УДК 338.26 DOI 10.21685/2227-8486-2020-1-9

ПРОАКТИВНЫЙ МОНИТОРИНГ СОБЫТИЙ НА ОСНОВЕ ПРЕДИКТИВНОГО АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

И. Н. Колесников, А. Г. Финогеев

PROACTIVE EVENT MONITORING BASED ON PREDICTIVE TIME SERIES ANALYSIS

I. N. Kolesnikov, A. G. Finogeev

Аннотация. *Предмет и цель работы.* Рассматриваются методы проактивного мониторинга дорожно-транспортной инфраструктуры на основе сбора и обработки больших данных о событиях на контролируемых участках дорог. В процессе мониторинга выполняется консолидация разнородных данных из множества открытых источников и извлечение характеристик событий с целью представления их динамики в виде временных рядов для предиктивного моделирования, анализа и оценки рисков возникновения и развития нештатных и аварийных происшествий с учетом влияния внешних факторов. *Методы.* Для достижения цели и задач исследований использовались методы сбора, консолидации и обработки больших данных (Big Data) с целью идентификации, классификации и кластеризации событий, сравнительного анализа спектров временных рядов характеристик и факторов влияния. Большие данные о дорожно-транспортных происшествиях поступают с фоторадарных комплексов фото- и видеофиксации правонарушений и транспортных средств, а также с открытых источников в сети Интернет и мобильных средств связи участников событий. *Результаты и выводы.* Рассмотрены методы и инструментальные средства для сбора и анализа больших данных о событиях для предиктивной аналитики, такие как Hadoop, MapReduce и NoSQL. Предложены способы сбора и консолидации разнородных данных для синтеза временных рядов, анализа и предиктивного моделирования. Приведен алгоритм обработки текстовых сообщений для перехода к векторной модели слов и машинного обучения системы прогнозирования на основе спектров временных рядов событий. Результаты мониторинга необходимы для превентивного реагирования на возможные негативные события и происшествия в дорожной среде для снижения аварийных ситуаций и оказания экстренной помощи. Рассматриваемая система проактивного мониторинга основывается на методах сбора и анализа больших данных и интеллектуального анализа данных, полученных из различных информационных источников. В системе предлагается использование прогностического моделирования рисков возникновения и развития событий за счет построения спектра временных рядов на основе векторного представления слов. Данные методы позволят системе проактивного мониторинга решать задачи оценки и рисков возникновения нештатных и аварийных событий на дорогах с учетом влияния внешних факторов, контролировать участки дорог и транспортные средства, отслеживать локации мест аварий и прочих деструктивных событий на дорогах.

Ключевые слова: большие данные, интеллектуальный анализ данных, временные ряды, предиктивная аналитика, прогностическое моделирование, Hadoop, MapReduce, NoSQL**.**

Abstract. *Subject and goals.* The article discusses methods of proactive monitoring of road transport infrastructure based on the collection and processing of big data about events on controlled sections of roads. In the process of monitoring, heterogeneous data is consolidated from many open sources and the characteristics of events are extracted in or-

der to present their dynamics in the form of time series for analysis and prognostic modeling of the risks of emergencies and emergencies taking into account the influence of external factors. *Methods*. To achieve the goal and objectives of the research, methods of collecting, consolidating and Big Data processing, identification, classification and clustering of events, a comparative analysis of the spectra of time series of their characteristics and influence factors are used. Big data on traffic accidents and incidents comes from distributed photo-radar photo and video recording complexes of offenses and vehicles, as well as from open sources on the Internet and from mobile means of communication of witnesses and participants in events. *Results and conclusions*. Methods and tools were selected for collecting and analyzing big data about events in a proactive monitoring system, such as Hadoop, MapReduce and NoSQLe. The main methods for collecting and consolidating heterogeneous data for intelligent analysis and predictive modeling are considered. An algorithm for collecting and preparing textual data using the vector representation of words for machine learning a prediction system based on spectra of time series of events is presented. The results of proactive monitoring are necessary for a proactive response to possible negative events and incidents in the road environment to reduce emergency situations and provide emergency assistance. The proactive monitoring system under consideration is based on the methods of collecting and analyzing big data, and the intellectual analysis of data obtained from various information sources. The system proposes the use of prognostic modeling of risks of occurrence and development of events by constructing a spectrum of time series based on a vector representation of words. These methods will allow the proactive monitoring system to solve the problems of assessing and the risks of emergency and emergency events on the roads, taking into account the influence of external factors, to control road sections and vehicles, to track the locations of accident sites and other destructive events on the roads.

Keywords: big data, data mining, time series, predictive analytics, predictive modeling, Hadoop, MapReduce, NoSQL.

