.

Обозначение	**Название документа**
Фрагмент текста документа	
ИД2	Сколько строить и в чем возить: транспортное машиностроение на пространстве
Разбалансировка рынка Большинство железнодорожных компаний стран пространства являются крупнейшими системообразующими элементами экономики и ключевыми звеньями транспортных систем своих государств. В России 2010–2019 гг. были достаточно плодотворными с точки зрения появления современной продукции для транспортного машиностроения. Правда, в последнее время из-за перенасыщения рынка грузовых вагонов динамика производства в отечественном железнодорожном машиностроении значительно ухудшилась. Такая тенденция наблюдается уже не первый год, а сейчас на нее наложились еще и последствия пандемии коронавируса, из-за которых произошло резкое снижение грузовых перевозок и сокращение инвестиционных планов транспортных компаний. По данным РЖД, всего в этом году планируется модернизировать 449 вагонов. Кроме того, 2 октября на Павелецком вокзале Москвы состоялась презентация нового концепта плацкартного пассажирского вагона для поездов дальнего следования. На фоне пандемии вагоностроители столкнулись с серьезным падением спроса и прогнозами на его замораживание в среднесрочной перспективе. Заместитель генерального директора ИПЕМ Владимир Савчук поясняет, что снижение спроса на новые грузовые вагоны в России прогнозировалось заранее и максимальные риски среди заводов были у НПК «Уралвагонзавод» (УВЗ, входит в «Ростех») из-за концентрации линейки выпускаемого предприятием подвижного состава в наиболее рискованных сегментах рынка – полувагонах и нефтебензиновых цистернах.	
ИД3	Договор с ОАО «РЖД» на оказание услуг, связанных с перевозкой грузов
5. Ответственность Сторон 5.1. Стороны несут ответственность за неисполнение или ненадлежащее исполнение своих обязательств по настоящему Договору в соответствии с действующим законодательством Российской Федерации. 5.2. За просрочку платежей по настоящему Договору ОАО «РЖД» вправе предъявить Клиенту требования об уплате пени в размере 0,1 % от суммы задолженности за каждый день просрочки. 5.3. В случае отсутствия в ОАО «РЖД» уведомления об изменении местонахождения, почтового адреса, других реквизитов Клиента, учредительных, уставных документов, а также непредставления Клиентом заверенных надлежащим образом копий подтверждающих документов в сроки, установленные подпунктами 9.1 и 9.2 настоящего Договора, ОАО «РЖД» вправе приостановить выполнение своих обязательств по настоящему Договору до предоставления Клиентом указанных сведений и документов. 5.4. Исполнитель по настоящему Договору не является Грузоотправителем и не несет ответственность, предусмотренную законодательством РФ как Грузоотправитель.	

Продолжение таблицы 3.

Обозначение	**Название документа**
Фрагмент текста документа	
ИД4	Отрывок из романа Л. Н. Толстого «Война и Мир»
Гостиная Анны Павловны начала понемногу наполняться. Приехала высшая знать Петербурга, люди самые разнородные по возрастам и характерам, но одинаковые по обществу, в каком все жили; приехала дочь князя Василия, красавица Элен, заехавшая за отцом, чтобы с ним вместе ехать на праздник посланника. Она была в шифре и бальном платье. Приехала и известная, как la femme la plus séduisante de Pétersbourg,[34] молодая, маленькая княгиня Болконская, прошлую зиму вышедшая замуж и теперь не выезжавшая в большой свет по причине своей беременности, но ездившая еще на небольшие вечера. Приехал князь Ипполит, сын князя Василия, с Мортемаром, которого он представил; приехал и аббат Морио и многие другие. — Вы не видали еще? или: — вы не знакомы с ma tante? [35] — говорила Анна Павловна приезжавшим гостям и весьма серьезно подводила их к маленькой старушке в высоких бантах, выплывшей из другой комнаты, как скоро стали приезжать гости, называла их по имени, медленно переводя глаза с гостя на ma tante,[36] и потом отходила.	

Эксперимент по определению оптимальных характеристик словаря векторной модели проводился для обеих коллекций ГОСТов, в качестве ИД был выбран наиболее отличающийся от них документ ИД4 – сопоставимый по объему отрывок из романа «Война и Мир».

Для подбора оптимальной метрики аномальности были выбраны следующие варианты построения словаря векторной модели документов:

- субсловарные термы (*n*-граммы из 2–4 символов);
- субсловарные термы (*n*-граммы из 2–4 символов) с редуцированием словаря (РС): удалены русские стоп-слова, часто встречающиеся и редкие термы;
- униграммы и биграммы;
- униграммы и биграммы с РС.

Векторизация документов осуществлялась в двух вариантах: с помощью алгоритма расчета TF-IDF и на основе количества вхождений слова в документ. Векторное преобразование TF-IDF осуществлялось с нормализацией L1 (нормализация вектора на сумму абсолютных значений всех его компонентов). Для токенизации применялись два метода: TweetTokenizer (далее - TT) и RegexpTokenizer (далее - RT). Результаты расчета расстояния Хеллингера между центром коллекции и документом ИД4 представлены в таблице 4.

