Организация сбора и обработки данных социодинамических процессов с возможной самоорганизацией и наличием памяти и анализ наблюдаемых характеристик их временных рядов

К.К. Отраднов, В.Н. Калинин, С.А. Лесько, И.В. Платонова

Аннотация: **В статье рассмотрены вопросы разработки специализированного программного обеспечения для сбора, обработки и хранения данных социодинамических процессов (изменение эмоциональной окраски комментариев пользователей к опубликованным новостям в сетевых массмедиа, и электоральных кампаний выборов Президента США в 2012 и 2016 годах). Показано, что для его создания можно использовать конвейерный принцип с реализацией микросервисной архитектуры, а для хранения данных, с учетом их специфики и происхождения, предпочтительнее применение графовых баз данных.**

На основе собранных данных были получены временные ряды наблюдаемых процессов. Их R/S анализ показал, что они обладают антиперсистентностью. Исследование зависимости математического ожидания, дисперсии и эксцесса амплитуд отклонений уровней ряда от размеров интервала времени расчета амплитуд ("скользящего окна") показало, что для математического ожидания наблюдается корневая зависимость дробной степени; для дисперсии - степенной закон с дробным показателем больше 1,5; а поведение эксцесса показывает наличие так называемых «тяжелых хвостов», его величина существенно больше чем у нормального распределения.

Полученные результаты указывают на то, что временные ряды рассматриваемых процессов обладают нестационарностью, нелокальностью, как по времени (имеют память), так и состоянию (проявляют самоорганизацию).

Ключевые слова: **временные ряды, самоорганизация, наличие памяти, нестационарность, фрактальность временного ряда, социодинамические процессы, графовые базы данных, микросервисная архитектура ПО.**

Статья получена 10 февраля 2024.

Работа выполнена при финансовой поддержке Российского научного фонда (РНФ), грант № 23-21-00153 «Анализ и моделирование динамики нестационарных временных рядов фрактальных процессов с реализацией памяти (последействия) и самоорганизацией на основе использования дифференциальных уравнений с дробными производными».

К. К. Отраднов, Институт радиоэлектроники и информатики, РТУ МИРЭА, Москва, Россия (e-mail: const.otradnov@yandex.ru).

В. Н. Калинин, кафедра прикладной информатики и интеллектуальных систем в гуманитарной сфере, ФГАОУ ВО РУДН, Москва, Россия (e-mail: vkalininz@mail.ru).

С. А. Лесько, Институт кибербезопасности и цифровых технологий, РТУ МИРЭА, Москва, Россия (e-mail: sergey@testor.ru).

И. В. Платонова, Физический факультет МГУ им. М.В. Ломоносова, Москва, Россия (e-mail: platonovaiv@mail.ru).

I. Введение

Социодинамические процессы имеют широкое распространение и большое разнообразие своих проявлений (например, активность пользователей в социальных сетях, электоральные кампании, сетевой маркетинг и т.д.). В частности, новости и блоги, под которыми пользователи социальных сетей и масс – медиа оставляют свои комментарии или высказывают к ним свое эмоциональное отношение, являются одними из важнейших явлений, которое может в онлайн режиме выполнять роль индикатора общественного мнения и настроения.

Эмоциональное состояние является очень важным фактором не только социологии, но и экономики. Поскольку позитивные настроения являются индикатором уверенности в будущем, увеличивают потребительскую активность и тем самым могут способствовать росту экономики. С другой стороны, депрессивные тенденции в общественных настроениях вызывают неуверенность в завтрашнем дне и тем самым могут способствовать снижению деловой активности и спаду в экономике.

Если в новостях или блогах затрагивается общественно значимая тема, то это как правило притягивает, как её сторонников, так и противников, которые вступают в дискуссии и оставляют взаимные комментарии. И чем более резонансной является новость или блог, тем выше активность пользователей и больше число комментаторов (появляется многоуровневая структура комментариев к комментариям), которые проявляют свое эмоциональное отношение к описываемым в новостях событиям.

Ещё одним очень важным типом социодинамических процессов являются электоральные процессы. Очевидно, что прогнозирование тенденций мирового развития в значительной степени определяется динамикой и результатами прохождения электоральных кампаний различного вида в ведущих мировых державах.

Изменения активности пользователей сетевых массмедиа и участников электоральных кампаний образуют временные ряды, анализ динамики которых и их характеристик, может играть важную роль с точки зрения прогнозирования направленности социодинамических и экономических процессов.

Сложные системы (а социальные сети, сетевые массмедиа и электоральные кампании относятся именно к таким системам) можно определить, как структуры, в которых в качестве хотя бы одного из элементов выступает человек. С одной стороны, действие множества случайных факторов приводит к стохастичности свойств таких систем. Кроме того, помимо стохастичности ещё возникает и неопределенность, связанная в некоторых случаях с иррациональностью поведением людей. А с другой стороны, наличие человеческого фактора создает предпосылки для самоорганизации таких системы и может определять существование памяти о их предыдущих состояниях. Всё это приводит к появлению организованной сложности (эмерджентности). Возникновение которой нельзя рассматриваться как итог простого суммирования характеристик элементов, а является результатом возникновения системных связей и адаптивного перераспределения функций между элементами. Следует отметить, что модели описания большинства процессов, наблюдаемых в сложных социальных и экономических системах, может быть отнесено к классу нестационарных временных рядов. Из-за нелинейности и изменчивости характеристик процессов, протекающих в сложных социальных и экономических системах традиционные методы анализа и моделирования, например, такие как интегрированная модель авторегрессии - скользящего среднего (ARIMA, модель Бокса – Дженкинса) и многие другие модели часто приводят к неточным или ошибочным результатам.

Всё это говорит о том, что изучение социодинамических процессов является очень важным, как с научной, так и практической точек зрения.

II. Создание специализированного программного обеспечения для сбора, обработки и хранения данных социодинамических процессов

Для сбора, обработки и хранения данных социодинамических процессов требуется разработать специализированное программное обеспечение. В данной разрабатываемой системе каждый этап получения, обработки и сохранения информации может быть представлен собственным модулем, являющимся независимой единицей и базирующимся на принципах микросервисной архитектуры [1]:

1) реализуют функционал согласно своему предназначению (осуществляют поиск и сбор данных, обработку);

2) GUI-сопряжение, служит для формирования данных графического интерфейса пользователя;

3) база данных;

4) самодиагностика (микросервисные архитектуры отличаются на порядок возрастающим количеством и сложностью взаимодействий, при этом многокомпонентная распределённая среда затрудняет обнаружение ошибок и сбоев; данных модуль призван нивелировать эти недостатки архитектуры);

5) коммуникации, обеспечивает эффективный обмен сообщениями между микросервисами;

6) документация, организация документации, как для пользователей, так и разработчиков;

7) командный интерфейс пользователя.