Введение

Современным трендом в управлении сложными системами является использование элементов искусственного интеллекта, включая методы интеллектуального анализа данных, машинного обучения и предиктивного моделирования. Данные элементы обуславливает переход к проактивной концепции управления на основе обработки больших данных [1]. Концепция позволяет предотвращать риски возникновения и развития аварий и катастроф на основе предиктивного анализа событий и синтеза упреждающих воздействий [2]. Методы проактивного управления широко используются в киберфизических системах, которые являются атрибутом четвертой промышленной революции «Индустрия 4.0» [3, 4]. В интеллектуальных киберфизических системах применяются технологии работы с большими данными, методы машинного обучения и прогнозирования, межмашинного (M2M) взаимодействия в среде Интернет вещей [5]. Важнейшим аспектом проактивного управления является решение прогнозных задач для синтеза превентивных мер предупреждения или минимизации рисков нештатных и аварийных ситуаций [6]. Поэтому основным принципом здесь является применение схемы действий на основе анализа данных о старых событиях (event-driven принцип) для предсказания новых событий, которая представляет собой методику обнаружить–предсказать–решить–действовать («detect–forecast–decide–act»). Типовой алгоритм работы проактивных систем управления включает этапы: а) идентификации и классификации событий, б) иден-

тификации факторов влияния для установления связи с событиями, в) анализа чувствительности событий к выявленным факторам, г) выбора прогностической модели для оценки рисков негативных событий, д) прогностического моделирования событий, е) сравнительного анализа результатов прогноза с аналогами, ж) выбора решений для минимизации рисков.

Результатом внедрения проактивных технологий является перевод лиц, принимающих решения, из подсистемы управления в подсистему конфигурирования и контроля с передачей им функций настройки, контроля и диагностики работы средств мониторинга [7]. Фактически субъект управления переходит на уровень координации и супервизорного контроля процесса оперативного мониторинга в сложных территориально-распределенных системах.

Методы сбора и обработки больших данных для анализа событий

Проведем анализ существующих подходов к мониторингу дорожно-транспортной инфраструктуры. При мониторинге дорожно-транспортной инфраструктуры используются традиционные методы и методы дистанционного мониторинга. К традиционным методам относятся:

– *регистрация проходящего автотранспорта людьми – операторами*. При этом человек должен регистрировать параметры безопасности дорожного движения и заносить свои наблюдения в полевой журнал. Для полного охвата больших территорий необходима одновременная работа большого числа операторов. На точность влияют квалификация оператора, усталость, невнимательность. Это приводит к ошибкам и некоторому искажению реальной ситуации;

– *опрос работников автотранспортных предприятий.* Этот метод позволяет достаточно быстро оценить наиболее напряженные участки дорожной сети, но имеет недостатки в точности.

К методам дистанционного мониторинга относятся:

– *регистрация проходящего автотранспорта с помощью различных датчиков*. К достоинствам можно отнести оперативность получаемой информации, а также точность измерений, которая повышается за счет автоматической регистрации, автоматической передачи данных в центр обработки, и минимизации влияния «человеческого фактора». К недостаткам можно отнести то, что единовременные затраты на установку датчиков, развитие инфраструктуры связи с центром обработки довольно высоки;

– *проактивный мониторинг дорожно-транспортной инфраструктуры.* Помимо данных с различной информацией, получаемой с различных датчиков, информация о дорожно-транспортной ситуации поступает из большого числа различных источников. Для обработки большого количества данных и использования их для прогнозирования будущей ситуации используются элементы искусственного интеллекта и методы интеллектуального анализа данных. К недостаткам можно отнести существенное усложнение системы и возможные ошибки прогнозов при неправильно выбранных моделях прогнозирования. К достоинствам – полная автоматизация процесса и возможность получения прогнозов дорожно-транспортной ситуации, позволяющая предотвращать различные инциденты.

Рассмотрим современный подход, которым является проактивный мониторинг объектов, процессов и событий в киберфизических системах с множеством пространственно-распределенных объектов. Примером систем являются интеллектуальные энергетические сети (Smart Energy Grid) [8], интеллектуальные производственные системы (Smart Manufacturing), системы интеллектуального управления городским и дорожным освещением (Smart Light), системы «умный» город (Smart City) [9], интеллектуальные системы мониторинга дорожно-транспортной инфраструктуры (Smart&Safe Road) [10]. В последнем случае объектами мониторинга являются участки дорог и компоненты дорожной инфраструктуры (знаки, остановки, переходы, устройства регулировки, устройства фото- и видеофиксации, транспортные средства, участники дорожного движения, придорожные объекты, системы дорожного освещения и т.д.) [11]. Сложность мониторинга связана с протяженностью и разбросом контролируемых объектов на большой территории. Поэтому система проактивного мониторинга должна иметь комплекс инструментальных средств сбора и анализа больших данных о событиях в дорожной среде для предиктивного анализа временных рядов событий и прогнозной оценки рисков происшествий. Примером событий являются: дорожно-транспортные происшествия, нарушения правил дорожного движения, пробки и заторы, поведение участников движения, нарушение дорожного покрытия, ремонтные работы, сбои в работе дорожного регулирующего оборудования, проблемы с уличным освещением, проблемы с дорожными знаками, ухудшение метеоусловий и т.п. Дорожно-транспортные происшествия являются наиболее массовыми негативными событиями [12]. Подобные технологии уже используются, например, для синхронизации работы светофоров и регулировки транспортных потоков на перекрестках [7], для оптимизации пропускной способности магистралей в зависимости от плотности трафика и пробок [13], для выявления скрытых закономерностей в данных о произошедших инцидентах с целью прогноза новых происшествий [14].