Таблица 4. Значения расстояния Хеллингера между центром коллекции и документом ИД4 (отрывок из «Войны и мира») при различных вариантах построения векторной модели

Словарь векторной модели	Коллекция: ГОСТ ИТ				Коллекция: ГОСТ ЖД			
	TF-IDF		Счетчик слов		TF-IDF		Счетчик слов	
	TT	RT	TT	RT	TT	RT	TT	RT
Субсловарный	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Субсловарный с РС	0.04	0.04	0.03	0.05	0.06	0.06	0.07	0.07
Униграммы и биграммы	0.03	0.03	0.05	0.05	0.04	0.04	0.09	0.1
Униграммы и биграммы с РС	0.03	0.03	0.05	0.05	0.04	0.04	0.05	0.05

Из таблицы видно, что наилучший результат (максимальное расстояние между распределениями) для обеих коллекций показала метрика аномальности на основе расстояния Хеллингера для субсловарных 2–4-грамм без редуцирования словаря по частотам. Дальнейший подбор методов выявления аномалий в текстовых коллекциях проводился при этих параметрах словаря.

В рамках подбора методов и параметров для обнаружения аномальных данных был проведен ряд экспериментов с применением следующих алгоритмов:

- локальный фактор выброса (LOF): алгоритм – поиск методом перебора, количество соседей – 2. Метод проверялся для разных значений метрики расстояния: cosine и L2;
- изоляционный лес (iForest): количество деревьев (эстиматоров) – 500;
- метод опорных векторов для одного класса (One-Class SVM): в качестве функции преобразования признакового пространства-ядра (kernel) рассматривались линейная функция (linear)/ сигмоида (sigmoid)/ радиальная базисная функция (RBF).

Использовалась реализация алгоритмов в Python-библиотеке Scikit-learn.

Указанные алгоритмы применялись для векторов документов, сформированных с помощью векторного преобразования TF-IDF для субсловарных 4-грамм без редуцирования словаря по частотам.

Для каждого из алгоритмов было исследовано влияние размерности векторов документов на результат работы. Проверялась работа алгоритма без снижения размерности, а также с применением следующих методов ее снижения:

- нелинейный метод снижения размерности UMAP: размерность конечного пространства – 300, метрика расстояния – косинусное расстояние;
- линейный метод на основе сингулярного разложения (Truncated SVD: размерность конечного пространства – 300). Метод тестировался при двух вариантах алгоритма разложения: итерационным методом Арнольди (SVD arpack) и ускоренным алгоритмом Халко (SVD random)).

Обнаружение аномалий проводилось каждым из алгоритмов поочередно в двух режимах, соответствующих виду задачи: поиск выбросов (обучение модели на коллекциях документов с добавленными инородными документами) и обнаружение новизны (обучение модели на коллекциях документов без добавленных инородных документов).

После получения функции принятия решения (decision-функции) по каждому алгоритму была произведена оценка качества ее работы: были определены документы коллекции, которые decision-функция отнесла к аномалиям с индексом выше, чем ИД2 (статья в СМИ). Результаты анализа представлены в таблице 5.

Из таблицы видно, что наилучший результат для обеих коллекций получен на полноразмерных векторах документов без использования методов снижения размерности для задачи обнаружения новизны. При таком варианте найдено наименьшее количество документов коллекции, отнесенных к аномальным с индексом выше, чем инородный документ. Также было обнаружено, что в данном варианте нет документов с индексом аномальности выше, чем у художественного текста (см. рис. 2). Наихудший результат получен при использовании нелинейного метода снижения размерности UMAP.

Наиболее стабильный результат для обеих коллекций вне зависимости от выбора метода снижения размерности показал алгоритм, основанный на методе опорных векторов (One-Class SVM). Работа метода была проверена на моделях обеих коллекций, построенных на векторах документов, сформированных с помощью преобразования TF-IDF для словаря из униграмм и биграмм с редуцированием по частотам (удалены часто встречающиеся и редкие термы) – такая векторная модель наиболее часто используется в ПИИС.

Эксперимент показал, что метод One-Class SVM с функцией преобразования признакового пространства RBF лучше определяет аномалии при нормализации векторной модели по L2, при которой вектор нормализуется на сумму квадратов всех его значений (см. табл. 6).

Также было обнаружено, что наилучшие результаты получены при детектировании инородных документов как новых для коллекций.

Таблица 5. Число документов, отнесенных к аномальным с индексом выше, чем у документа ИД2 (статья в СМИ), различными алгоритмами на основе полноразмерной векторной модели

Снижение размерности	Вид задачи	iForest	LOF		One-Class SVM		
			cosine	L2	sigmoid	linear	RBF
Коллекция: ГОСТ ИТ							
Без снижения	Выбросы	13	460	496	2	2	291
	Новизна	31	3	5	2	2	292
UMAP	Выбросы	1029	792	863	1102	593	818
	Новизна	1036	123	117	1102	593	818
SVD arpack	Выбросы	170	902	897	2	2	380
	Новизна	146	9	11	2	2	380
SVD random	Выбросы	156	881	864	2	2	384
	Новизна	182	7	10	2	2	384
Коллекция: ГОСТ ЖД							
Без снижения	Выбросы	12	20	36	5	5	202
	Новизна	7	2	1	5	5	200
UMAP	Выбросы	443	312	305	458	156	417
	Новизна	442	255	246	458	156	416
SVD arpack	Выбросы	112	20	41	5	5	194
	Новизна	151	2	3	5	5	192
SVD random	Выбросы	129	18	41	5	5	194
	Новизна	136	2	2	5	5	193