Преимуществом микросервисной архитектуры является возможность ведения параллельной (модульной) разработки и апробации решений, а также возможность использования различных стеков технологий для реализаций той или иной части решения. Это позволяет различным разработчикам практически независимо друг от друга создавать различные программные модули для сбора и обработки данных.

При выборе основного языка программирования рекомендуется предпочтительно использовать Python [2 - 4]. Этот выбор обусловлен относительной легкостью освоения, удобством работы с текстовыми данными и наличием обширного набора готовых вспомогательных библиотек, таких как Requests, BeautifulSoup, lxml, re, multiprocessing, tqdm, sqlite3, datetime, math, NumPy, pandas, Plotly и др.

Сбор данных происходит по следующему алгоритму:

1) выбирается количество статей или диапазон времени, за которое должен быть произведен сбор (парсинг) данных с новостного портала;

2) определяется количество свободных виртуальных приватных серверов (VPS) (worker-node) и указывается их количество;

3) проводится равномерное разделения количества новостей или количества дней между всеми узлами (виртуальными приватными серверами);

4) выбирается количество потоков на каждом узле (обычно равное количеству потоков сервера минус 1);

5) в цикле от первого до последнего дня или по количеству новостей (для части значений, выданных на сервер) генерируются ссылки на определенные дни или на кол-во статей.

6) далее ссылки равномерно делятся на количество потоков, после чего каждый поток обрабатывает свой объем ссылок и возвращает результат.

Потоки работают следующим образом:

1) получив ссылки, в цикле обрабатывается по одной ссылке, для каждой ссылки отправляется GET запрос, и возвращается ответ в виде HTML страницы;

2) HTML страница распарсивается с помощью библиотек BeautifulSoap и lxml;

3) по полученным данным происходит поиск (поиск по DOM элементам), где ищется информация о статье: заголовок статьи, текст статьи и остальные данные;

4) после полученной информации по статье, идет поиск информации по комментариям, для этого в полученных HTML данных ищется ссылка на комментарии, выделяется, после чего идет отправка GET запроса на получения ответа по этой ссылке;

5) полученный ответ также распарсивается, после чего выделяется структура комментариев;

6) после выделения структуры комментариев, идет сбор информации о комментарии, такой как: текст комментария, дата и время публикации комментариев, пользователь оставивший комментарий и так далее;

7) после сбора данных о комментарии, собирается информация о вложенности комментариев внутри других комментариев. Это необходимо, так как уровень комментария не присутствует в метаданных новостного

портала и понять, как один комментарий относится к другому по уровню комментирования можно только визуально или с помощью HTML просмотрев вложения одного в другой (ответы комментаторов на комментарий);

8) после чего из собранной ранее информации с комментариев по пользователям, удаляются повторы, производится поиск в общей базе данных о наличии данного пользователя в ней, при отсутствии записи, выполняется GET запрос на страницу пользователя, из которой собираются необходимы данные, такие как: имя пользователя, количество рекомендаций и другие;

9) после сбора данных по всем этапам для новости (новость, комментарии, пользователи), данные заносятся в базу данных;

10) после обработки всех ссылок, worker-node возвращает информацию по всем собранным данным и объединяет их в единую базу данных.

Процесс размещения данных в базе данных происходит следующим образом:

1) после выполнения задания сервер передает полученный результат в «control-node»;

2) затем «control-node» устанавливает подключение к СУБД и заносит данные о новости и комментарии;

3) при добавлении объекта «пользователь» происходит проверка на его существование. Если пользователь уже существует, данные не записываются, чтобы избежать дублирования и переполнения базы данных.

Для проведения исследования активности пользователей было необходимо выполнить предобработку текстов, которая включает в себя токенизацию, нормализацию текстов и удаление стоп слов из текстов.

Предобработка тестов была выполнена с помощью программного кода, написанного на языке Python. Для этого были использованы следующие библиотеки:

1) docx - предназначена для создания и обновления файлов с расширением .docx.

2) re – библиотека для работы с регулярными выражениями в Python.

3) pymorphy2 – морфологический анализатор для русского языка.

4) nltk – библиотека для обработки естественного языка.

Из библиотеки pymotphy2 был взят класс MorphAnalyzer, данный класс был необходим для использования метода данного класса – normal_forms, который принимает в себя слово и приводит его к нормальной форме.

Из библиотеки nltk был взят словарь со стоп-словами русского языка.

Для создания векторных моделей TF-IDF в лексическом подходе была использована библиотека sklearn.

Более того, помимо предложенных ранее алгоритмов, применялся метод векторизации текста Word2vec. Для обучения и создания векторной модели текста использовалась python библиотека gensim, предназначенная для неконтролируемого тематического моделирования, индексирования документов, поиска по сходству и других задач обработки естественного языка с применением современных методов статистического машинного обучения.

Из библиотеки genism была взята модель Word2vec, которая использует следующие параметры:

1) min_count –отвечает за минимальную частоту вхождения слова в текстах.

2) window –отвечает за максимальное расстояние между текущим и предсказанным словом в предложении.

3) vector_size - отвечает за максимальную размерность векторов слов.

4) negative –отвечает за добавление шума в выборку слов, а именно сколько «шумовых слов» будет использоваться в выборке.

5) alpha –отвечает за начальную скорость обучения.

6) min_alpha –отвечает за минимальную скорость обучения, так как в процессе обучения скорость будет линейно снижаться.

7) sample –отвечает за порог для настройки того, какие высокочастотные слова будут случайным образом сокращены.

8) sg – отвечает за алгоритм обучения 1 – skip gram, 0 – CBOW.

Программа анализа временных рядов для определения их нестационарности, наличия самоорганизации и памяти. Проведение расчетов и визуализация данных включают в себя следующие технологические этапы: построение временных рядов; их обработка; построение графиков; построение гистограмм амплитуд активностей пользователей; зависимостей матожидания, дисперсии, асимметрии, эксцесса амплитуд активностей от интервала времени их определения; определение показателя Херста; вычисление параметров разработанных моделей; построение графов сетей комментариев пользователей; анализ параметров графов; сентимент анализ текстов пользователей.

Python имеет несколько библиотек для визуализации данных. Наиболее популярная из них - Matplotlib (подготовка данных; выбор типа графика; создание графика; настройка внешнего вида графика; представление графика и сохранение в файл).

Графовая база данных [5]. В результате изучения различных баз данных после детального исследования был сделан выбор в пользу использования для создания системы их хранения графовых баз данных. Которые в сравнении с традиционными таблицами имеют большую гибкость и адаптацию к изменяющимся структурам собираемых, обрабатываемых и хранимых данных.