В данной статье решается задача сбора больших данных о событиях, представления динамики характеристик событий и внешних факторах в виде временных рядов для сравнительного анализа и прогностического моделирования рисков возникновения и развития новых инцидентов на контролируемых участках дорог. Технологии интеллектуального анализа позволяют выявлять скрытые закономерности в множестве данных и связать их с влиянием разных факторов для прогнозирования вероятности появления и развития негативных событий. В ходе предиктивного анализа временных рядов идентифицируется и прогнозируется вероятность неблагоприятного развития событий, а методы машинного обучения позволяют находить механизмы превентивного реагирования на инциденты.

Как известно, термин «большие данные» (Big Data) означает совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объемов и значительного многообразия для получения результатов, воспринимаемых человеком и эффективных в условиях их непрерывного прироста [15]. К основным свойствам больших данных чаще всего относят [16]:

1. Сверхбольшой объем информации, который генерируется, собирается и хранится от множества источников данных разных типов.

2. Высокая скорость генерации и обработки данных в режиме реального времени, которая позволяет принимать наиболее адекватные решения для конкретной ситуации с учетом воздействий на процесс управления.

3. Многообразие информации, генерируемой из множества источников в различных форматах, с разной структурой и размером, относящимся к любым категориям и аспектам управления, что требует необходимости предварительной классификации, стратификации, кластеризации, консолидации и т.д.

Исследования, проведенные в 2018 г., показали, что более 55 % компаний в мире готовы к внедрению в бизнес-процессы инструментальных средств по работе с большими данными [17]. Прежде всего такие технологии и системы могут быть использованы в финансовой сфере, в государственном секторе, в медицинской индустрии, на предприятиях IT-области и в интернет-компаниях.

Для работы с большими данными необходимы способы хранения и обработки информации, отличные от традиционных OLAP систем [18]. На первый план выходят распределенные системы обработки и хранения больших данных класса Business Intelligence. Их появление связано с развитием технологических возможностей для сбора, обработки и хранения огромных массивов данных непосредственно на сетевых узлах. Выбор средств обработки больших данных обусловлен характеристиками, которые включают объем данных, структурированность (таблицы реляционных баз данных), вид мультимедийных данных, скорость генерации, изменчивость, достоверность, актуальность, сложность и т.п. В табл. 1 приведены различия между подходами к аналитической обработке обычных и больших данных.

Таблица 1

Традиционная OLAP аналитика и аналитика больших данных

OLAP аналитика	Big data аналитика
Постепенный анализ небольших пакетов данных	Обработка сразу всего массива доступных данных
Редакция и сортировка данных перед обработкой	Данные обрабатываются в их исходном виде
Старт с гипотезы и ее тестирования относительно данных	Поиск корреляций по всем данным до получения искомой информации
Данные собираются, обрабатываются, хранятся и лишь затем анализируются	Анализ и обработка больших данных в реальном времени по мере поступления

Платформа для работы с большими данными должна обеспечивать [19]:

– горизонтальную масштабируемость вычислительной системы (реализуется через возможность модульного расширения);

– отказоустойчивость (обеспечивается путем резервирования);

– адаптивность (реализуется за счет настройки на данные из конкретной предметной области);

– локализацию (выполняется через внедрение распределенной обработки данных в местах их сбора).

Распространенным подходом на текущий момент к распределенной обработке больших данных являются схема и метод MapReduce (рис. 1) от компании Google, где в качестве модели хранилища используется нереляционная

(NoSQL) модель на платформе Hadoop [20]. Преимуществами NoSQL модели являются использование различных типов хранилищ, интегрируемость, масштабируемость.

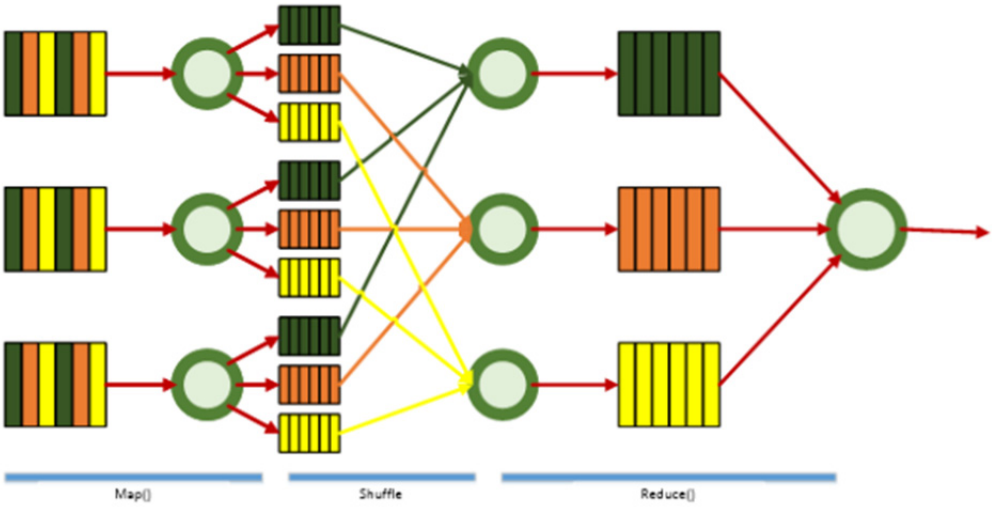

Рис. 1. Схема работы с данными MapReduce[1]

Согласно модели MapReduce обработка данных выполняется в 3 этапа:

1. На первом этапе (Map) реализуется обработка и фильтрация данных с помощью функции, определяемой пользователем. Функция аналогична операции Map в программировании и применяется к входному потоку данных.