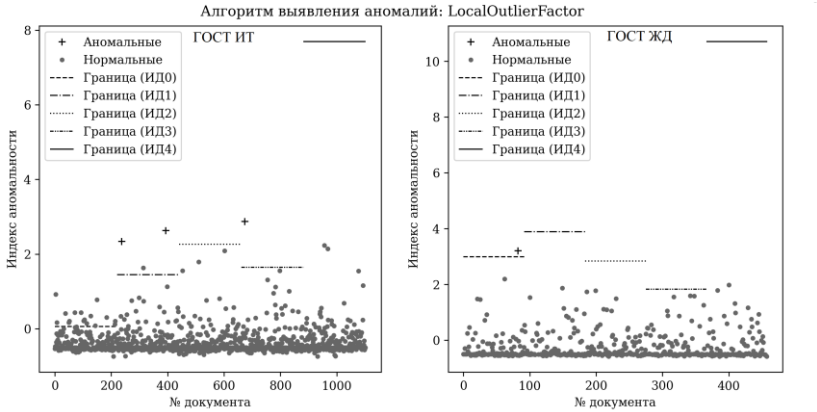

Рис. 2. Индекс аномальности документов в коллекциях ГОСТ ИТ и ГОСТ ЖД. Алгоритм Local Outlier Factor, векторная модель TF-IDF без понижения размерности

Таблица 6. Число документов, отнесенных к аномальным с индексом выше, чем у документа ИД2 (статья в СМИ), алгоритмом One-Class SVM на основе полноразмерной векторной модели

Снижение размерности	**Вид задачи**	**One-Class SVM (L1)**			**One-Class SVM (L2)**		
		sigmoid	**linear**	**RBF**	**sigmoid**	**linear**	**RBF**
Коллекция: ГОСТ ИТ							
Без снижения	Выбросы	4	110	51	8	8	6
	Новизна	0	0	51	3	2	0
SVD arpack	Выбросы	3	159	53	10	9	272
	Новизна	0	0	52	3	3	247
SVD random	Выбросы	0	136	53	10	9	264
	Новизна	2	0	53	5	4	242
Коллекция: ГОСТ ЖД							
Без снижения	Выбросы	125	386	5	3	3	1
	Новизна	0	0	4	0	0	0
SVD arpack	Выбросы	117	366	5	3	3	2
	Новизна	0	0	4	0	0	0
SVD random	Выбросы	118	364	4	3	3	2
	Новизна	0	0	5	0	0	0

Поскольку в ПИИС в зависимости от задачи могут использоваться разные способы векторного преобразования документов, то дальнейшие эксперименты проводились для векторной модели документов, полученной с помощью тематического моделирования.

Для обеих коллекций векторная модель документов строилась с методом тематического моделирования с аддитивной регуляризацией (Additive Regularization Topic Modelling, ARTM). При формировании словаря тематической модели была учтена информация о совстречаемости слов. Далее из 5 инородных документов была сформирована матрица «документ-тема» путем построения векторов на основе имеющихся тем.

Для получения decision-функции были проверены те же методы, что и для векторной модели текста TF-IDF: iForest, LOF, One-Class SVM. Для оценки качества работы функции также были определены документы коллекции, которые были отнесены к аномалиям с индексом выше, чем ИД2 (статья в СМИ). Результаты анализа представлены в таблице 7.

Таблица 7. Число документов, отнесенных к аномальным с индексом, превышающим индекс аномальности документа ИД2 (статья в СМИ), различными алгоритмами на основе тематической модели коллекции

Вид задачи	**iForest**		**LOF**		**One-Class SVM**	
	estimators: 400	**estimators: 500**	**cosine**	**L2**	**sigmoid**	**linear**
Коллекция: ГОСТ ИТ						
Выбросы	0	2	64	68	191	80
Новизна	0	0	6	97	191	77
Коллекция: ГОСТ ЖД						
Выбросы	0	0	339	401	305	27
Новизна	0	0	18	425	98	17

Из таблицы видно, что определение аномалий для тематической модели структурированных нормативно-технических документов на русском языке наилучшим образом произведено с помощью алгоритма изолирующего леса (Isolation Forest), при использовании этого метода все 5 ИД идентифицируются как аномальные: практически нет документов, превышающих ИД по индексу аномальности (см. рис. 3).

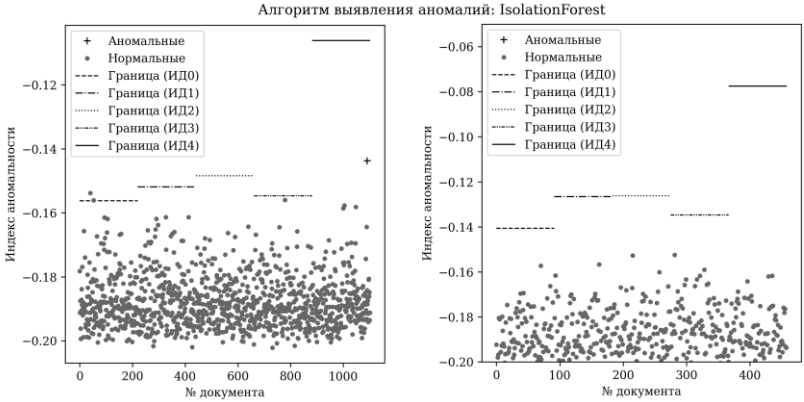

Рис. 3. Индекс аномальности для коллекций ГОСТ ИТ и ГОСТ ЖД. Алгоритм Isolation Forest, тематическая векторная модель (ARTM)

При использовании тематических векторов коллекций инородные документы изолируются алгоритмом гораздо раньше, чем нормальные объекты (выбросы попадают в листья на небольшой глубине дерева).

5. Результаты. Результаты эксперимента продемонстрировали эффективность предложенного подхода. При поиске аномалий в однородных текстовых коллекциях подбор параметров оптимизационным способом обеспечивает максимальную точность решения для конкретного набора данных с учетом предоставленного алгоритму образца инородного документа.