С развитием онлайн-социальных сетей возникает растущая потребность в обработке графовых данных, необходимо учитывать структуру собираемой, хранимой и обрабатываемой информации, а также взаимосвязи между её отдельными единицами. Например, в каждой статье сетевых массмедиа, будь то на новостном портале или в социальной сети, содержатся следующие данные: идентификатор (id) статьи на портале; заголовок статьи; веб-адрес новости; дата и время публикации (timestamp); текст новости (context); количество комментариев и просмотров (метаданные); комментарии пользователей новостного ресурса; краткая информация о пользователях, оставляющих комментарии (никнейм,

интересы, статистика и метаданные).

В графовых базах данных данные организованы в виде графа, где узлы представляют объекты, а ребра - связи между ними. Каждый узел может иметь несколько свойств, а каждое ребро может иметь свойства, описывающие характер связи между узлами [5].

Основное различие между табличными и реляционными базами данных заключается в методе обработки запросов. В реляционных базах данных используется язык SQL для выполнения запросов, который предоставляет возможность выбора данных из таблиц, их объединения и выполнения прочих операций. В графовых базах данных запросы обрабатываются посредством языка запросов к графу, который позволяет выбирать узлы и ребра текущего графа, а также проводить разнообразные математические преобразования над графом.

Главным преимуществом графовых баз данных перед реляционными является возможность эффективно работать с данными, имеющими сложную структуру связей, что и определило их выбор при реализации проекта.

Графовая база данных может эффективно хранить информацию о пользователях, их профилях и взаимосвязях между ними, такими как дружба, подписки, лайки и комментарии. Каждый пользователь представлен узлом графа, а связи между ними - ребрами.

Графовые базы данных позволяют использовать большой набор алгоритмов для анализа графовых структур. Это включает в себя поиск кратчайших путей, выявление центральных узлов (важных пользователей), обнаружение сообществ (групп пользователей с похожими интересами) и многое другое.

Графовые базы данных могут использоваться для создания рекомендательных систем, которые предлагают пользователям контент или друзей на основе их предыдущих действий и интересов в социальной сети.

Графовые алгоритмы могут помочь выявлять аномалии в сети, такие как необычные активности или попытки мошенничества, анализируя структуру связей между пользователями.

Графовые базы данных позволяют создавать наглядные визуализации социальных сетей, что облегчает понимание и визуальное представление структуры и взаимосвязей.

Проектирование графовой базы данных состоит из следующих этапов:

1) Определение сущностей и связей: Определение сущностей, которые будут представлены как узлы в графе. Определение связей (рёбра) между сущностями и их характеристики. Например, для социальной сети связями могут быть "дружба", "подписка" и т.п.

2) Разработка схемы графа: Создание схемы графа, с определением типов узлов и связей, и их атрибуты (свойства). Необходимо определить, какие атрибуты будут храниться в узлах, а какие — в связях, для оптимизации доступа к данным.

3) Учет запросов и использования: Учет типичных запросы и операций, которые будут выполняться в графовой базе данных. Данный этап поможет определить, какие индексы и структуры данных нужны для оптимизации запросов.

Пример проектирования базы данных для сбора данных из социальной сети. В социальной сети присутствуют такие объекты как:

1) сообщество;

2) пост;

3) комментарий;

4) пользователь.

Кроме того, важно помнить о том, что:

1) сообщество состоит из различного количества постов;

2) под постами пишут комментарии;

3) один пользователь может написать сколько угодно комментариев;

4) пользователь может "дружить" (быть подписчиком) с другим пользователем.

В связи с этим, при проектировании реляционной базы данных, необходимо создать 4 таблицы, для хранения каждого объекта социальной сети, и пятую промежуточную таблицу, для организации взаимосвязи типа "друг" между пользователями.

Диаграмма реляционной базы данных представлена на рисунке 1.

Рис. 1. Проектирование реляционной базы данных для сбора данных из социальной сети.

В отличии от реляционной базы данных, для графовой базы данных необходимо определить:

1) типы объектов (узлов) и их метаданные;

2) взаимосвязи узла.

Так, для графовой базы данных, определяем 4 типа узлов (рисунок 2):

1) узел сообщество;

2) узел пост;

3) узел комментарий;

4) узел пользователь.

Также определяется 4 типа взаимосвязей (рисунок 3):

1) сообщество - пост;

2) пост - комментарий;

3) комментарий - пользователь;

4) пользователь - пользователь.

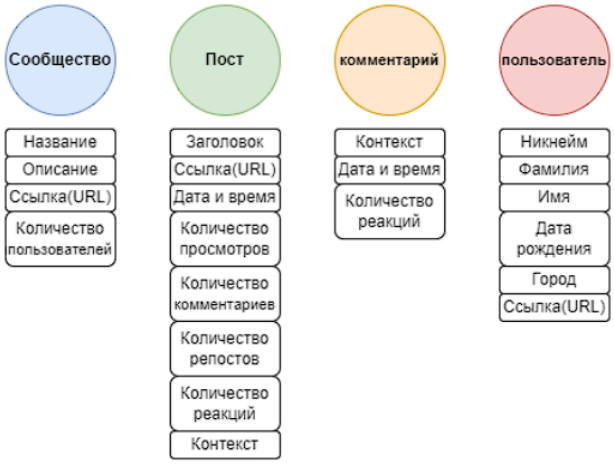


Рис. 2. Типы объектов и их метаданные в графовой базе данных.

Рис. 3. Взаимосвязи в графовой базе данных.

В итоге, можем получить результат модели графовой базы данных, представленный на рисунке 4.

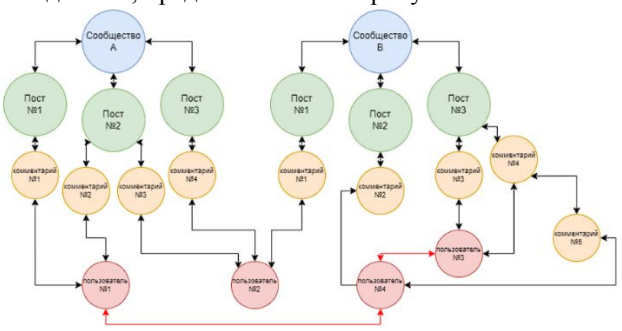


Рис. 4. Модель графовой базы данных.

Расматриваемые нами модули в совокупности создают микросервисную архитектуру и включают в себя:

Master-node - обеспечивает взаимодействие, распределение задач на worker-node и получение отчетов от Control-node. Он также имеет доступ к предыдущим результатам, сохраненным в Backup-node.

Worker-node - это виртуальная машина или отдельный сервер с фиксированным объемом CPU и RAM. Он выполняет задачи по получению и предварительной обработке данных. Для этого применяются алгоритмы, созданные на основе MapReduce.

