МАШИННЫЙ МОНИТОРИНГ ТЕКСТОВЫХ ЧАТОВ И ПРЕДСКАЗАНИЕ АНОМАЛИЙ

Мозаидзе Е. С.[1]
(*ФГБОУ ВО Белгородский государственный технологический университет им. В.Г. Шухова, Белгород*)
Зуев С. В.[2]
(*ФГАОУ ВО Крымский федеральный университет им. В.И. Вернадского, Симферополь*)

Целью работы является разработка нового метода предсказания аномалий в текстовых чатах, не использующего корпусы текстов. Поставленные задачи: краткое представление статистического описания повторяемости аномалий, развитого в прошлых работах авторов, введение метода парных (обобщенных) N-грамм на коллекциях «существительное – глагол», синтез указанных методов в новый метод предсказания аномалий в системах обмена короткими сообщениями, тестирование метода. Предложен новый метод предсказания аномалий в потоке текстовых сообщений, не использующий корпус текстов для обучения, и, кроме того, допускающий онлайн-обучение. Материалом для работы были чаты, группы и каналы в Telegram, на которые подписан один из авторов работы, с большим объемом текстового материала. Метод использует статистическое распределение повторения аномалий, а также метод тематического моделирования на основе статистики пар «существительное – глагол». Оба метода предложены ранее в работах авторов. Проведенный эксперимент показал соответствие результатов, предсказанных с помощью предлагаемого метода, фактически зарегистрированным аномалиям. Применение предложенного метода может быть полезно в исследованиях и анализе появления аномалий в сложных социальных системах, взаимодействие в которых отражается в коммуникациях через социальные сети и мессенджеры. Подобного рода задачи являются актуальными как для государственных структур, так и для бизнеса, и могут позволить сгладить острые социальные и производственные проблемы. Особенно полезен предложенный метод для журналистов – он позволяет определить время наиболее вероятного появления значимых социальных явлений.

Ключевые слова: предсказание аномалий, тематическое моделирование, вероятности редких событий, повторяемость редких событий, аномалии в текстовых чатах.

---

[1] *Елена Сергеевна Мозаидзе, аспирантка* (*mozaidze95@mail.ru*).
[2] *Сергей Валентинович Зуев, к.ф.-м.н., доцент* (*sergey.zuev@bk.ru*).

1. Введение

Мониторинг текущей активности пользователей в чатах с помощью ботов широко распространен и применяется во всех мессенджерах. В основном это делается для модерации общения и решения задач, подобных следующим:

– проверка действий и активности пользователей;
– проверка длины комментариев;
– бан нарушителей;
– запрет отправлять в чат картинки, стикеры, голосовые сообщения;
– возврат исправившихся нарушителей;
– борьба со спамом;
– отправление приветственного сообщения новым участникам;
– выявление замаскированных нарушителей.

Однако, в последнее время возникла потребность в решении задачи выявления и предсказания аномального поведения пользователя в чате. Работы, посвященные этому вопросу, имеются как в российской [1, 3, 5] так и в зарубежной [8, 9] литературе.

Понимание аномального поведения варьируется в зависимости от конкретной задачи. Поэтому общая постановка вопроса о детектировании аномалии пока представляется в весьма абстрактном виде: выявление аномалий – это распознавание редких данных, событий или наблюдений, которые вызывают подозрения ввиду существенного отличия от большей части данных [7, 14]. В такой интерпретации аномалий ключевым является редкость событий, понимание которой основано на наличии информации о повторяемости таких и других событий: повторяемость редких событий должна характеризоваться большими промежутками времени. В то же время выделяемые события должны иметь существенное семантическое содержание, так как иначе можно, например, установить событием появление в тексте большой буквы «Ы»: это будет действительно редкое событие, но не аномалия, так как под ним не содержится какого-либо смысла.

Обычно для обнаружения аномалий используются такие алгоритмы, как K-Means, KNN, OPTICS [10, 11, 12]. Эти методы машинного обучения могут решать обозначенные выше задачи вы-

явления аномалий поведения и активно используются для различных направлений деятельности. Но в исходном виде все они не подходят для работы с непрерывными потоками данных высокой интенсивности в реальном времени. Особенно ярко это проявляется при работе с данными, имеющими значительную изменчивость. Стандартный подход для использования алгоритмов управляемого обучения (Supervised Learning) требует сначала обучить модель на заранее подготовленной обучающей выборке, затем протестировать корректность предсказаний на тестовой выборке и, наконец, интерполировать результаты на более широкое множество реальных данных [17, 18].

При работе с временными рядами такой подход можно применять с серьезными ограничениями, так как:

– данные временных рядов жестко упорядочены;
– эти данные непрерывно меняются во времени.

Поэтому использовать модели, ранее обученные на старых данных, не разумно.

Можно пытаться переучивать модель через определенные интервалы времени. Однако такое решение малоэффективно из-за того, что обрабатываемых данных, как правило, очень много и одновременно обычно развернуто сразу несколько экземпляров моделей для работы с разными временными рядами. Поэтому процесс частого изменения модели связан с высокими дополнительными затратами и необходимостью контролировать процесс очередного обучения [13]. При высокой волатильности данных можно применять только такие алгоритмы, которые переобучаются быстрее, чем поступает очередная порция данных обучающей выборки из потока. Класс подобных моделей принято относить к области онлайн-обучения (Online Learning) [6, 16, 19]. Для скоростной высокоинтенсивной работы удобно использовать фреймворки, ориентированные на распределенные вычисления при работе с большими данными. Одним из наиболее популярных фреймворков является Apache Spark, предназначенный для распределенной обработки структурированных и неструктурированных данных [4].