При этом выбор алгоритма зависит от целей и задач ПИИС: если для документов предполагается строить только векторную модель по TF-IDF, то выбор метода определяется большой размерностью пространства признаков. Как показал эксперимент, в этом случае снижение размерности нецелесообразно: наиболее эффективным оказались методы One-Class SVM (на основе метода опорных векторов) и LOF (локальный фактор выброса), работающие с полноразмерными векторами в режиме обнаружения новизны. Наиболее стабильный результат для обеих тестовых коллекций вне зависимости от выбора метода снижения размерности показал алгоритм One-Class SVM.

Эксперимент также показал, что наилучший результат можно получить при формировании словаря из субсловарных единиц в виде 2–4-грамм. Однако работа метода была также проверена при векторизации преобразованием TF-IDF по словарю из униграмм и биграмм с редуцированием по частотам - такая векторная модель наиболее часто используется в ПИИС. Экспериментально была установлена зависимость между выбором ядра и способом нормализации: линейное и сигмоидное ядра в One-Class SVM лучше работают при нормализации L1, тогда как преобразование пространства RBF эффективно при нормализации L2.

В случае, если в рамках ПИИС строится тематическая модель, аномалии эффективно определяются методом iForest (изолирующий лес), причем результат оказывается даже лучше, чем у оптимального алгоритма при векторизации методом TF-IDF.

Тематическое моделирование может быть рассмотрено как метод снижения размерности пространства: вместо векторов размерностью десятки тысяч компонентов (в соответствии со словарем коллекции) пространство формируется из векторов документов размерностью в несколько сотен тематических компонентов (что соответствует общепринятому порядку оптимального числа тем в тематической модели). Другие методы снижения размерности приводят к потере вместе с избыточной размерностью и некоторой части значимой информации - в частности, семантических признаков текстовых документов. Тематическое же моделирование позволяет сохранить эту информацию, поскольку уменьшение размерности модели происходит не за счет уплотнения векторов, а благодаря разложению исходной матрицы «документ-

терм» на две матрица меньшего размера - «терм-тема» и «тема-документ».

Тематическая модель также позволяет обнаружить аномальные элементы в процессе кластеризации. Аномалиями могут считаться документы, которые, несмотря на регуляризацию, «плохо» раскладываются по полученной модели: для них не обнаруживается подходящей темы с достаточным весом, вместо этого вектор такого документа содержит низкие значения весов по многим темам коллекции. Такие документы могут рассматриваться как выбросы. Однако при таком подходе возникает необходимость эмпирического выбора величины порога, которая в общем случае зависит от свойств коллекции. Пороговое значение необходимо выбирать таким образом, чтобы соблюсти баланс между числом кластеризуемых документов и числом документов, определяемых как аномальные. Кроме того, на практике этот подход невозможен для обнаружения новизны, поскольку полное переобучение тематической модели при поступлении каждого нового документа в ПИИС нецелесообразно.

Если же векторизовать новый документ по имеющемуся набору тем, то для обнаружения аномалий необходимо формализовать алгоритм принятия решения относительно качества разложения конкретного документа по имеющейся тематической модели. Это и делают алгоритмы обнаружения аномалий, проверенные экспериментально в данной работе. Примечательно, что стратегия поиска аномалий на основе тематической модели ARTM оказалось наиболее эффективной, при применении методов снижения размерности к модели TF-IDF таких результатов добиться не удавалось.

Таким образом, эксперимент подтвердил, что задача выявления единичных аномальных документов по отношению к заданной коллекции не имеет универсального решения. Однако возможно применение различных стратегий выявления аномалий в зависимости от потребности ПИИС.

Предложенный оптимизационный подход показал высокую эффективность при поиске смысловых аномалий на исследуемых данных. При этом необходимо отметить существующие ограничения текущего исследования. Для достижения высокой точности при поиске аномалий необходимо использовать знания о типе и структуре данных и об их предметной области. Это затрудняет оценку и сравнение с результатами других исследований: насколько нам известно, в настоящее время нет опубликованных исследований, посвященных поиску аномалий на уровне документов в коллекциях юридически значимых русскоязычных текстов. Предложенный подход оптимизации алгоритма обнаружения

аномалий с использованием заведомо инородного документа является новым, что также является причиной невозможности сопоставления полученного результата с предыдущими работами. Опубликованные исследования, посвященные обнаружению смысловой аномальности (новизны) на уровне документов, демонстрируют, что эффективность предлагаемых методов сильно зависит от обрабатываемого корпуса документов. В частности, в работе [2] приводятся оценки точности предложенного нейросетевого метода определения семантической новизны на уровне 0,75, 0,76 и 0,86 в зависимости от набора данных. Оптимизационный подход, представленный в данной статье, позволяет добиться абсолютной точности (близкой к 100%) при достаточном уровне однородности коллекции. Такая разница в эффективности демонстрирует зависимость оценок методов от качества тестовых данных. Это позволяет утверждать, что в прикладных задачах этап подготовки данных и выбор самого алгоритма поиска аномалий могут оказывать сопоставимо сильное влияние на результат анализа. В то же время при определении аномальности на семантическом уровне важно разграничивать случаи, когда ошибка вызвана работой системы (алгоритма), и ситуации ошибки из-за недостаточной формализации качества «аномальности» (когда результат может быть субъективно воспринят как ошибочный). Предложенный подход позволяет объективизировать смысловую аномальность текстовых документов по отношению к заданной коллекции, благодаря чему снижается число ошибок и повышается точность решения.