Control-node - отслеживает выполнение задач, предоставляет информацию о состоянии сервера, времени выполнения и ожидаемом времени (по предварительным расчетам). Control-node принимает задания от worker-node и передает их на master-node, только если задание полностью выполнено; регулярно сохраняет промежуточные данные на backup-node, которые удаляются после обработки всех данных.

Backup-node выполняет две основные функции: служит хранилищем для временных значений, которые могут быть использованы в будущем; сохраняет готовые результаты и отчеты, после завершения процесса или задачи.

Certificate authority and trusted storage – выполняет роль корневого удостоверяющего центра, где хранятся конфиденциальные данные, и обеспечивает безопасность системы.

Использованные библиотеки.

1) py2neo предназначена для взаимодействия с графовой базой данных Neo4j, использующей графовую модель данных для хранения и обработки информации. Эта библиотека предоставляет удобные инструменты для создания, обновления, удаления и выполнения запросов к данным в Neo4j с использованием языка программирования Python. Она упрощает взаимодействие с базой данных, предоставляя объектно-ориентированный API для работы с узлами и отношениями. Кроме того, библиотека обеспечивает поддержку асинхронного взаимодействия с Neo4j, что позволяет выполнять запросы асинхронно и эффективно обрабатывать большие объемы данных.

2) NetworkX представляет собой библиотеку, созданную для работы с графами и анализа сетей. Она обеспечивает удобные средства для формирования, изменения и изучения структур графов. NetworkX разработана с упором на простоту использования и понимания, предоставляя простой и интуитивно понятный интерфейс для взаимодействия с графами. Эта библиотека поддерживает разнообразные виды графов, такие как направленные и ненаправленные, мультиграфы (с возможностью наличия нескольких ребер между одной и той же парой узлов) и графы с весами (где ребра имеют числовые значения).

3) Matplotlib - это библиотека, которая предоставляет возможности для формирования статических, интерактивных и анимированных графиков, а также визуализации данных. Она предоставляет разнообразные инструменты для построения графиков и диаграмм, что делает ее неотъемлемым инструментом для визуализации данных и исследования научных результатов. Matplotlib предлагает простой и интуитивно понятный интерфейс для создания графических изображений, позволяя пользователю настраивать внешний вид графиков и добав-

лять различные элементы, такие как легенды и подписи. Эта библиотека поддерживает широкий спектр типов графиков, включая линейные, столбчатые, точечные, гистограммы, круговые диаграммы и многие другие.

4) Gephi - это открытое программное обеспечение, предназначенное для визуализации и анализа сложных сетей и графовых данных. Оно предоставляет средства для исследования и визуализации взаимосвязей между данными в форме графов. Gephi обладает удобным и интуитивно понятным интерфейсом, что позволяет пользователям создавать, настраивать и визуализировать графовые структуры без необходимости программирования. Эта программа способна обрабатывать и визуализировать обширные и сложные сети, что делает ее полезной для анализа социальных сетей, транспортных сетей, биологических сетей и других сложных систем.

5) Plotly - это библиотека, предназначенная для разработки интерактивных графиков и визуализаций данных. Она предоставляет инструменты для создания высококачественных графиков и диаграмм, которые легко интегрируются в веб-приложения и дашборды. Основной фокус Plotly направлен на создание интерактивных графиков, что обеспечивает возможность взаимодействия пользователей с данными на графике. Это включает в себя функции, такие как приближение, перемещение, выделение и другие интерактивные возможности. Библиотека поддерживает различные типы графиков, включая линейные, столбчатые, точечные, гистограммы, круговые диаграммы, тепловые карты и многие другие.

III. Обработка данных и анализ наблюдаемых временных рядов

В ряде исследований [6 - 9] было показано, что наблюдаемые на практике временные ряды социальных процессов обладают фрактальностью, а системы, динамику которых они описывают обладают памятью и проявляют самоорганизацию. Если, например, проанализировать зависимость матожидания и дисперсии амплитуд изменения уровней временных рядов от интервала времени расчета этих амплитуд, то наблюдаются сложные зависимости. Например, их дисперсия зависит от размера «скользящего» окна, как корень дробной степени, что существенно отличается, например, от нормального закона распределения.

Для сбора данных и проведения исследований был выбран информационный ресурс «РИА Новости», что связано с его популярностью в российском обществе, он занимают 1 место среди российских медиаресурсов по версии «br-analytics». (https://br-analytics.ru/mediatrends/media/?period=202203) и входят в ТОП-3 российских самых цитируемых информационных агентств в СМИ и социальных медиа (https://www.mlg.ru/ratings/media/federal/11110/#internet).

На рисунке 5 представлена часть данных временных рядов эмоционального отношения пользователей портала «РИА Новости» к опубликованным в течении суток новостям, за период в 1460 дней с 01.01.2019 по 31.12.2022. В качестве объектов исследования были выбраны следующие эмоции: «нравится» - ряд номер 1 и «не нравится» - ряд номер 2.

Рис. 5. – Временные ряды эмоционального отношения пользователей портала «РИА Новости» к опубликованным в течении суток новостям.

Для предварительного анализа динамики временных рядов и определения их характеристик может быть использован метод нормированного размаха Хёрста [10], позволяющий определить их фрактальную размерность и классифицировать тип поведения. Можно пронормировать размах величин уровней ряда R стандартным отклонением значений величин уровней ряда S и представить их отношение (нормированный размах) в виде уравнения: $R / S = C\tau H$ $(log[R / S] = H * log[\tau] + log[C])$, где C – некоторая константа, τ – число наблюдений (уровней ряда), составляющих рассматриваемый временной ряд (ВР), а показатель степени H – это так называемый коэффициент или показатель Хёрста. Наличие изломов в зависимости R/S (τ) может свидетельствовать о наличие характерных временных масштабов и/или периодичностей. Величина коэффициента H позволяет дать классификацию временных рядов по характеру их поведения [11]. На рисунке 6 для примера показана зависимость $log[R / S]$ от $log[\tau]$ для временных рядов отношения к новостям по эмоциям: «нравится» - рисунок 6a и «не нравится» - рисунок 6b.