2. На втором этапе (Shuffle) результаты выполнения пользовательской функции раскладываются по корзинам, где корзина соответствует ключу для результата операции Map и служит входными данными для этапа Reduce.

3. На третьем этапе (Reduce) каждая корзина является источником данных для функции Reduce, которая задается пользователем для вычисления окончательного результата.

Платформа Hadoop для хранения и управления данными включает:

– распределенную файловую систему Hadoop Distributed File System (HDFS) для иерархического хранения неструктурированных данных;

– прикладной программный интерфейс (API) для создания приложений обработки больших данных в серверном кластере;

– систему управления данными Apache Hadoop YARN.

Платформа позволяет снизить время обработки неструктурированной информации и стоимость оборудования за счет использования типовых модулей расширения, что, в свою очередь, решает проблемы горизонтального масштабирования и отказоустойчивости.

[1] Introduction to MapReduce. – URL: https://developerzen.com/introduction-to-mapreduce-for-net-developers-1030e070698a.

В качестве источников больших данных в первую очередь следует отметить открытые источники в сети Интернет, такие как новостные сайты, социальные сети, мессенджеры, системы видео- и фотохостинга и т.п., где ежедневно выкладывается огромный объем мультимедийной информации. Другими источниками являются доступные данные из корпоративных информационных систем предприятий, а именно, транзакционная деловая информация, архивы, базы данных и т.п. Также к источникам следует отнести множество киберфизических объектов в сети Интернет вещей, включающих измерительные приборы, датчики, камеры видеонаблюдения, SCADA системы, мобильные средства связи. В дорожно-транспортной инфраструктуре в качестве источников больших данных выступают фоторадарные комплексы фото- и видеофиксации дорожно-транспортных происшествий (ДТП) (рис. 2), камеры видеонаблюдения, средства безопасности дорожного движения, средства навигации, интеллектуальные транспортные системы, мобильные средства связи участников дорожного движения. В разрабатываемой системе мониторинга к основным средствам сбора относятся фоторадарные комплексы, открытые источники в сети Интернет и мобильные системы свидетелей событий. Все процедуры сбора данных можно разделить:

1. Сбор данных с систем слежения за элементами дорожно-транспортной инфраструктуры (измерительных устройств, фоторадарных комплексов, дорожных камер видеонаблюдения и т.п.). Процедура представляет собой процесс автоматизированного получения сенсорных данных посредством опроса фоторадарных комплексов, камер видеонаблюдения и сенсорных узлов, расположенных на автомагистралях или вблизи них.

2. Сбор данных с внешних систем (метеорологических станций, систем навигации, интеллектуальных транспортных систем и т.п.). Примером являются данные с навигационного оборудования транспортных средств и систем реагирования (навигаторы, транспондеры, GPS/ГЛОНАСС модули, системы ЭРА-ГЛОНАСС, eCall, ЭВАК, E911 и т.п. Данная процедура позволяет выявлять факторы, влияющие на риски ДТП в зависимости от текущей обстановки на дорогах, погодных условий, состояния дорожного покрытия, времени суток, дорожного трафика и т.д.

3. Сбор данных с мобильных средств связи участников дорожного движения и сторонних наблюдателей. Такая информация часто представляет собой фотографии или видеоролики, которые затрудняют извлечение данных. Однако если они сопровождаются текстовыми сообщениями, то это облегчает задачу. Проблемой является наличие множества источников с данными об одном событии, что требует очистки недостоверной информации и исключения дубликатов.

4. Сбор данных, размещенных в источниках сети Интернет (социальные сети, форумы, мессенджеры, web-ресурсы). Процедура включает извлечение и анализ текстовых сообщений, поиск и распознавание изображений на фотографиях и видеокадрах, регистрацию событий по извлеченной информации, сопоставление данных с различных ресурсов для удаления дубликатов и подтверждения достоверности событий и т.д.

5. Сбор и обработка данных со специализированных сервисов типа Yandex-карты, Google карты, Yandex-навигатор, Navitel и т.п.

Рис. 2. Фоторадарные комплексы фото- и видеофиксации ДТП

Вся информация, полученная в рамках процедур, представляет набор больших разнородных данных, включая телеметрические (сенсорные) данные, текстовые сообщения, фотографии, кадры из видеороликов, данные мобильных приложений и онлайн сервисов. Поэтому перед анализом данных необходима очистка зашумленной информации, унификация, структурирование, консолидация данных. Консолидация данных включает: поиск ассоциаций и корреляций, нормализацию, исключение дубликатов, интеграцию текстовой и графической информации о событии, оценку актуальности и достоверности и т.п.