На двух исследуемых коллекциях метод показал высокую эффективность, которая, однако, должна быть проверена в реальном применении в рамках прикладной информационной системы. Изучение откликов пользователей на предложенные системой оценки аномальности документов позволит подтвердить на практике точность представленного метода и его применимость на различных наборах данных. Составление соответствующих метрик для встраивания в систему является одним из возможных направлений дальнейших исследований.

Кроме того, для прикладного применения предложенного алгоритма целесообразно автоматизировать процесс определения порогового значения аномальности, который устанавливается через выбор документа, заведомо инородного для коллекции. Таким документом может являться внешний документ, который не соответствует коллекции тематически и стилистически. В рамках данной работы подбор такого инородного документа происходил вручную, на основе экспертной оценки. Без привлечения эксперта пороговое значение аномальности может устанавливаться на основе максимального значения скоринга

аномальности документов коллекции, с некоторым сдвигом этого значения, вычисленного при начальных параметрах алгоритма. Эта гипотеза также требует дальнейшего изучения и экспериментальной проверки.

Общее направление дальнейших исследований определяется движением в сторону т.н. машиночитаемого права и включает в себя решение задач, связанных с обнаружением смысловых аномалий в юридически значимых текстах: выявление новшеств в проектах российских нормативно-правовых документов и в локальных нормативных актах, формальное представление семантического содержания русскоязычных регулятивных документов [39], поиск отдельных аномальных структур в юридических текстах, обнаружение нормотворческих коллизий.

6. Заключение. В работе исследовалась задача обнаружения аномалий как документов, обладающих новизной по отношению к однородной текстовой коллекции, с учетом требований, который накладываются на выбор методов решения задачи прикладными интеллектуальными информационными системами (ПИИС), работающими с юридически значимыми документами. К этим требованиям относится высокая точность решения, реализуемая эффективность алгоритмов, воспроизводимость результата и его объяснимость.

Предложена стратегия выбора оптимального метода поиска аномалий и подбора его параметров в зависимости от предполагаемой в ПИИС векторной модели коллекции документов.

К коллекции добавляется заведомо инородный документ, с учетом которого определяются критерии оптимизации: максимальное различие распределений расстояний между документами коллекции до ее центра и до инородного документа; отсутствие (либо минимальное число) документов исходной коллекции, которые по значению индекса аномальности превосходят инородный документ. Точкой отсчета для индекса аномальности считается центр рассматриваемой коллекции.

Для точечного выявления аномальности необходимо определить в коллекции пограничный документ, имеющий максимальную величину индекса аномальности среди всех документов коллекции (функция определения аномальности выбирается в зависимости от векторной модели документов). В рамках ПИИС аномальность нового документа определяется в рекомендательном режиме, при этом может быть установлен допустимый лимит превышения индекса аномальности нового документа относительно пограничного документа.

Эксперимент проводился на двух однородных коллекциях юридически значимых русскоязычных документах: государственные стан-

дарты в сфере информационных технологий и в сфере железнодорожного сообщения. В качестве образцов инородных документов использовались отрывок из художественного произведения, статья в СМИ, научная статья, финансовая отчетность и шаблон договора.

Результаты эксперимента показали, что предложенный метод подбора алгоритмов и параметров по минимизации выбросов относительно заведомо инородного документа позволяет выбрать оптимальный алгоритм обнаружения аномалий для каждой текстовой коллекции, а полученная таким образом оценка аномальности документа соответствует экспертной оценке.

Для задач поиска аномалий в рамках процесса тематической кластеризации юридически значимых текстовых документов эффективен метод изолирующего леса (Isolation Forest), поскольку инородные документы в силу отсутствия явно выделенной темы изолируются быстрее, чем документы, принадлежащие семантически однородному кластеру.

Если для решения прикладной задачи используется разреженная полноразмерная векторная модель на основе словарной статистики, то после выбора оптимальных параметров словаря модели для ПИИС рекомендуется применить метод опорных векторов в модификации One-Class SVM с соответствующей функцией преобразования признакового пространства. При снижении размерности наилучший результат показывают линейные методы.

В ходе дальнейших исследований планируется протестировать предложенный метод на практике в рамках действующей прикладной информационной системы и оценить применимость метода через встроенные метрики, учитывающие отклики пользователей на предложенные системой оценки аномальности документов. Также целесообразно автоматизировать процесс определения порогового значения аномальности, используя данные о скоринге аномальности документов исходной коллекции. Другим направлением исследований станет доработка алгоритма поиска аномалий методом тематической кластеризации. В частности, требует проверки предположение, что использование словосочетаний вместо n-грамм при построении словаря модели улучшит определение аномальных тем.

В рамках дальнейших исследований решение актуальной прикладной задачи выявления смысловых аномалий в юридически значимых, в том числе регулятивных, документах позволит сделать шаг в сторону интеллектуализации нормотворческой деятельности и машиночитаемого права.