Общее понимание аномалии как явления в системах исследовалось, в частности, в работах Н. Талеба [14], где в основном рассматривались экономические системы.

В индустрии средств защиты информации уже несколько лет развиваются новые и более совершенные подходы выявления атак с помощью анализа поведения и аномалий сетевого трафика. Анализ аномалий выявляет существенные отклонения трафика сетевых устройств от «нормального» профиля трафика для данного устройства или группы устройств. Эти алгоритмы предполагают наличие обучения и статистического анализа для построения и обновления «нормального» профиля трафика. Примерами сетевых аномалий являются внезапное увеличение интернет-трафика рабочей станции или изменение структуры трафика (например, увеличение шифрованного SSL-трафика) в сравнении с обычными ежедневными показателями для данной рабочей станции. Для выявления плохого поведения и аномалий в большинстве случаев достаточно анализировать основные параметры трафика (телеметрию) [2].

В 2022 году в исследовании [20] было предложено статистическое распределение для повторяемости редких событий («черных лебедей»). Позже оно было экспериментально проверено на очень большом датасете и полностью подтвердилось, но статья с этими результатами еще находится в печати.

В настоящей работе будет предложен метод оценивания аномалий в текстовых чатах, основанный на предложенном в работе [20] статистическом распределении и на разработанном авторами ранее методе обобщенных *N*-грамм на коллекциях «существительное – глагол».

Цель исследования: разработка нового метода предсказания аномалий в текстовых чатах, не использующего корпусы текстов.

Задачи работы: краткое представление статистического описания повторяемости аномалий, представление метода *N*-грамм на парах «существительное – глагол», синтез методов в новый метод обнаружения аномалий в текстовых данных, тестирование метода, описание результата.

2. Материалы и методы

Для интеллектуального анализа данных в больших системах могут использоваться, в том числе, и статистические методы, ос-

нованные на распределениях, соответствующих задаче. С помощью такого распределения можно, в частности, установить априорные вероятности появления аномалии. При этом детектировать саму аномалию в большой системе в реальном времени может оказаться затруднительно: слишком большой объем данных нужно проанализировать. Если статистическое распределение покажет повышение вероятности аномалии в системе, то далее можно воспользоваться методами классификации или кластеризации, чтобы уточнить прогноз аномалии, а в малых системах (например, чатах с небольшой историей сообщений) можно воспользоваться бэггингом для создания ансамбля и далее использовать те же методы, что и для больших систем.

2.1. РАСПРЕДЕЛЕНИЕ ДЛЯ ДИНАМИЧЕСКОЙ ОЦЕНКИ ВЕРОЯТНОСТИ АНОМАЛИИ

Введем несколько определений.

Серия тестов, или *k-серия* – случайная последовательность фиксированной длины k с двумя возможными исходами (условно 0 и 1), вероятности которых определены и постоянны.

Знаковое событие в серии тестов – это значение 1, появившееся в k-серии. Под *событием* далее понимается произвольный элемент k-серии.

Измерение далее понимается как количественно выраженный результат наблюдения.

Аномалия – это измерение, которое по своим свойствам (числовым значениям) определенным образом отличается от других.

В [20] используется следующая модель аномалии (там она называется инцидентом): аномалии предшествует конечная серия тестов, в которых либо происходят, либо не происходят знаковые события. Длина серии k, число a зарегистрированных в ней знаковых событий, постоянная вероятность p появления события в тесте находятся в списке параметров распределения. Семантическое содержание знакового события может отсутствовать: суть события не важна для вывода распределения. Семантическое содержание аномалии является обязательным (об этом упоминалось выше).

Распределение основано на следующих предположениях:

– измерение имеет фиксированную продолжительность, включающую целое число k-серий;

– при появлении в серии тестов меньше, чем a знаковых событий аномалии не происходит, но в противном случае аномалия происходит обязательно.

В этих предположениях распределение вероятностей аномалии, как показано, определяется вероятностью в пространстве k-серий. Для измерений важен еще параметр s, который характеризует продолжительность измерения в системе, т.е. такое количество k-серий, которое приводит к получению информации о системе (на меньшем промежутке времени наблюдатель не может диагностировать аномалию). Параметр s является еще одним атрибутом распределения. Конкретно, если обозначить через $F_{a,s,k,p}(n)$ вероятность обнаружения первой аномалии в n-м измерении после зарегистрированной аномалии, то ее зависимость от номера измерения имеет вид

$$F_{a,s,k,p}(n) = P_{a,s,k,p}(n-1) \cdot \Big(1 - P_{a,s,k,p}(n)\Big), \quad (1)$$

где

$$P_{a,s,k,p}(n) = \frac{\sum_{m=1}^{\left[\frac{ns}{k}\right]} mC_{ns-m(k-1)}^{m}\left(\sum_{i=0}^{a-1} C_k^i p^i (1-p)^{k-i}\right)^m}{\sum_{m=1}^{\left[\frac{ns}{k}\right]} mC_{ns-m(k-1)}^{m}}. \quad (2)$$

К сожалению, асимптотики этой формулы установить пока не удалось, и поэтому в настоящей работе для вычислений будет использоваться именно это выражение. Характер зависимости $F_{a,s,k,p}(n)$, как показано, таков, что вероятность, как правило, монотонно падает при увеличении n. Но при определенных значениях атрибутов a, s, k, p, может присутствовать максимум при определенном n_{max}; этот максимум и соответствует наиболее вероятному появлению аномалии (см. рис. 1).