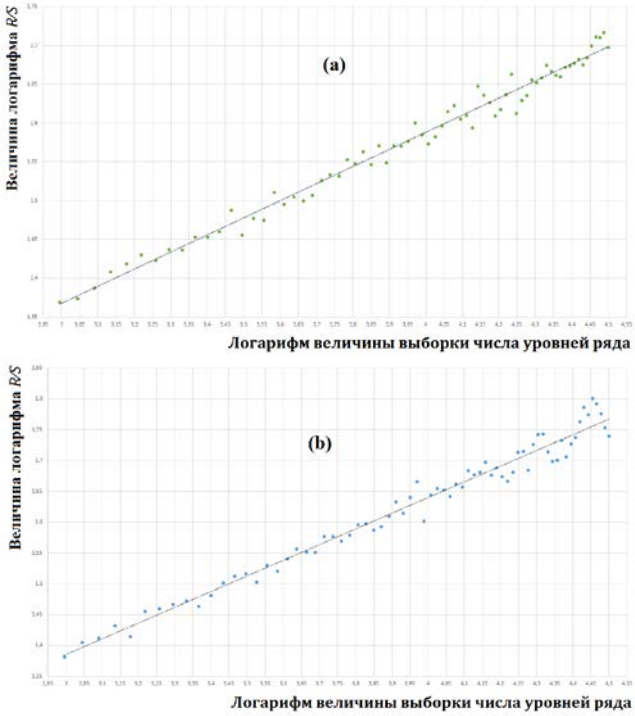


Рис. 6. Определение показателя Хёрста временных рядов эмоционального отношения пользователей портала «РИА Новости» к опубликованным в течении суток новостям за период в 1460 дней с 01.01.2019 по 31.12.2022.

Используя стандартный алгоритм метода Хёрста для расчета зависимости величины логарифма R/S от логарифма величины выборки уровней временного ряда (τ) эмоционального отношения пользователей портала «РИА Новости» к опубликованным новостям можно получить следующее линейные уравнения: для эмоции «нравится» $y = 0{,}22x + 0{,}70$ со значением коэффициента корреляции $R^2 = 0{,}98$; для эмоции «не нравится» $y = 0{,}24x + 0{,}63$ со значением коэффициента корреляции $R^2 = 0{,}98$.

Во всех случаях величина H существенно меньше 0,5 (для «нравится» - 0,22 и «не нравится» - 0,24) и, следовательно, наблюдаемые временные ряды является антиперсистентным (эргодическим). Поскольку величины коэффициента Хёрста существенно отлична от 0,5, то из этого следует, что структура данных рядов обладает фрактальностью, а описываемые им процессы могут иметь краткосрочную память [11].

Для дальнейшей обработки наблюдаемых данных и определения свойств функций распределения (плотности вероятности), например, величины амплитуды изменения эмоционального отношения пользователей сетевых новостных ресурсов, можно использовать следующий алгоритм:

1) Сначала можно задать размер «скользящего окна» (интервал времени между наблюдаемыми значениями эмоционального отношения пользователей сетевых ресурсов), например, одни сутки, двое суток, трое и т.д. и выбрать из временного ряда данные для заданного временного диапазона (размера «скользящего окна»).

2) Далее по выбранным данным можно рассчитать амплитуды изменения величины эмоционального отношения пользователей сетевых новостных ресурсов для различных выбранных временных интервалов (размеров «скользящего окна»).

3) Затем, необходимо отсортировать по возрастанию (от отрицательного к положительному) наборы значений, полученные для каждого из измеряемых интервалов, и для каждого размера «скользящего окна» построить гистограммы плотности распределения амплитуд отклонений величины эмоционального отношения пользователей сетевых новостных ресурсов.

4) Потом можно по полученным гистограммам рассчитать моменты распределения (среднее значение: математическое ожидание, дисперсию, асимметрию, эксцесс) для выбранных интервалов времени расчета амплитуд отклонений эмоционального отношения пользователей сетевых ресурсов (размеров «скользящего окна»).

5) Далее можно построить зависимости для математического ожидания, дисперсии амплитуд отклонений эмоционального отношения пользователей сетевых ресурсов от интервалов времени расчета (размеров «скользящего окна»).

Исследование гистограмм, а также зависимости математического ожидания, дисперсии, асимметрии (третий момент распределения) и эксцесса (четвертый момент распределения) амплитуд отклонений активности пользователей сетевых ресурсов от интервалов времени расчета (размеров «скользящего окна») позволяет определить, являются ли изучаемые временные ряды нестационарными и использовать наблюдаемые характеристики для построения модели их эволюции, что необходимо для прогнозирования их динамики. Наблюдаемое значение математического ожидания, дисперсии, асимметрии и эксцесса можно рассчитать по следующим формулам:

$$\mu(t) = \frac{\sum_{j=1}^{N} x_j(t)}{\sum_{l=1}^{M} n_l}, \quad \sigma^2(t) = \frac{\sum_{j=1}^{N}\{x_j(t)-\mu(t)\}^2}{\sum_{l=1}^{M} n_l}, \quad As(t) = \frac{\sum_{j=1}^{N}\{x_j(t)-\mu(t)\}^3}{\sigma^3 \sum_{l=1}^{M} n_l}, \; Ex(t) = \frac{\sum_{j=1}^{N}\{x_j(t)-\mu(t)\}^4}{\sigma^4 \sum_{l=1}^{M} n_l},$$

а $\sum_{l=1}^{M} n_l$ – вычисляется по числу значений амплитуд n_l.

Исследование зависимости математического ожидания, дисперсии, асимметрии *(третий момент распределения)* и эксцесса *(четвертый момент распределения)* амплитуд отклонений активности пользователей сетевых ресурсов от интервалов времени расчета (размеров «скользящего окна») позволяет определить, являются ли изучаемые временные ряды нестационарными и использовать наблюдаемые характеристики для построения модели их эволюции, что необходимо для прогнозирования их динамики.

Рисунок 7 показывает, что величины математического ожидания амплитуд эмоционального отношения (кривая 1 - «нравится» и кривая 2 - «не нравится») пользователей зависят от интервала времени расчета этих амплитуд («скользящего окна») сложным образом, что указывает на то, что исследуемые временные ряды являются не стационарными. Например, при выполнении нормального закона распределения для стационарного временного ряда $\rho(x,t) = \frac{1}{2\sqrt{\pi \mathrm{D}t}} \cdot e^{-\frac{x^2}{4 \cdot D \cdot t}}$, величина математического ожидания $\mu(t)$ должна была бы быть равна нулю:

$$\mu(t) = \int_{-\infty}^{+\infty} x \cdot \rho(x,t)dx = \frac{1}{2\sqrt{\pi \mathrm{D}t}} \int_{-\infty}^{+\infty} x \cdot e^{-\frac{x^2}{4 \cdot D \cdot t}} dx = 0$$

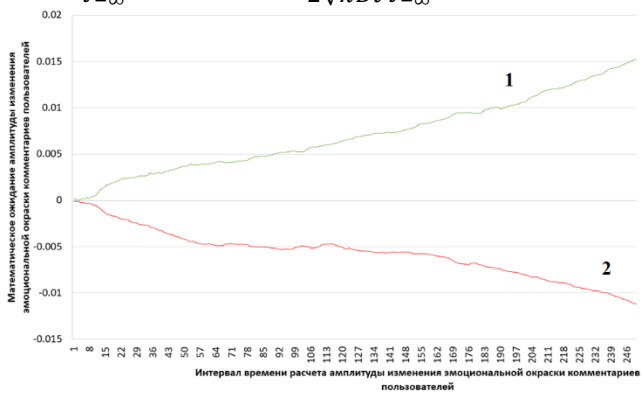


Рис. 7. Зависимость величины математического ожидания амплитуд изменения уровней временных рядов эмоционального отношения пользователей портала «РИА Новости» к опубликованным в течении суток новостям за период в 1460 дней с 01.01.2019 по 31.12.2022.