Методы представления данных о событиях в виде временных рядов

Для прогностического моделирования рисков негативного развития событий предлагается информацию о событиях из разных источников, а также динамику изменения факторов влияния на события представить в виде временных рядов [20, 21]. Сравнение временных рядов характеристик событий и факторов позволяет установить зависимости появления событий от влияния конкретных факторов. Временной ряд для системы мониторинга представляет собой последовательность числовых значений величин, изображений и текстов, для которых известен момент времени, в который они были получены [22]. Для предиктивного анализа событий сначала требуется определить функциональную зависимость, адекватно описывающую его временной ряд. Затем необходимо выявить фактические изменения характеристик событий во временном ряду, влияющие на формирование прогноза [23]. Таким образом можно определить интервалы с аномальными отклонениями числа ДТП от среднего значения для установления корреляции с изменениями значений факторов в соответствующих временных рядах. Сравнительный анализ временных рядов появления и развития инцидентов и временных рядов изменения факторов влияния позволяет установить корреляции между ними. Для оценки чувствительности событий к влиянию факторов применяется метод многофакторного дисперсионного анализа. Таким образом, определяются наиболее сильные факторы влияния, которые выбираются в качестве входных данных для модели прогнозирования динамики инцидентов. Методика обработки текстовых данных для представления в виде временного ряда включает следующие этапы [24]:

1. Регистрация событий на контролируемом участке дороги посредством автоматизированных средств фото- и видеофиксации либо посредством сообщений, фото- и видеоматериалов в источниках сети Интернет.

2. Фиксация информации о событии с извлечением метаданных с текстовых сообщений, фотографий и видеороликов, которые содержат временные и геопространственные метки событий для привязки описаний к временной шкале и к картографической основе.

3. Фиксация информации о параметрах факторов влияния на события (температуры, давления, влажности, скорости ветра, гололеда, времени суток, освещенности, состояния дороги и обочины, плотности трафика, количества транспортных средств и пешеходов, скоростного режима, наличия разметки, количества полос движения, наличия препятствий движению и пробок, наличия средств регулирования движения, состояния участников движения и т.д.).

4. Синтез векторной модели слов для описания события, его идентификации и классификации события по типу, подтверждение факта его появления и достоверности с помощью фото- и видеоматериалов, полученных с разных источников в фиксированный интервал времени и в конкретном местонахождении информационных источников.

5. Поиск сообщений о событии из множества источников по временной и геопространственной метке, а также по ключевым словам. Синтез множества векторных моделей текстовых сообщений о событии для формирования временного ряда векторов слов с момента появления сообщений, фиксация области распространения сообщений по карте местности и синтез динамической графовой модели распространения сообщений в социальных сетях.

6. Выбор и регистрация характеристик описания события и возможных факторов влияния на его появление и развитие события в зависимости от его типа с последующей фиксацией их значений в моменты времени для синтеза спектра временных рядов (паттерна события) для отобранных характеристик и факторов.

7. Поиск информации об аналогичных событиях в другие временные интервалы на данном участке и/или на других участках магистрали. Сравнительный анализ ранее сохраненных паттернов временных рядов с описаниями аналогов с временными рядами характеристик данного события и факторов влияния для расчета и подтверждения средней статистической вероятности корреляций между показателями временных рядов.

8. Кластеризация паттернов временных рядов схожих событий и векторных моделей текстовых сообщений о них в пространствах признаков и факторов влияния. Определение центров кластеризации и фиксация признаков события и параметров влияния на его появление. Определение наилучших соответствий временных рядов по каждому событию по формуле скользящего среднего для прогнозирования значений факторов, указывающих риски повторения аналогичных событий.

В результате имеем ряд средних прогнозных значений факторов влияния по типам и кластерам событий. Для оценки степени влияния каждого фактора они ранжируются и определяется степень важности события. Показатель важности события с набором характеристик и значений факторов влияния, изменяющихся во времени, представляется в виде спектра временных рядов (паттерна события). Комплекс паттернов событий используется для последующего сравнительного анализа (бенчмаркинга) временных рядов и представляет собой элемент обучения системы проактивного мониторинга.

Результаты представления текстовых данных о событиях в виде векторов слов

Для сокращения сложности решения задач анализа разных описаний событий, извлекаемых из множества источников, выполняется переход к представлению сообщений в виде векторов слов. Вектор слов – это численное представление группы семантически связанных слов или фраз. Метод перехода к векторной модели готовит описания событий к извлечению метаданных. Проблемой является то, что информация об одном и том же событии во множестве источников имеет разные форматы данных и неструктурированный вид. Векторное представление является компактным и унифицированным типом данных для хранения и обработки. Оно учитывает контекст и позволяет структурировать данные о событии путем представления в виде системы векторов (рис. 3).