В частности, из распределения (1) следует, что повторение аномалии гораздо более вероятно сразу после уже случившейся аномалии или в относительно короткий период времени после нее. Это говорит о том, что редкая аномалия в чате, когда одна тема резко вырывается вперед, не приходит одна: если такое событие случилось, то в скором времени оно, скорее всего, повто-

рится. Если такого рода аномалии приводят к значимым социально-политическим последствиям (как, например, в официальных чатах административных органов) или экономическим последствиям (в чатах корпораций), то от машинного анализатора требуется как можно раньше определить тему, которая выйдет в «супер-топ», т.е. станет второй аномалией. Это можно сделать, комбинируя предсказанное время второй аномалии и обычный подсчет динамики роста популярности тем: аномальной будет наиболее популярная и быстро растущая в преддверии срока повторения аномалии тема. Определив ее, например, за 1–2 дня до попадания ее в супер-топ, можно принять меры и предотвратить нежелательные последствия.

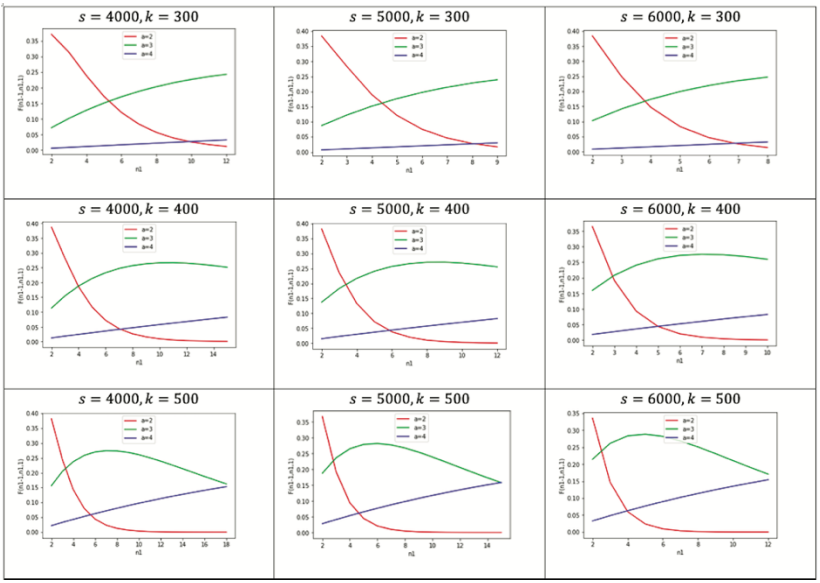

Рис. 1. Вероятность первого повторения при $p = 0{,}001$ [20]

Для использования формулы (1) необходимо определить значения атрибутов для конкретной системы. Это можно сделать методом наименьших квадратов при достаточном количестве данных из рассматриваемой системы или ансамбля подобных систем. Конечно, для этого требуется много данных, и это является

слабым местом предлагаемого метода. Однако то, что метод может быть распространен на статистику в ансамбле систем, позволяет преодолеть этот недостаток, если имеется возможность оперативного обмена информацией между сходными чатами (например, между многими городскими администрациями). Альтернативой является длительное наблюдение в одной системе – это реализовано в эксперименте, который освещен в этой работе.

2.2. РАБОТА С РАСПРЕДЕЛЕНИЕМ ПОВТОРЕНИЯ АНОМАЛИИ

Пусть имеется набор данных $\{x_i\}$, полученный в результате длительного наблюдения за системой, – набор измерений. Это очень простой набор: каждый экземпляр представляет собой 0 или 1, а именно: 0, если в момент времени, определенный индексом i, аномалии не было, а если было, то 1. Длительность наблюдений обеспечивает то, что индекс i пробегает большой ряд значений: $i = 0, \dots, I - 1$, причем $I \gg 1$.

Будем использовать описанную выше модель аномалии. То есть каждой аномалии предшествует последовательный набор из s серий тестов (k-серий), в котором имеется хотя бы одна серия с a или более знаковых событий. Напомним, что внутри набора из s серий тестов измерения невозможны, т.е. события внутри этого промежутка времени ненаблюдаемы, но параметры s, k, a влияют на статистику аномалий.

Упомянутое выше распределение указывает на повышенную вероятность повторения аномалии в более короткий срок, чем срок, который предсказывается статистикой аномалий. Например, если за период 10 000 наблюдений аномалии встретились 40 раз, то статистически следует ожидать следующую аномалию через 250 наблюдений после происшедшей. Но, согласно распределению, она с большой вероятностью наступит гораздо раньше, а период между несколькими (от 2) последовательными аномалиями будет значительно больше, чем 250 наблюдений. Последовательность аномалий, разделенных промежутками времени, намного меньшими, чем время, разделяющие такие последова-

тельности, будем кратко называть *аномальной серией*. Собственно, распределение, полученное в работе [20], показывает наличие таких аномальных серий.

Так как в исследовании речь идет о повторении аномалии, то в первую очередь разобьем выборку на аномальные серии (согласно распределению, они должны быть): число тестов внутри этих аномальных серий будет значительно меньше, чем между ними. Если такие серии выделить не удается, то либо аномалии уединенные (и тогда распределение (1) не работает), либо 1 в последовательности измерений не соответствует аномалии – это измерение не такое уж редкое, т.е. не аномальное.