Где x - величина амплитуды, а t - интервал времени её расчета (величина «скользящего окна»).

Рисунок 8 показывает, что зависимости величин дисперсии амплитуд эмоционального отношения пользователей по комментированию новостей от интервала времени расчета этих амплитуд («скользящего окна») имеют сложный нелинейный характер (кривая 1 - «нравится» и кривая 2 - «не нравится»), и при малых значениях интервала расчета амплитуд их дисперсия не стремится

к нулю.

При выполнении нормального закона, для дисперсии амплитуд должна была бы наблюдаться линейная зависимость от величины «скользящего окна»:

$$\sigma^2(t) = \int_{-\infty}^{+\infty} x^2 \rho(x,t)dx = \frac{1}{2\sqrt{\pi \mathrm{D}t}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{4 \cdot D \cdot t}} dx = 2Dt$$

Отклонение от линейного закона в данном случае также говорит о том, что наблюдаемые временные ряды являются нестационарными. Здесь скорее наблюдается зависимость типа $\sigma^2(t) \sim \sqrt[\beta]{t}$, где β – некоторое дробное число. Такого типа зависимости свидетельствуют о нелокальности во времени и последействии (процесс обладает памятью).

Рис. 8. Зависимость величины дисперсии амплитуд изменения уровней временных рядов эмоционального отношения пользователей портала «РИА Новости» к опубликованным в течении суток новостям за период в 1460 дней с 01.01.2019 по 31.12.2022.

Поведение эксцесса представлено на рисунке 9. Эксцесс характеризует «хвост» распределения. При больших положительных величинах эксцесса функция распределения медленнее убывает при удалении от среднего значения, чем при малых. Для нормального распределения эксцесс равен 3. При величине эксцесса больше трех график плотности распределения будет лежать выше графика нормального распределения, а меньше трех – ниже. В рассмотренных случаях наблюдается, что при небольших величинах интервала времени расчета амплитуд величина эксцесс их распределения имеет большое значение, а при больших «скользящих окнах» медленно убывает к значению равному 1 для эмоции «нравится» - кривая 1; значению равному 3 для эмоции «не нравится» - кривая 2, что может свидетельствовать о нелокальности по состоянию уровней временного ряда и самоорганизации процессов.

Рис. 9. Зависимость величины эксцесса распределения амплитуд изменения уровней временных рядов эмоционального отношения пользователей портала «РИА Новости» к опубликованным в течении суток новостям за период в 1460 дней с 01.01.2019 по 31.12.2022.

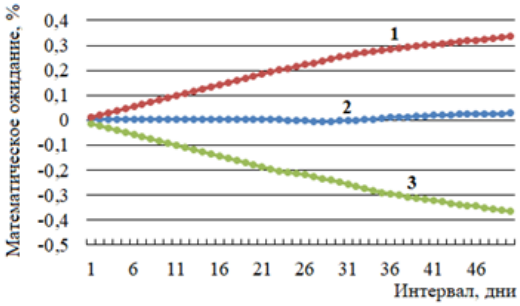
Помимо анализа активности пользователей по комментированию новостей в социальных массмедиа был проведен анализ предпочтений избирателей президентских электоральных кампаний в США в 2012 и 2016 годах (данные взяты с ресурса: http://www.realclearpolitics.com/epolls/).

Например, для электоральной кампании по выборам Президента США в 2012 году были получены следующее линейные уравнения для колебаний предпочтений избирателей: Обама y=0,29x+0,66 со значением коэффициента корреляции R^2=0,97; Ромни y=0,22x+0,99 со значением коэффициента корреляции R^2=0,99.

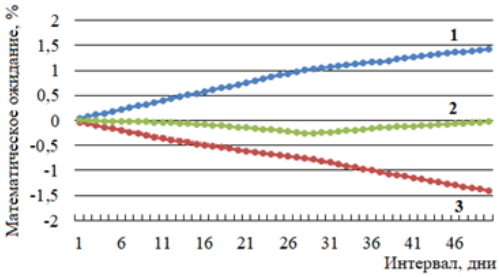
В 2016 году: Трамп y=0,30x+0,66 со значением коэффициента корреляции R^2=0,98; Клинтон (Хиллари) y=0,36x+0,48 со значением коэффициента корреляции R^2=0,979.

Для данных, полученных при наблюдении рассмотренных нами избирательных кампаний, с использованием метода нормированного размаха, были получены схожие значения показателя Хёрста лежащие в диапазоне 0,2-0,35. Для всех рядов наблюдается явление антиперсистентности (показатели Хёрста меньше 0,5).

На рисунке 10 представлены зависимости математических ожиданий амплитуд изменения уровней временных рядов предпочтений избирателей в избирательных кампаниях 2012 и 2016 годов от интервала времени расчета этих амплитуд.



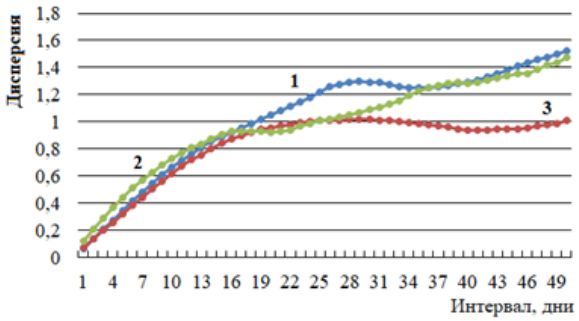
1 – Роумни, 2 – Обама, 3 – не определившиеся



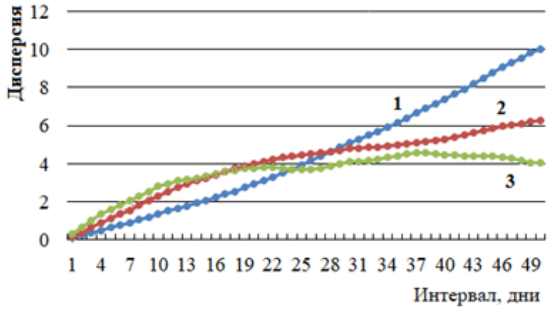
1 – Трамп, 2 – не определившиеся, 3 – Клинтон

Рис. 10. Зависимость величины математического ожидания амплитуд изменения уровней временных рядов предпочтений избирателей в избирательных кампаниях 2012 и 2016 годов от интервала времени расчета этих амплитуд.