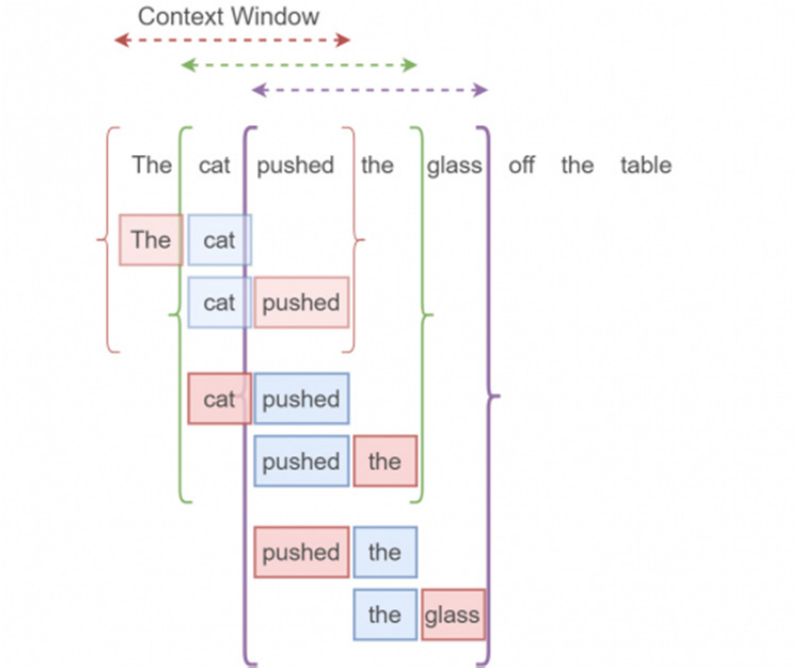

Рис. 3. Пример векторного представления фразы[1]

Для представления текстовых данных в виде множеств векторов слов в системе используется алгоритм ***Word2Vec***, этапы работы которого включают:

1. Синтез кортежей данных в формате [входное слово, выходное слово], где слово представлено в виде двоичного вектора длины *n*, где *i*-е значение кодируется единицей на *i*-й позиции и нулями – на всех остальных (код *one-hot*);
2. Синтез модели обучения, где вход и выход получает *one-hot* вектора;
3. Определение функции потерь, которая предсказывает верное слово для оптимизации модели обучения;

[1] Neurohive. Word2Vec. – URL: https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/

4. Определение качественных характеристик модели после согласования векторных представлений похожих слов, т.е. определение, насколько точно, адекватно и качественно работает модель.

В ряде случаев для синтеза векторных представлений необходимо выполнить оптимизацию алгоритма *Word2Vec*. В аналогичных задачах часто в качестве критерия оптимизации алгоритма выбирается логистическая функция кросс-энтропийных потерь (*softmax cross entropy loss*). Однако использование функции для оптимизации алгоритма *Word2Vec* не является практичным, так она хорошо подходит в основном для решения бинарных задач с двумя результатами [25]. В текстовых сообщениях число слов в фразах может измеряться сотнями, поэтому для вычисления логистической функции *softmax* требуется рассчитать потери в кросс-энтропии по всем выходам. Поэтому будем использовать семплированную логистическую функцию потерь в качестве альтернативы (*sampled softmax loss*). Для ее расчета сначала вычисляется функция перекрестной энтропии между истинным значением контекста для целевого слова и значением предсказанного слова, соответствующего истинному значению контекста. Затем добавляется кросс-энтропийная потеря k отрицательных семплов (целевое слово + слово вне контекста), которые отбираются в соответствии с распределением шума. Далее определяем функцию потерь L следующим образом:

$$L = SigCrEnt\left(Prediction, Correct\,Word\right) + \sum_{1}^{K} E_{noise} SigCrEnt\left(Prediction, Noise\right),$$

где *SigCrEnt* – это ошибка, которую можно определить только на одном выходе. Решение задачи возможно тогда, когда словарь описания событий становится большим, как в нашем случае. Пример разбора модели представлен на рис. 4.

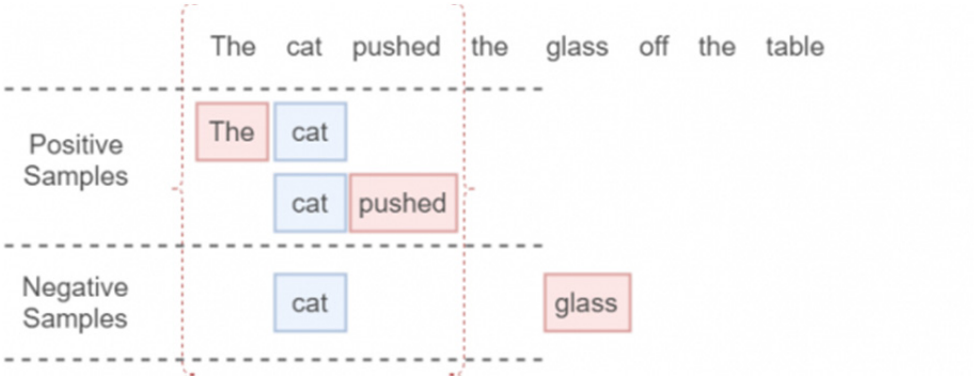

Рис. 4. Пример разбора фразы алгоритмом *Word2Vec*[1]

Заключение

Целью статьи является описание процесса проактивного мониторинга на основе сбора и подготовки больших данных о событиях с различных информа-

[1] Neurohive. Word2Vec. – URL: https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/

ционных источников для последующего описания их динамики в виде векторного представления слов и спектров временных рядов и прогностического моделирования рисков их возникновения и развития в зависимости от разных факторов влияния. Основными задачами проактивного мониторинга являются:

– оценка и прогнозирование рисков возникновения и развития нештатных и аварийных событий с учетом влияния внешних факторов;

– контроль участков дорог и транспортных средств;

– отслеживание изменений факторов влияния на негативные события посредством анализа и прогностического моделирования временных рядов;

– локализация мест аварий и прочих деструктивных событий и т.д.