Алгоритм поиска всех аномальных серий в $\{x_i\}$ приведен в Приложении.

Обозначим найденные аномальные серии через $S_{ij} = \{x_i, \dots, x_j\}$, где $x_i = x_j = 1$, так как серия должна начинаться и заканчиваться аномалией. Каждой такой серии S_{ij} поставим в соответствие число n_{ij}, определенное как средний номер отсчета, на котором аномалия повторилась:

$$n_{ij} = \left[\frac{1}{k}(j_1 + j_2 + \dots + j_k) - i\right],$$

где $j_1, \dots, j_k$ – индексы, для которых $x_{j_1}, \dots, x_{j_k} = 1$ в серии S_{ij}, причем $j_k = j$. Чаще всего $k = 1$ или 2, т.е аномальная серия включает 2–3 аномалии.

Имеем набор значений n_{ij}, которые могут быть использованы для поиска параметров распределения (1), т.е. величин a, s, k, p: для этого достаточно найти максимум функции этих переменных на заданных n_{ij}. Поскольку распределение (1) показывает, что в случае одной и той же рассматриваемой природы аномалии числа n_{ij} не зависят от i, то их распределение одинаково для всех аномальных серий. Таким образом, имеем множество одинаково распределенных случайных величин, среднее значение которых, согласно центральной предельной теореме, распределено нормально.

Предсказание следующего повторения будет, собственно, максимумом функции (1) с параметрами a, s, k, p, т.е. n_{ij} для нового значения i. Если у нас несколько следующих подряд ано-

мальных серий, то, согласно сказанному выше о среднем значении n_{ij}, его можно считать предсказанием возникновения аномалии с нормально распределенным отклонением.

Иначе говоря, если для дальнейших действий не требуется знать явного вида распределения (1), а нужен лишь его максимум, то решать задачу оптимизации и определять a, s, k, p не нужно: можно обойтись статистическими данными за последнее время. Если же необходимо искать продолжительность периода, когда вероятность остается высокой, то все-таки придется находить параметры распределения, но это не является предметом настоящей работы.

2.3. МЕТОД ОБОБЩЕННЫХ N-ГРАММ НА КОЛЛОКАЦИЯХ СУЩЕСТВИТЕЛЬНОЕ – ГЛАГОЛ

Традиционно *N*-граммы используются для генерации текста. При этом текст получается не очень хорошего качества и требует доработки человеком. Для тематического моделирования коротких сообщений без использования больших корпусов текста можно применять *N*-граммы, построенные на коллокациях, имеющих вид пар «существительное – глагол», и получать для каждого сообщения, содержащего хоть одну такую пару или ее часть, его кластер – тему. Для этого достаточно иметь текст, представляющий собой набор сообщений, каждое из которых есть последовательность предложений, и выделять в каждом предложении пары по следующему алгоритму.

1. Создать пару из двух None.
2. Просматривать предложение по словам и перейти к шагам 3 и 4 только для существительных или глаголов.
3. Если последняя пара не заполнена (содержит хотя бы одно значение None), то установить слово в пару на свое место (первое место – существительное, второе – глагол).
4. Если последняя пара заполнена, то образовать новую пару со словом на своем месте и None на другом месте.
5. По окончании просмотра текста удалить пары, содержащие None.

Далее пары группируются в списки: каждому сообщению соответствует свой список пар.

Дальнейшие действия направлены на группирование сообщений по их главным темам, т.е. на тематическое моделирование. Тематическое моделирование является задачей, родственной кластеризации: требуется найти тему (аналог – центр кластера) и сообщения (аналог – точки данных), которые наиболее близки к этой теме. Для этого построен алгоритм, основанный на следующих трех принципах:

– тема сообщения имеется либо среди существительных в парах этого сообщения, либо среди существительных в парах связанных с ним сообщений;

– сообщения связаны, если у них:
  - имеются общие пары (сильная связь);
  - имеются общие существительные (средняя связь);
  - имеются общие глаголы (слабая связь).

– общей темой является существительное с наибольшим весом в группе связанных сообщений.

По аналогии с построением N-грамм, где $(N+1)$-е слово определялось частотой появления слова после данных N слов в базовом тексте, такой метод можно назвать парными N-граммами, но применять его именно для тематического моделирования, а не для генерации текста. В самом методе N-грамм (даже обобщённых) нет никакой новизны. В то же время его конкретная реализация в применении к русскоязычным текстам на коллокациях вида «существительное – глагол» является новой. Применение такого рода N-грамм для тематического моделирования в сочетании с предсказанием аномалий не встречается в литературе.

Множество тем – существительных – представляет собой множество ключей словарей, значениями в которых выступают пары «существительное – глагол» с их весом. Множество тем может быть изначально пустым, но при необходимости в него уже можно заложить некоторые образующие кластеров, т.е. задать темы и соответствующие им пары с весами. Далее для каждой пары в каждом сообщении реализуется следующий алгоритм:

1. Если существительное новой пары содержится хотя бы в одной паре темы, то новая пара включается во все такие темы, а также во вновь созданные для данного сообщения темы с весами, равными 1.

2. Если существительное новой пары не содержится ни в одной паре существующих тем, но глагол содержится хотя бы в одной паре существующих тем, то пара включается в эти темы с весами, равными $0 < \mu < 1$ (задаваемый постоянный параметр), а также создается новая тема по существительному и новая пара включается в нее и во все вновь созданные для данного сообщения темы с весами, равными 1.