Рисунок 10 показывает, что исследованные зависимости близки к линейным или имеют степенной закон с дробным показателем близким к 0,9.



1 – Обама, 2 – не определившиеся, 3 – Роумни

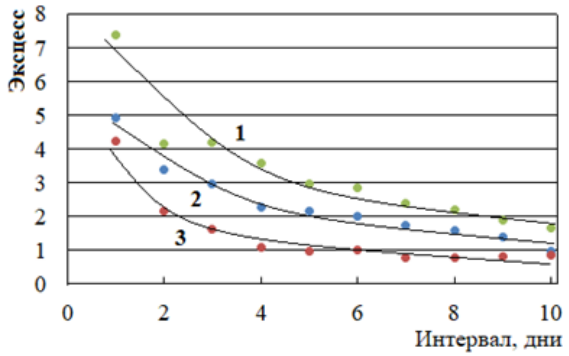


1 – Трамп, 2 – Клинтон, 3 – не определившиеся

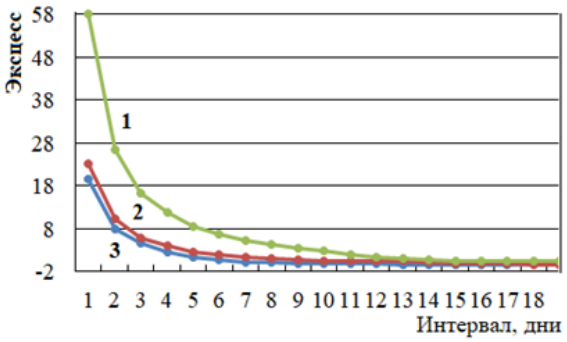
Рис. 11. Зависимость величины дисперсии амплитуд изменения уровней временных рядов предпочтений избирателей в избирательных кампаниях 2012 и 2016 годов от интервала времени расчета этих амплитуд.

На рисунках 11 и 12 показаны зависимости дисперсии и эксцесса амплитуд изменения уровней временных рядов предпочтений избирателей в избирательных кампаниях 2012 и 2016 годов от интервала времени расчета этих амплитуд.

Рисунок 11 показывает, что зависимости дисперсии амплитуд близки к степенному закону с дробным показателем больше 1,5.



1 – не определившиеся, 2 – Обама, 3 – Роумни



1 – не определившиеся, 2 – Клинтон, 3 – Трамп

Рис. 12 Зависимость величины эксцесса амплитуд изменения уровней временных рядов предпочтений избирателей в избирательных кампаниях 2012 и 2016 годов от интервала времени расчета этих амплитуд.

Обработка полученных данных показывает, что:

1) величины математического ожидания амплитуд изменения активности пользователей по комментированию новостей зависят от интервала времени расчета этих амплитуд («скользящего окна»), что указывает на то, что исследуемые временные ряды являются не стационарными, а их параметры не могут быть описаны нормальным законом распределения;

2) величины дисперсии амплитуд изменения активности пользователей по комментированию новостей зависят от интервала времени расчета этих амплитуд («скользящего окна») сложным образом: пропорционально дробной степени из интервала времени их расчета. Дробная зависимость от интервала времени указывает на то, что исследуемые процессы имеют нелокальность во времени - t (обладают последействием или памятью);

3) исследования эксцесса распределения амплитуд показывают наличие так называемых «тяжелых хвостов», существенно больше, чем у нормального распределения (для него эксцесс равен 3). При больших положительных величинах эксцесса функция распределения медленнее убывает при удалении от среднего значения, чем при малых. При величине эксцесса больше трех график плотности распределения будет лежать выше графика нормального распределения, а меньше трех – ниже. Это указывает на то, что рассматриваемые процессы обладают не только нелокальностью по времени – t, но и нелокальностью по состоянию – обозначим его x. Отметим, что данные на рисунках 5 и 8 отложены от уровня эксцесса нормального распределения.

IV. Заключение и выводы

Динамика поведения сложных систем имеет очень сложный характер, включая наличие памяти и возможность самоорганизации. Необходимо отметить, что одновременно могут наблюдаться различные по природе процессы, движущие причины которых могут иметь скрытый характер.

Проведенные исследования сложных социальных процессов, например, электоральных кампаний и активности пользователей в социальных сетях показывают, что наблюдаемые на практике временные ряды обладают фрактальностью, а системы, динамику которых они

описывают обладают памятью и проявляют самоорганизацию. Если, например, проанализировать зависимость матожидания и дисперсии амплитуд изменения уровней временных рядов активности пользователей социальных сетей и сетевых массмедиа от интервала времени расчета этих амплитуд ("скользящего окна"), то наблюдаются сложные зависимости.

Например, для математического ожидания наблюдается корневая зависимость дробной степени, для дисперсии - степенной закон с дробным показателем больше 1,5. Дробные зависимости от времени указывают на наличие нелокальности по этой переменной.

Исследования эксцесса показывает наличие так называемых «тяжелых хвостов», его величина существенно больше, чем у нормального распределения. Такое поведение эксцесса указывает на наличие нелокальности по состояниям уровней временных рядов.

Полученные результаты указывают на то, что рассматриваемые процесс имеют память и возможность самоорганизации, а их временные ряды обладают нестационарностью, а также нелокальностью, как по времени, так и состоянию.

Нелокальные процессы характеризуются тем, что переход в данное состояние системы (или процесса) зависит не только от локальных характеристик процесса или поведения системы в окрестности рассматриваемой точки в данный момент времени, но и от принимаемых значений на всем исследуемом интервале в предыдущие моменты времени, т.е. зависит глобально от распределения по всем состояниям и от предыстории процесса (памяти). Нелокальность по времени влияет на плотность вероятности распределения состояний в начальное время, что может приводить к самоорганизации, а нелокальность по состоянию влияет на асимптотическое поведение плотность вероятности обнаружить некоторое состояние x в момент времени t при больших временах.

Для разработки специализированного программного обеспечения для сбора, обработки и хранения данных социодинамических процессов, учитывая их специфику можно применять конвейерный принцип с реализацией микросервисной архитектуры и использовать графовые базы данных.

Благодарности

Работа выполнена при финансовой поддержке Российского научного фонда (РНФ), грант № 23-21-00153 «Анализ и моделирование динамики нестационарных временных рядов фрактальных процессов с реализацией памяти (последействия) и самоорганизацией на основе использования дифференциальных уравнений с дробными производными».