Платформа для сбора и анализа данных использует различные инструменты, например такие, как Apache HBase, Apache Hadoop, Apache Storm, Apache Spark, библиотеки алгоритмов интеллектуального анализа и машинного обучения MLlib (Apache Spark) и Mahout (Apache Hadoop). Библиотеки MLlib и Mahout включают программные реализации алгоритмов интеллектуального анализа и прогностического моделирования с поддержкой технологии MapReduce. Интеллектуализация распределенной системы проактивного мониторинга дорожно-транспортной инфраструктуры на основе сбора и обработки больших данных о происходящих событиях необходима для повышения безопасности дорожного движения. В процессе анализа негативных событий и сравнения паттернов временных рядов происходит выявление схожих участков дорожно-транспортной инфраструктуры по количеству и виду дорожных происшествий. Кластеризация позволяет выделить критические и аварийные участки и представить их на картографической основе с цветовой дифференциацией опасных зон. В процессе анализа временных рядов с моментами инцидентов также определяются временные интервалы, в которые происходит аномальное отклонение количества происшествий от средних показателей. В результате сопоставления временных рядов выявляются факторы, которые с высокой вероятностью становятся определяющими для аномального изменения дорожно-транспортных ситуаций на контролируемых участках. Конечной целью является выявление критических пространственно-временных зон и факторов, которые вызывают возникновение и реализацию рисков дорожных инцидентов.

Результаты работы получены при финансовой поддержке РФФИ в рамках грантов № 18-010-00204-а, 18-07-00975-а, 19-013-00409-а. Результаты исследований, представленные в разделе 2, получены за счет средств Российского научного фонда (проект № 20-71-10087).

Библиографический список

1. Lawrence, M. What Is Proactive Monitoring? Small Business – Chron.com. – URL: http://smallbusiness.chron.com/proactive-monitoring-73438.html (дата обращения: 21.01.2020).
2. Proactive Management of Complex Objects Using Precedent Methodology / A. Bakhmut, A. Krylov, M. Krylova, M. Okhtilev, P. Okhtilev, B. Sokolov // Artificial Intelligence and Algorithms in Intelligent Systems / ed. by R. Silhavy. – 2019. – Vol 764.
3. Lee, E. A. The Past, Present and Future of Cyber-Physical Systems: A Focus on Models / E. A. Lee // Sensors. – 2015. – Vol. 15. – P. 4837–4869.

4. Design, Modelling, Simulation and Integration of Cyber Physical Systems: Methods and Applications / P. Hehenberger, B. Vogel-Heuser, D. Bradley, B. Eynard, T. Tomiyama, S. Achiche // Computers in Industry. – 2016. – Vol. 82. – P. 273–289.
5. Hersent, O. The Internet of Things: Key Applications and Protocols / O. Hersent, D. Boswarthick, O. Elloumi. – Willey, 2012. – 370 p.
6. Monnin, M. Proactive Fleet Health Monitoring and Management / M. Monnin, J. Leger, D. Morel // Engineering Asset Management / ed. by J. Lee, J. Ni, J. Sarangapani, J. Mathew. – Lecture Notes in Mechanical Engineering. – London : Springer, 2011. – URL: https://doi.org/10.1007/978-1-4471-4993-4_28 (дата обращения: 21.01.2020)
7. Proactive behavior-based system for controlling safety risks in urban highway construction megaprojects / Li. Yongkui et al. // Automation in Construction. – 2018. – Vol. 95. – P. 118–128.
8. Ouzounis, G. Smart cities of the future / G. Ouzounis, Y. Portugali // The European Physical Journal Special Topics. – 2012. – Vol. 214 (1). – P. 481–518.
9. Multiagent Intelligent System of Convergent Sensor Data Processing for the Smart&Safe Road / A. Finogeev, A. Bershadsky, A. Finogeev, L. Fionova, M. Deev // Intelligent System / ed. by W. Chatchawal. – 2018. – Ch. 5. – P. 102–121.
10. Persia, L. Management of Road Infrastructure Safety / L. Persia, D. Usami et al. // Transportation Research Procedia. – 2016. – Vol. 14. – P. 3436–3445.
11. Department for Transport, Reported road accidents and casualties, Great Britain 2011, Table RAS 30070. – URL: https://www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents, last accessed 2020/01/21
12. Manikonda, P. Intelligent traffic management system / P. Manikonda, A. Yerrapragada, S. Annasamudram. – P. 119–122. – URL: https://10.1109/STUDENT.2011.6089337 (last accessed: 2020/01/21)
13. Industry Article: Proactive Event Processing in Action: A Case Study on the Proactive Management of Transport Processes / Z. Feldman et al. // Proceedings of the Seventh ACM International Conference on Distributed Event-Based Systems. – Arlington, Texas, USA, 2013. – P. 97–106.
14. eWeek / Preimesberger, Chris Hadoop, Yahoo, 'Big Data' Brighten BI Future. – URL: https://www.eweek.com/storage/hadoop-yahoo-big-data-brighten-bi-future (дата обращения: 02.01.2020).
15. Маликова, С. Big Data: тенденции развития, опасности и перспективы / С. Маликова // Экономика и жизнь. – 2018. – № 17–18 (9733). – URL: https://www.eg-online.ru/article/372363/ (дата обращения: 24.12.2019).
16. Gantz, J. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East – United States / J. Gantz, D. Rainsel // IDC Country brief, 2013.
17. Иванов, П. Д. Технологии Big Data и их применение на современном промышленном предприятии / П. Д. Иванов, В. Ж. Вампилов // Наука и инновации. – 2014. – Вып. 8. – URL: http://engjournal.ru/articles/1228/1228.pdf (дата обращения: 03.01.2020).
18. Big Data: Принципы работы с большими данными. – URL: https://habr.com/ru/company/dca/blog/267361/ (дата обращения: 05.01.2020).
19. Атаманов, Ю. С. Введение в Big Data / Ю. С. Атаманов, В. С. Гончарук, С. Н. Гордеев // Молодой ученый. – 2017. – № 11. – С. 33–34. – URL: https://moluch.ru/archive/145/40562/ (дата обращения: 12.01.2020).
20. Кендэл, М. Временные ряды / М. Кендэл. – Москва : Финансы и статистика, 2015. – 200 с.
21. Бриллинджер, Д. Временные ряды. Обработка данных и теория / Д. Бриллинджер. – Москва, 2017. – 653 с.
22. Афанасьев, В. Н. Анализ временных рядов и прогнозирование / В. Н. Афанасьев, М. М. Юзбашев. – Москва : Финансы и статистика : Инфра-М, 2015. – 320 с.