3. Если ни существительное, ни глагол новой пары не содержатся в парах существующих тем, то формируется новая тема по существительному и новая пара включается в нее, а также во все вновь созданные для данного сообщения темы с весами, равными 1.

После того как все пары всех сообщений обработаны этим алгоритмом, образуется словарь тематической модели: с ключами – темами (существительными) и значениями в виде словаря с ключами – парами «существительное – глагол», а значениями – вероятностями вхождения пары в тему. Вероятности получаются нормированием весов: вес делится на сумму всех весов пары во всех темах – так получается вероятность принадлежности пары данной теме.

Итогом тематического моделирования является свернутый по значениям вероятностей пар словарь тематической модели: в нем каждому ключу-теме соответствует числовое значение веса этой темы, равное сумме всех вероятностей входящих в нее пар. Отсортированный по убыванию значений, этот объект даст искомые кластеры с их весами.

Материалом для работы послужили чаты, группы и каналы в Telegram, на которые подписан один из авторов работы. Объем текстового материала составил около 50 Мб, что соответствует примерно 2 млн слов, собранным за 5 лет.

Пример. Сообщение «Надеюсь, вернулся уже из всех командировок. 14 мая, как мне сказали, будет круглый стол про личные данные и этическое регулирование. Меня спрашивают – пойду ли я. У меня вопрос к тебе – я тебе там нужен?»

Построенное множество пар (местоимения здесь отнесены к существительным, но в темы не отображаются): (командировка, вернуться), (я, сказать), (стол, быть), (я, спрашивать), (я, идти), (ты, нужен).

Остальные отфильтрованы как неполные. В этот день после подсчета весов тем темы, фигурирующие в этом сообщении, были на следующих позициях в общем рейтинге тем (в скобках – позиция в предыдущий день):

стол – 34 (36),

командировка – 122 (145).

Так случилось, возможно потому, что командировка в этот день обсуждалась и с другими собеседниками, а до этого встречалась редко. Стандартное отклонение, посчитанное к этому дню, равно 5,78, а изменение позиции слова «командировка» равно 23, что превышает стандартное отклонение почти в 4 раза. Это наверняка аномалия (то, как точно определить аномалию, будет разъяснено ниже в алгоритме).

3. Результаты

3.1. АЛГОРИТМ ВЫЧИСЛЕНИЯ НАИБОЛЕЕ ВЕРОЯТНОГО ВРЕМЕНИ НОВОЙ АНОМАЛИИ

Объединяя приведенные выше методы, получаем следующий метод оценки наиболее вероятного времени появления повторной аномалии в текстовых чатах.

1. Определяем источник(и) сообщений и квант времени τ (обычно сутки).

2. Выбираем сообщения за каждый промежуток τ и записываем их в список M_j, где j – номер взятого промежутка времени.

3. Проходим по списку M_j , выполняя алгоритм парных N-грамм. В результате получаем обновленный, сортированный по убыванию значений словарь тем.

4. Вычисляем изменения позиций тем по отношению к промежутку времени $j-1$: строим словарь D_j с ключами-темами и значениями – изменениями их позиций.

5. Определяем тему с наибольшим изменением позиции вверх. Сравниваем это изменение с $K\sigma_j$, где K – глобальный коэффициент, устанавливаемый в качестве параметра, а σ_j – стандартное отклонение в выборке значений словаря D_j. Обычно достаточно принять $K=2$, но если при этом аномалии возникают слишком часто, то можно увеличить значение K. Если позиция

темы выросла более чем на $K\sigma_j$, то распознаем эту тему как аномальную в период j.

6. Вычисляем все такие j, для которых появлялись аномальные темы. Записываем их в список A.

7. Определяем аномальные серии элементов, как описано выше: эти серии (обычно из 2–3 элементов) будут соответствовать одному пику вероятности повторения аномалии.

8. Для каждой выделенной серии из более чем одной точки считаем, что среднее время всех точек, кроме первой, соответствует максимуму кривой $F_{a,s,k,p}(n)$, т.е. $n_{max} = \langle j \rangle_1$, где среднее берется по всем элементам группы, кроме первого.

9. Считаем, что $n_{max} = n_{max}$ для следующей серии, которая начнется, как только будет зарегистрирована следующая аномалия. Обучение закончено: параметр n_{max} определен.

10.При дальнейшем функционировании предиктора аномалий – обнаруживать аномалии по критерию, указанному в п. 5 с тем же значением K, что было принято при обучении, и выдавать прогноз в виде n_{max} периодов τ, основанный на нескольких последних сериях измерений (число серий измерений можно установить вручную, в эксперименте ниже оно равно 5).

Выполнение этого алгоритма производится на протяжении всей работы предиктора, и значение прогнозного максимума вероятности аномалии может регулярно изменяться, но, как следует из соображений, приведенных при выводе распределения (1) в работе [20], если причина возникновения аномалий в чате не меняется, то и величина n_{max} сильных изменений не претерпит. Причина возникновения аномалий может измениться только при изменении системных характеристик чата (значительно изменится число абонентов, изменится характер деятельности хозяина чата и т.п.).

Проведенный эксперимент имеет следующий сценарий. Период измерения – 1 день. То есть определение тем и их сортировка производятся каждый день. Скользящим образом определяется значение стандартного отклонения σ в дневном изменении позиций тем (исчезающие и появляющиеся темы не учитываются при расчете стандартного отклонения). Если в текущий

день какая-то тема изменила позицию больше, чем на $K\sigma$ вверх, то этот день объявляется аномальным (аномалией). Задача состоит в том, чтобы предсказать следующий аномальный день.