Литература

1. *Mahapatra A., Srivastava N., Srivastava J.* Contextual anomaly detection in text data // Algorithms. 2012. vol. 5. no. 4. pp. 469-489.
2. *Ghosal T. et al.* Novelty goes deep. A deep neural solution to document level novelty detection // Proceedings of the 27th International Conference on Computational Linguistics, 2018. pp. 2802–2813.
3. *Zhao L., Zhang M., Ma S.* The nature of novelty detection // Information Retrieval. 2006. vol. 9. no. 5. C. 521–541.
4. *Guzman J., Poblete B.* On-line relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model // Proceedings of the ACM SIGKDD workshop on outlier detection and description. 2013. pp. 31-39.
5. *Lau J. H. et al.* Word sense induction for novel sense detection // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012. pp. 591-601.
6. *Гурина А.О., Гузев О.Ю., Елисеев В.Л.* Обнаружение аномальных событий на хосте с использованием автокодировщика // International Journal of Open Information Technologies. 2020. Т. 8. №. 8.
7. *Goldstein M., Dengel A.* Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm // KI-2012: Poster and Demo Track. 2012. pp. 59-63.
8. *Zhao Y., Nasrullah Z., Li Z.* Pyod: A python toolbox for scalable outlier detection // arXiv preprint arXiv:1901.01588. 2019.
9. *Denning D.E.* An intrusion-detection model // IEEE Transactions on software engineering. 1987. no. 2. pp. 222-232.
10. *Markou M., Singh S.* Novelty detection: a review—part 1: statistical approaches // Signal processing. 2003. vol. 83. no. 12. pp. 2481-2497.
11. *Chandola V., Banerjee A., Kumar V.* Anomaly detection: A survey // ACM computing surveys (CSUR). 2009. vol. 41. no. 3. pp. 1-58.
12. *Pimentel M.A.F. et al.* A review of novelty detection // Signal Processing. 2014. vol. 99. pp. 215-249.
13. *Faria E.R. et al.* Novelty detection in data streams // Artificial Intelligence Review. 2016. vol. 45. no. 2. pp. 235-269.
14. *Ruff L. et al.* A unifying review of deep and shallow anomaly detection // Proceedings of the IEEE. 2021.
15. *Hendrycks D., Mazeika M., Dietterich T.* Deep anomaly detection with outlier exposure // arXiv preprint arXiv:1812.04606. 2018.
16. *Gorokhov O., Petrovskiy M., Mashechkin I.* Convolutional neural networks for unsupervised anomaly detection in text data // International Conference on Intelligent Data Engineering and Automated Learning. Springer, Cham, 2017. pp. 500-507.
17. *Yang Y. et al.* Topic-conditioned novelty detection // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002. pp. 688-693.
18. *Ng K.W. et al.* Novelty detection for text documents using named entity recognition // 2007 6th international conference on information, communications & signal processing. IEEE, 2007. pp. 1-5.
19. *Amplayo R.K., Hong S.L., Song M.* Network-based approach to detect novelty of scholarly literature // Information Sciences. 2018. vol. 422. pp. 542-557.

20. *Li Z. et al.* COPOD: copula-based outlier detection // arXiv preprint arXiv:2009.09463. 2020.
21. *Mikolov T., Yih W., Zweig G.* Linguistic regularities in continuous space word representations // Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013. pp. 746-751.
22. *Краснов Ф.В., Смазневич И.С.* Фактор объяснимости алгоритма в задачах поиска схожести текстовых документов // Вычислительные технологии. 2020. Т. 25. №. 5. С. 107-123.
23. *Schubert E., Gertz M.* Intrinsic t-stochastic neighbor embedding for visualization and outlier detection // International Conference on Similarity Search and Applications. Springer, Cham, 2017. pp. 188-203.
24. McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction // arXiv preprint arXiv:1802.03426. 2018.
25. *Narayan A., Berger B., Cho H.* Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability // bioRxiv. 2020.
26. *Campos G.O. et al.* On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study // Data mining and knowledge discovery. 2016. vol. 30. №. 4. pp. 891-927.
27. *Amarbayasgalan T., Jargalsaikhan B., Ryu K.H.* Unsupervised novelty detection using deep autoencoders with density-based clustering // Applied Sciences. 2018. vol. 8. no. 9. pp. 1468.
28. *Campello R.J.G.B. et al.* Hierarchical density estimates for data clustering, visualization, and outlier detection // ACM Transactions on Knowledge Discovery from Data (TKDD). 2015. vol. 10. no. 1. pp. 1-51.
29. *Ankerst M. et al.* OPTICS: Ordering points to identify the clustering structure // ACM Sigmod record. 1999. vol. 28. no. 2. pp. 49-60.
30. *Karypis G., Han E.H., Kumar V.* Chameleon: Hierarchical clustering using dynamic modeling // Computer. 1999. vol. 32. no. 8. pp. 68-75.
31. *Karypis G., Kumar V.* A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices // University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center, Minneapolis, MN. 1998. vol. 38.
32. *Kannan R. et al.* Outlier detection for text data // Proceedings of the 2017 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2017. pp. 489-497.
33. *Zhang J., Ghahramani Z., Yang Y.* A probabilistic model for online document clustering with application to novelty detection // Advances in neural information processing systems. 2004. vol. 17. pp. 1617-1624.
34. *Manevitz L. M., Yousef M.* One-class SVMs for document classification // Journal of machine Learning research. 2001. vol. 2. no. Dec. pp. 139-154.
35. *Zimek A., Campello R.J.G.B., Sander J.* Ensembles for unsupervised outlier detection: challenges and research questions a position paper // ACM SIGKDD Explorations Newsletter. 2014. vol. 15. no. 1. pp. 11-22.
36. *Marques H.O. et al.* Internal evaluation of unsupervised outlier detection // ACM Transactions on Knowledge Discovery from Data (TKDD). 2020. vol. 14. no. 4. pp. 1-42.
37. *Liu F.T., Ting K.M., Zhou Z.H.* Isolation Forest // 2008 Eighth IEEE international conference on data mining. IEEE, 2008. pp. 413-422.