Библиография

[1] Митра Р., Надареишвили И. Микросервисы. От архитектуры до релиза. O'Reilly, 2024.
[2] Маккинли У. Python и анализ данных. Пер. с англ. Слинкин А.А. М.: ДМК Пресс, 2015.
[3] Хеллман Д. Стандартная библиотека Python 3. Справочник с примерами. Диалектика. 2-е изд. 2019.
[4] Персиваль Г. Python. Разработка на основе тестирования. 2018.
[5] Робинсон Я., Эифрем Э., Вебер Д. Графовые базы данных. Новые возможности для работы. ДМК-Пресс, 2016.
[6] Zhukov D., Khvatova T., Zaltsman A. Stochastic Dynamics of Influence Expansion in Social Networks and Managing Users' Transitions from One State to Another // Proceedings of the 11 th European Conference on Information Systems Management, ECISM 2017, The University of Genoa, Italy, 14 -15 September, 2017, pp. 322 – 329.
[7] Sigov A.S., Zhukov D.O., Khvatova T.Yu., Andrianova E.G. A Model of Forecasting of Information Events on the Basis of the Solution of a Boundary Value Problem for Systems with Memory and Self-Organization // Journal of Communications Technology and Electronics. 2018, Vol. 18, №2, pp. 106 – 117. DOI: 10.1134/S1064226918120227
[8] Zhukov D., Khvatova T., Istratov L. A stochastic dynamics model for shaping stock indexes using self-organization processes, memory and oscillations // Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics, ECIAIR 2019, Oxford, UK, 31 October–1 November 2019, pp. 390 – 401.
[9] Zhukov D., Khvatova T., Istratov L. Analysis of non-stationary time series based on modelling stochastic dynamics considering self-organization, memory and oscillations // ITISE 2019 International Conference on Time Series and Forecasting. Proceedings of Papers, 25-27 September 2019, Granada (Spain), Vol. 1, pp. 244 – 254.
[10] Hurst H.E. Long - term storage capacity of reservoirs. // Transactions of American Society of Civil Engineers. 1951. Vol. 116. P. 770.
[11] Mandelbrot B. B. The Fractal Geometry of Nature. W. H. Freeman, Sun Francisco, 1982.

Об авторах:
Отраднов Константин Константинович, старший преподаватель, институт радиоэлектроники и информатики, МИРЭА - Российский технологический университет
Калинин Владимир Николаевич, педагог ДО, кафедра прикладной информатики и интеллектуальных систем в гуманитарной сфере, ФГАОУ ВО Российский Университет Дружбы Народов им. Патриса Лумумбы
Лесько Сергей Александрович, профессор, д.т.н., институт кибербезопасности и цифровых технологий, МИРЭА - Российский технологический университет
Платонова Ирина Вячеславовна, доцент, к.ф-м.н. физический факультет, МГУ им. М.В. Ломоносова

Organization of data collection and processing of sociodynamic processes with possible self-organization and memory availability and analysis of observed characteristics of their time series

K.K. Otradnov, V.N. Kalinin, A.S. Lesko, I.V. Platonova

***Abstract* - The article discusses the development of specialized software for collecting, processing and storing data from sociodynamic processes (changing the emotional color of user comments on published news in online media, and electoral campaigns of the US presidential elections in 2012 and 2016). It has been shown that for its creation it is possible to use the pipeline principle with the implementation of a microservice architecture, and for storing data, taking into account their specifics and origin, the use of graph databases is preferable.**

Based on the collected data, time series of observed processes were obtained. Their R/S analysis showed that they had antipersistance. A study of the dependence of the mathematical expectation, variance and excess of the amplitudes of deviations of series levels from the dimensions of the amplitude calculation time interval ("sliding window") showed that for the mathematical expectation there is a root dependence of fractional degree; for dispersion - the power law with a fractional indicator greater than 1.5; and the behavior of the excess shows the presence of the so-called "heavy tails," its magnitude is significantly greater than that of the normal distribution.

The obtained results indicate that the time series of the processes under consideration have unsteady, non-locality, both in time (have memory) and state (show self-organization).

***Keywords:* time series, self-organization, memory availability, unsteadiness, time series fractality, sociodynamic processes, graph databases, microservice software architecture.**

REFERENCES

[1] Mitra R, Nadareishvili I. Microservices. From architecture to release. O'Reilly, 2024. 336 p. [rus]
[2] McKinley W. Python and data analysis. Per. with English. Slinkin A.A. M.: DMK Press, 2015. [rus].
[3] Hellman D. Python Standard Library 3. Reference book with examples. Dialectics. 2nd ed. 2019. [rus]
[4] Percival G. Python. Development based on testing. 2018. [rus]
[5] Robinson Yang, Eifrem Emil, Weber Jim, Graph Databases. New opportunities for work, DMK-Press, 2016 ISBN: 978-5-97060-201-0.
[6] Zhukov D., Khvatova T., Zaltsman A. Stochastic Dynamics of Influence Expansion in Social Networks and Managing Users' Transitions from One State to Another. In Proceedings of the 11 th European Conference on Information Systems Management, ECISM 2017, The University of Genoa, Italy, 14 -15 September, 2017, pp. 322 – 329.
[7] Sigov A.S., Zhukov D.O., Khvatova T.Yu., Andrianova E.G. A Model of Forecasting of Information Events on the Basis of the Solution of a Boundary Value Problem for Systems with Memory and Self-Organization. Journal of Communications Technology and Electronics, vol. 18, no 2, pp. 106–117, 2018
[8] Zhukov D., Khvatova T., Istratov L. A stochastic dynamics model for shaping stock indexes using self-organization processes, memory and oscillations. In Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics, ECIAIR 2019, Oxford, UK, 31 October–1 November 2019, pp. 390 – 401, 2019.
[9] Zhukov D., Khvatova T., Istratov L. Analysis of non-stationary time series based on modelling stochastic dynamics considering self-organization, memory and oscillations. In ITISE 2019 International Conference on Time Series and Forecasting. Proceedings of Papers, 25-27 September 2019, Granada (Spain), Vol. 1, pp. 244 – 254, 2019.
[10] Hurst H.E. Long - term storage capacity of reservoirs. Transactions of American Society of Civil Engineers, vol. 116, p. 770, 1951.
[11] Mandelbrot B. B. The Fractal Geometry of Nature. W. H. Freeman, Sun Francisco, 1982.

About the authors:
Konstantin K. Otradnov, senior lecturer, MIREA - Russian Technological University
Vladimir N. Kalinin, teacher of further education, Patrice Lumumba Peoples' Friendship University of Russia
Sergey A. Lesko, Professor, MIREA - Russian Technological University
Irina V. Platonova, associate professor, Faculty of Physics, M.V. Lomonosov Moscow State University