В эксперименте использовалось значение $K = 1{,}85$. В истории записей по всем чатам, группам и каналам было 1712 дней и было зафиксировано 52 аномалии, из которых только две не имели соседей по серии (стояли уединенно). Аномалии были разделены на 5 последовательных серий: $S_{33,36}, S_{95,99}$, ..., т.е. вычисление параметров производилось по итогам 5 усредненных значений сроков повторения аномалии (см. рис. 2). Результат показан на рис. 3.

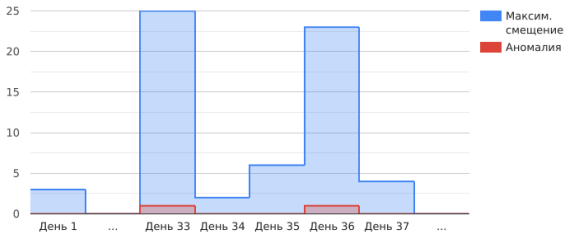

Рис. 2. Построение аномальной серии
(*показано одно повторение аномалии*)

Аномальных серий в эксперименте зафиксировано 24 и все они включали не более 3 аномалий. Например, в первой аномальной серии случаи повышения какой-либо темой своей позиции более чем на текущее значение $K\sigma$ регистрировались на 33 и 36 день после начала наблюдения. То есть аномальная серия содержала две аномалии и промежуток времени между ними равен 3. Как видно из графика на рис. 3, были случаи, когда этот промежуток времени был равен 1 и даже 0 (т.е. две аномалии произошли в один день). Но среднее значение ожидаемого наступления аномалии после уже зарегистрированной колебалось между 6 и 8 днями. Все 24 аномальные серии (изображены зелеными точками) расположены в первых пяти измерениях (с 0 по 4). Для измерения 5 фактических значений нет, есть только предсказание.

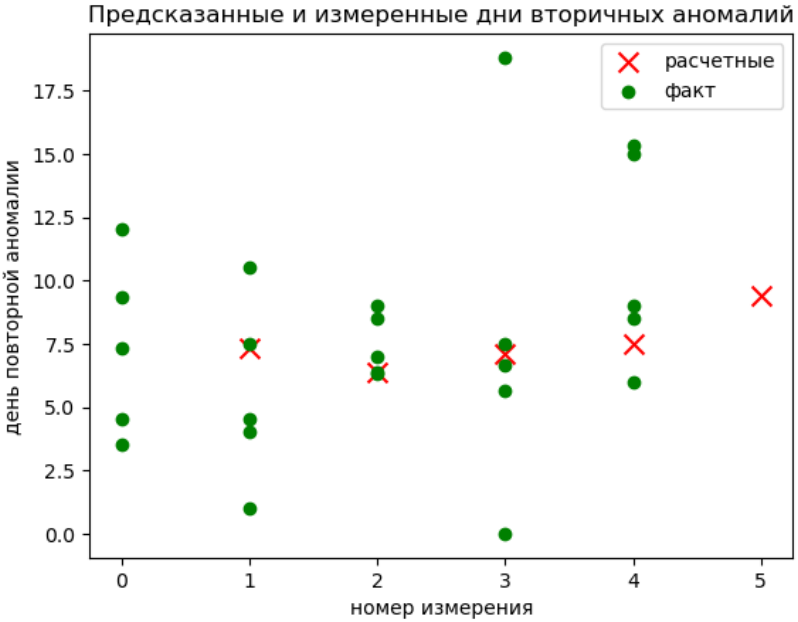

Рис. 3. Предсказанные (*крест*) *и фактические* (*точка*) *дни повторения аномалий*

На диаграмме видно, что предсказания дня повторения аномалии в целом соответствуют тенденциям появления фактически регистрируемых аномалий, хотя погрешность может быть существенной. Это объясняется тем, что максимум распределения (1) для этих аномалий не является острым.

Если аномалии наступают независимо друг от друга, а зависят от наличия каких-то наблюдаемых признаков в цепи событий, то случайный процесс появления аномалий стационарен при условии стационарности процессов появления признаков. В этом случае вероятность появления аномалии после уже случившейся будет иметь максимум через число дней, равное математическому ожиданию нормального распределения. Это математическое ожидание на всей выборке можно оценить как $\mu = \frac{i_l - i_f}{n_a - 1}$, где i_l – индекс последней аномалии; i_f – индекс первой аномалии; n_a – число аномалий. В нашем случае $\mu = 32{,}45$, что явно

намного выше предсказанного значения, обусловленного нахождением аномалии в аномальной серии. Иными словами, *p*-значение для гипотезы повторения в сериях будет значительно больше, чем для гипотезы нормального распределения без учета аномальных серий. Это говорит о том, что гипотеза аномальных серий подтверждается на этой выборке.

Оценим точность предсказания дня повторения аномалии в нашей модели и, для сравнения, в авторегрессионной модели ARIMA (с тем же критерием аномалии).

Точность метода можно оценить по успешности предсказания аномалии в определенном периоде после случившейся аномалии. Предложенный нами метод предсказывает наиболее вероятный день повторения аномалии. После первых пяти аномальных серий, которые произошли за 453 дня, наша модель показала, что следующая аномалия повторится на $\mu = 7$ день со стандартным отклонением 1,47, т.е. с 2 по 12 день по уровню 3σ. Из 5 следующих аномальных серий таких было 4, т.е. для них можно считать $Acc = 80\%$. Далее можно построить еще три показателя *Acc* по имеющимся данным. Общий показатель по всем измерениям равен $Acc = 76\%$.