38. *Краснов Ф.В.* Сравнительный анализ точности методов визуализации структуры коллекции текстов // International Journal of Open Information Technologies. 2021. Т. 9. №. 4. С. 79-84.
39. *Пименов В.И., Воронов М.В.* Формализация регулятивных текстов // Информатика и автоматизация. 2021. № 3 (20). С. 562–590.

Краснов Федор Владимирович — кандидат технических наук, эксперт, департамент семантических систем, NAUMEN R&D. Область научных интересов: интеллектуальная аналитика текстов. Число научных публикаций – 69. fkrasnov@naumen.ru, www.naumen.ru; 620028, Екатеринбург, ул. Татищева, 49А; р.т.: +7 (981)781-48-47.

Смазневич Ирина Сергеевна — бизнес-аналитик, департамент семантических систем, NAUMEN R&D. Область научных интересов: применение интеллектуальных алгоритмов в прикладных информационных системах. Число научных публикаций – 2. ismaznevich@naumen.ru, www.naumen.ru; 620028, Екатеринбург, ул. Татищева, 49А; р.т.:+7(916)722-66-39.

Баскакова Елена Николаевна — ведущий системный аналитик, департамент семантических систем, NAUMEN R&D. Область научных интересов: применение интеллектуальных алгоритмов в прикладных информационных системах. enbaskakova@naumen.ru, www.naumen.ru; 620028, Екатеринбург, ул. Татищева, 49А; р.т.: +7(903)258-75-93.

DOI 10.15622/ia.20.4.5

F. KRASNOV, I. SMAZVEVICH, E. BASKAKOVA

OPTIMIZATION APPROACH TO SELECTING METHODS OF DETECTING ANOMALIES IN HOMOGENEOUS TEXT COLLECTIONS

Krasnov F., Smaznevich I., Baskakova E. **Optimization Approach to Selecting Methods of Detecting Anomalies in Homogeneous Text Collections.**

Abstract. The problem of detecting anomalous documents in text collections is considered. The existing methods for detecting anomalies are not universal and do not show a stable result on different data sets. The accuracy of the results depends on the choice of parameters at each step of the problem solving algorithm process, and for different collections different sets of parameters are optimal. Not all of the existing algorithms for detecting anomalies work effectively with text data, which vector representation is characterized by high dimensionality with strong sparsity.

The problem of finding anomalies is considered in the following statement: it is necessary to checking a new document uploaded to an applied intelligent information system for congruence with a homogeneous collection of documents stored in it. In such systems that process legal documents the following limitations are imposed on the anomaly detection methods: high accuracy, computational efficiency, reproducibility of results and explicability of the solution. Methods satisfying these conditions are investigated.

The paper examines the possibility of evaluating text documents on the scale of anomaly by deliberately introducing a foreign document into the collection. A strategy for detecting novelty of the document in relation to the collection is proposed, which assumes a reasonable selection of methods and parameters. It is shown how the accuracy of the solution is affected by the choice of vectorization options, tokenization principles, dimensionality reduction methods and parameters of novelty detection algorithms.

The experiment was conducted on two homogeneous collections of documents containing technical norms: standards in the field of information technology and railways. The following approaches were used: calculation of the anomaly index as the Hellinger distance between the distributions of the remoteness of documents to the center of the collection and to the foreign document; optimization of the novelty detection algorithms depending on the methods of vectorization and dimensionality reduction. The vector space was constructed using the TF-IDF transformation and ARTM topic modeling. The following algorithms have been tested: Isolation Forest, Local Outlier Factor and One-Class SVM (based on Support Vector Machine).

The experiment confirmed the effectiveness of the proposed optimization strategy for determining the appropriate method for detecting anomalies for a given text collection. When searching for an anomaly in the context of topic clustering of legal documents, the Isolating Forest method is proved to be effective. When vectorizing documents using TF-IDF, it is advisable to choose the optimal dictionary parameters and use the One-Class SVM method with the corresponding feature space transformation function.

Keywords: Anomaly Detection, Novelty Detection, Outlier Detection, Homogeneous Text Collections, Sparse Space Dimension Reduction, Topic Modeling.

Krasnov Fedor — Ph.D., Expert, Department of Semantic Systems, NAUMEN R&D. Research interests: Intelligent text analysis. The number of publications – 69; e-mail: fkrasnov@naumen.ru, www.naumen.ru; 49A, Tatishcheva street, «Tatishchevsky» Business Center, 4th floor, Yekaterinburg, 620028, Russian Federation; office phone: +7 (981)781-48-47.

Smaznevich Irina — Business analyst, Department of Semantic Systems, NAUMEN R&D, Research interests: Use of intelligent algorithms in applied information systems. The number of publications – 2; e-mail: ismaznevich@naumen.ru, www.naumen.ru; 49A, Tatishcheva street, «Tatishchevsky» Business Center, 4th floor, Yekaterinburg, 620028, Russian Federation; office phone: +7(916)722-66-39.

Baskakova Elena — System analyst, Department of Semantic Systems, NAUMEN R&D. Research interests: Use of intelligent algorithms in applied information systems; e-mail: enbaskakova@naumen.ru, www.naumen.ru; 49A, Tatishcheva street, «Tatishchevsky» Business Center, 4th floor, Yekaterinburg, 620028, Russian Federation; office phone: +7(903)258-75-93.