Теперь посмотрим на прогноз ARIMA. После обучения на 453 днях модель ARIMA выдала последовательность аномалий, которая разделилась на аномальные серии иначе, чем последующие реальные данные. Если говорить, как и в нашей модели, только о повторении аномалии, т.е. о прогнозной длине аномальной серии, то в модели ARIMA после первого обучения получилось только 13 аномальных серий (остальные аномалии предсказаны уединенными). Поэтому разделить на измерения таким же образом, как в случае нашей модели, не получится и метрику качества нужно считать, взяв среднюю повторяемость по этим 13 сериям, а стандартные отклонения – те, которые были в реальных данных (те же, что взяты выше для нашей модели). Тогда для обученной на 453 днях модели ARIMA, получим $\mu_{av} \approx 14$ и в интервал 3σ попадут только 6 из 19 реальных серий, т.е. $Acc = 32\%$.

Если обучить модель ARIMA на данных, полученных после регистрации следующих 5 аномальных серий, то качество прогноза станет лучше (по крайней мере коэффициент детерминации значительно вырастет), но все равно из оставшихся 14 реальных аномальных серий, только 5 будут соответствовать предсказанию. То есть $Acc = 36\%$.

Надо сказать, что рассматриваемые аномалии не относятся к одинаковым темам; в исследовании делался упор на аномалию как таковую, а какая именно тема резко поднимает свою значимость – не имело значения. Это подчеркнуто выше, когда говорилось о процедуре использования предсказания срока повторения аномалии. То есть применительно к результату эксперимента надо сказать, что повторную аномалию следует искать в течение примерно двух недель от уже случившейся, а наиболее вероятно – на 7–9 день. С учетом того, что аномалии в этом чате возникали, в среднем, чуть реже, чем раз в месяц, можно сказать, что период между сериями примерно равен двум месяцам. То есть после двух недель увеличения риска повторения аномалии далее следует примерно полуторамесячный период «спокойствия». Но, что самое главное, в течение двух недель аномалия почти всегда возникала и, если ее вовремя выявить, то можно было принять меры.

4. Заключение

В работе предложен новый метод предсказания возникновения аномалий в потоке текстовых сообщений. Решение основано на модели аномалии, предложенной в работе [20] и учитывающей скрытые причины аномалий. В работе использовался алгоритм тематического моделирования коротких сообщений на основе коллокаций типа «существительное – глагол». Путем объединения метода динамической оценки риска повторения редких событий и метода тематического моделирования с помощью пар «существительное – глагол» построен алгоритм предсказания срока наиболее вероятного наступления следующей аномалии после уже зарегистрированной в системе.

Применение предложенного метода может быть полезно в исследованиях и анализе появления аномалий в сложных социальных системах, взаимодействие в которых отражается в коммуникациях через социальные сети и мессенджеры.

Приложение. Алгоритм поиска аномальных серий в выборке

Алгоритм проще всего проиллюстрировать в виде программы на языке Python.

```
def aseries(x: list):
    """
    Определяет аномальные серии в длинной серии бинарных исходов.
    Аргумент
    ----------
        x (list): серия измерений с двумя исходами (0 или 1)
    Возврат
    -------
        Список пар чисел: [индекс начала аномальной серии в списке x,
        количество измерений в серии, помимо первого]
    """
    dx = {}
    for i,xi in enumerate(x[:-1]):
        if xi:
            dx[i]=x[i+1:].index(1)+1
    vdx = list(dx.values())
    adx = sum(vdx)/len(vdx)
    aseries_points = []
    for k in dx:
        if dx[k]<adx:
            aseries_points.append([k,dx[k]])
             return aseries_points
```

Литература

1. КУЗОВЛЕВ В.И., ОРЛОВ А.О. *Выявление аномалий при прогнозном анализе данных* // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. – 2016. – № 5. – С. 75–85.

2. МИКОВА С.Ю., ОЛАДЬКО В.С. *Сетевые аномалии и причины их возникновения в экономических информационных системах* // Издательский центр «ИУСЭР. – Экономика и социум. – 2015. – №3 (16) – С. 76–81.
3. САВЕНКОВ П.А., ИВУТИН А.Н. *Методы анализа естественного языка в задачах детектирования поведенческих аномалий* // Известия ТулГУ. – Технические науки. – 2022. – №3. – С. 358–366.
4. САВИЦКИЙ Д.Е., ДУНАЕВ М.Е., ЗАЙЦЕВ К.С. *Выявление аномалий при обработке потоковых данных в реальном времени* // Int. Journal of Open Information Technologies. – 2022. – №6. – С. 70–76.
5. ЧАСТИКОВА В.А., КОЗАЧЁК К.В., ГУЛЯЙ В.Г. *Методы обработки естественного языка в решении задач обнаружения атак социальной инженерии* // Вестник Адыгейского государственного университета. – Сер. 4: Естественно-математические и технические науки. – 2021. – №4 (291). – С. 95–108.
6. https://hadoop.apache.org/ (дата обращения: 01.10.2021).
7. BENMAHDI D., RASOLOFONDRAIBE L., CHIEMENTIN X. et al. *RT-OPTICS: real-time classification based on OPTICS method to monitor bearings faults* // Journal of Intelligent Manufacturing. – June 2019. – Vol. 30, Iss. 5. – P. 2157–2170.
8. BORJ P.R., RAJA K., BOURS P. *Online grooming detection: A comprehensive survey of child exploitation in chat logs* // Knowledge-Based Systems. – 2023. Vol. 259. 110039. – DOI: https://doi.org/10.1016/j.knosys.2022.110039 (accessed 1 July 2023).
9. GUPTA A., MATTA P., PANT B. *Identification of Cyber-criminals in Social Media using Machine Learning* // Int. Conf. on Smart Generation Computing, Communication and Networking (SMART GENCON). – Bangalore, India. – 2022. – P. 1–6. – DOI:10.1109/SMARTGENCON56628.2022.10084119.
10. LEMAIRE V., ALAOUI ISMAILI O., CORNU´EJOLS A. et al. *Predictive k-means with localmodels* // In: Workshop LDRC–2020 (Workshop on Learning Data Representation for Clus-

tering) in PAKDD–2020 (The 24th Pacific-Asia Conf. On Knowledge Discovery and DataMining). – May 2020. – Singapore. – P. 11–16.
11. MD TAHMID RAHMAN LASKAR, JIMMY XIANGJI HUANG, SMETANA V. et al. *Extending Isolation Forest for Anomaly Detection in Big Data via K-Means* // ACM Trans. – Cyber-Phys. Syst. 5, 4, Article 41. – 2021. – 26 p. – DOI: https://doi.org/10.1145/3460976.
12. SARVANI A., VENUGOPAL B., DEVARAKONDA N. *Anomaly Detection Using K-means Approach and Outliers DetectionTechnique* // In: Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, Springer, Singapore. – 2019 – P. 742.
13. SHERIFF M.Z., NOUNOU M.N. *Improved Fault De-tection and Process Safety Using Multiscale Shewhart Charts* // Chem. Eng. Process Technol. – 2017. – Vol. 8(2). – P. 1–16. – DOI: 10.4172/2157-7048.100032.
14. TALEB N.N. *Black Swan and Domains of Statistics* // The American Statistician. – 2007. – Vol. 61, No. 3.
15. TSIGKRITIS T., GROUMAS G., SCHNEIDER M. *On the Use of k-NN in Anomaly Detection* // Journal of Information Security. – 2018. – Vol. 9. – P. 70–84.
16. https://spark.apache.org/ (дата обращения: 01.10.2021).
17. VANNEL Z., DONGHYUN K., DAEHEE S., *Ahyoung Leea An unsupervised anomaly detection frame-work for detecting anomalies in real time through network system's log files analysis* // High-Confidence Computing. – 2021. – Vol. 1, Iss. 2.
18. WANG Z., ZHOU Y.H., LI G.M. *Anomaly Detection by Using Streaming K-Means and Batch K-Means* // 5th IEEE Int. Conf. on Big Data Analytics (IEEE ICBDA 2020). – Xiamen, China, 8–11 May 2020. – P. 11–17.
19. ZIMEK A., SCHUBERT E. *Outlier Detection* // *Encyclopedia of Database Systems*. – Springer New York, 2017. – DOI: 10.1007/978-1-4899-7993-3_80719-1.
20. ZUEV S., KABALYANTS P. *On the black swan risk dynamical evaluation* // Int. Journal of Risk Assessment and Management. – 2022. – Vol. 25, No. 1/2. – P. 56–66.

MACHINE MONITORING OF TEXT CHATS AND DETECTION OF ANOMALIES

Elena Mozaidze, Belgorod State Technological University named after V.G. Shukhov, Belgorod, graduate student (mozaidze95@mail.ru).
Sergei Zuev, V.I. Vernadsky Crimean Federal University, Simferopol, Cand.Sc., associate professor (sergey.zuev@bk.ru).

Abstract: The aim of the work is to develop a new method for detecting anomalies in text chats that does not use text corpora. Tasks: a brief presentation of the statistical description of the recurrence of anomalies developed in the authors' previous works, the introduction of the method of paired (generalized) N-grams, the synthesis of these methods into a new method for detecting anomalies in short message exchange systems, the method testing. A new method for detecting anomalies in the flow of text messages is proposed, which does not use a corpus of texts for learning, and, in addition, allows online learning. The material for the work was chats, groups and channels in Telegram, to which one of the authors of the work is subscribed. The volume of text material was about 50 MB, which corresponds to about 2 million words collected over 5 years. The method uses a statistical distribution of the repetition of anomalous events, as well as a method of thematic modeling based on the statistics of noun-verb pairs. Both methods were proposed earlier in the authors' works. The experiment showed that the results predicted by the proposed method correspond to the actually registered anomalies. The application of the proposed method can be useful in research and analysis of the appearance of anomalies in complex social systems, the interaction in which is reflected in communications through social networks and messengers. Such tasks are relevant both for government agencies and for business, and can help to smooth out acute social and industrial problems. The proposed method is seemed especially useful for the journalism because it allows you to determine the time of the most likely appearance of significant social phenomena.

Keywords: anomaly detection, topic modeling, probabilities of rare events, repeatability of rare events, anomalies in text chats.

УДК 004.8
ББК 22.17
DOI: 10.25728/ubs.2024.109.4

Статья представлена к публикации членом редакционной коллегии Н.Н. Бахтадзе.

Поступила в редакцию 19.10.2023.
Опубликована 31.05.2024.