References

1. Mahapatra A., Srivastava N., Srivastava J. Contextual anomaly detection in text data. *Algorithms*. 2012. vol. 5. no. 4. pp. 469-489.
2. Ghosal T. et al. Novelty goes deep. A deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018, pp. 2802–2813.
3. Zhao L., Zhang M., Ma S. The nature of novelty detection. *Information Retrieval*. 2006. vol. 9. no. 5. C. 521–541.
4. Guzman J., Poblete B. Online relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model. *Proceedings of the ACM SIGKDD workshop on outlier detection and description*. 2013. pp. 31-39.
5. Lau J. H. et al. Word sense induction for novel sense detection. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012. pp. 591-601.
6. Gurina A.O., Guzev O.Ju., Eliseev V.L. [Detection of anomalous events on the host using an autoencoder] Obnaruzhenie anomal'nyh sobytij na hoste s ispol'zovaniem avtokodirovshhika. *International Journal of Open Information Technologies*. 2020. vol. 8. no. 8.
7. Goldstein M., Dengel A. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*. 2012. pp. 59-63.
8. Zhao Y., Nasrullah Z., Li Z. Pyod: A python toolbox for scalable outlier detection. *arXiv preprint.* arXiv:1901.01588. 2019.
9. Denning D.E. An intrusion-detection model. *IEEE Transactions on software engineering.* 1987. no. 2. pp. 222-232.
10. Markou M., Singh S. Novelty detection: a review—part 1: statistical approaches. *Signal processing*. 2003. vol. 83. no. 12. pp. 2481-2497.
11. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. *ACM computing surveys (CSUR).* 2009. vol. 41. no. 3. pp. 1-58.
12. Pimentel M.A.F. et al. A review of novelty detection. *Signal Processing*. 2014. vol. 99. pp. 215-249.
13. Faria E.R. et al. Novelty detection in data streams. *Artificial Intelligence Review*. 2016. vol. 45. no. 2. pp. 235-269.
14. Ruff L. et al. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE.* 2021.
15. Hendrycks D., Mazeika M., Dietterich T. Deep anomaly detection with outlier exposure. *arXiv preprint.* arXiv:1812.04606. 2018.
16. Gorokhov O., Petrovskiy M., Mashechkin I. Convolutional neural networks for unsupervised anomaly detection in text data. *International Conference on Intelligent Data Engineering and Automated Learning.* Springer, Cham, 2017. pp. 500-507.
17. Yang Y. et al. Topic-conditioned novelty detection. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002. pp. 688-693.
18. Ng K.W. et al. Novelty detection for text documents using named entity recognition. *2007 6th international conference on information, communications and signal processing*. IEEE, 2007. pp. 1-5.

19. Amplayo R.K., Hong S.L., Song M. Network-based approach to detect novelty of scholarly literature. *Information Sciences*. 2018. vol. 422. pp. 542-557.
20. Li Z. et al. COPOD: copula-based outlier detection. *arXiv preprint.* arXiv:2009.09463. 2020.
21. Mikolov T., Yih W., Zweig G. *L*inguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013. pp. 746-751.
22. Krasnov F.V., Smaznevich I.S. [The explicability factor of the algorithm in the problems of searching for the similarity of text documents]. *Vychislitel'nye tehnologii*. [Computational technologies]. 2020. vol. 25. no. 5. pp. 107-123.
23. Schubert E., Gertz M. *I*ntrinsic t-stochastic neighbor embedding for visualization and outlier detection. *International Conference on Similarity Search and Applications*. Springer, Cham, 2017. pp. 188-203.
24. McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint.* arXiv:1802.03426. 2018
25. Narayan A., Berger B., Cho H. Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *bioRxiv*. 2020.
26. Campos G.O. et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*. 2016. vol. 30. №. 4. pp. 891-927.
27. Amarbayasgalan T., Jargalsaikhan B., Ryu K.H. Unsupervised novelty detection using deep autoencoders with density-based clustering. *Applied Sciences*. 2018. vol. 8. no. 9. P. 1468.
28. Campello R.J.G.B. et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD).* 2015. vol. 10. no. 1. pp. 1-51.
29. Ankerst M. et al. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record*. 1999. vol. 28. no. 2. pp. 49-60.
30. Karypis G., Han E. H., Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*. 1999. vol. 32. no. 8. pp. 68-75.
31. Karypis G., Kumar V. A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center, Minneapolis, MN. 1998. vol. 38.
32. Kannan R. et al. Outlier detection for text data. *Proceedings of the 2017 siam international conference on data mining. Society for Industrial and Applied Mathematics*, 2017. pp. 489-497.
33. Zhang J., Ghahramani Z., Yang Y. A probabilistic model for online document clustering with application to novelty detection. *Advances in neural information processing systems*. 2004. vol. 17. pp. 1617-1624.
34. Manevitz L. M., Yousef M. One-class SVMs for document classification. *Journal of machine Learning research*. 2001. vol. 2. no. Dec. pp. 139-154.
35. Zimek A., Campello R.J.G.B., Sander J. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *ACM SIGKDD Explorations Newsletter*. 2014. vol. 15. no. 1. pp. 11-22.
36. Marques H. O. et al. Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2020. vol. 14. no. 4. pp. 1-42.
37. Liu F.T., Ting K.M., Zhou Z.H., Isolation Forest. *2008 Eighth IEEE international conference on data mining.* IEEE, 2008. pp. 413-422.
38. Krasnov F.V. [Comparative Analysis of the Accuracy of Methods for Visualizing the Structure of a Text Collection]. *International Journal of Open Information Technologies.* 2021. vol. 9. no. 4. pp. 79-84. (In Russ.)
39. Pimenov V.I., Voronov M.V. [Formalization of regulatory texts]. *Informatika i avtomatizacija.* [Computer Science and Automation]. 2021. no. 3(20). pp. 562–590. (In Russ.)