23. Колесников, И. Н. Прогнозирование временных рядов посредством привязки событий / И. Н. Колесников // Моделирование, оптимизация и информационные технологии. – 2019. – № 7 (4). – С. 12–21. – DOI 10.26102/2310-6018/2019.27.4.039. – URL: https://moit.vivt.ru/wp-content/uploads/2019/11/Kolesnikov_4_19_1.pdf
24. Лукашин, Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов / Ю. П. Лукашин. – Москва : Финансы и статистика, 2015. – 416 с.
25. Neurohive. Word2Vec. Как работать с векторными представлениями слов. – URL: https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/ (дата обращения: 16.01.2020)

References

1. Lawrence M. *What Is Proactive Monitoring? Small Business – Chron.com*. Available at: http://smallbusiness.chron.com/proactive-monitoring-73438.html (accessed Jan. 21, 2020).
2. Bakhmut A., Krylov A., Krylova M., Okhtilev M., Okhtilev P., Sokolov B. *Artificial Intelligence and Algorithms in Intelligent Systems*. 2019, vol. 764.
3. Lee E. A. *Sensors*. 2015, vol. 15, pp. 4837–4869.
4. Hehenberger P., Vogel-Heuser B., Bradley D., Eynard B., Tomiyama T., Achiche S. *Computers in Industry*. 2016, vol. 82, pp. 273–289.
5. Hersent O., Boswarthick D., Elloumi O. *The Internet of Things: Key Applications and Protocols*. Willey, 2012, 370 p.
6. Monnin M., Leger J., Morel D. *Engineering Asset Management*. Lecture Notes in Mechanical Engineering. London: Springer, 2011. Available at: https://doi.org/10.1007/978-1-4471-4993-4_28 (accessed Jan. 21, 2020)
7. Yongkui Li. et al. *Automation in Construction*. 2018, vol. 95, pp. 118–128.
8. Ouzounis G., Portugali Y. *The European Physical Journal Special Topics*. 2012, vol. 214 (1), pp. 481–518.
9. Finogeev A., Bershadsky A., Finogeev A., Fionova L., Deev M. *Intelligent System*. 2018, ch. 5, pp. 102–121.
10. Persia L., Usami D.et al. *Transportation Research Procedia*. 2016, vol. 14, pp. 3436–3445.
11. *Department for Transport, Reported road accidents and casualties, Great Britain 2011, Table RAS 30070*. Available at: https://www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents, last accessed 2020/01/21
12. Manikonda P., Yerrapragada A., Annasamudram S. *Intelligent traffic management system*. Pp. 119–122. Available at: https://10.1109/STUDENT.2011.6089337 (last accessed 2020/01/21)
13. Feldman Z. et al. *Proceedings of the Seventh ACM International Conference on Distributed Event-Based Systems*. Arlington, Texas, USA, 2013, pp. 97–106.
14. *eWeek*. Preimesberger, Chris Hadoop, Yahoo, 'Big Data' Brighten BI Future. Available at: https://www.eweek.com/storage/hadoop-yahoo-big-data-brighten-bi-future (accessed Jan. 02, 2020).
15. Malikova S. *Ekonomika i zhizn'* [Economy and life]. 2018, no. 17–18 (9733). Available at: https://www.eg-online.ru/article/372363/ (accessed Dec. 24, 2019). [In Russian]
16. Gantz J., Rainsel D. *IDC Country brief*, 2013.
17. Ivanov P. D., Vampilov V. Zh. *Nauka i innovatsii* [Science and innovation]. 2014, iss. 8. Available at: http://engjournal.ru/articles/1228/1228.pdf (accessed Jan. 03, 2020). [In Russian]
18. *Big Data: Printsipy raboty s bol'shimi dannymi* [Big Data: Principles of working with big data]. Available at: https://habr.com/ru/company/dca/blog/267361/ (accessed Jan. 05, 2020). [In Russian]

19. Atamanov Yu. S., Goncharuk V. S., Gordeev S. N. *Molodoy uchenyy* [Young scientist]. 2017, no. 11, pp. 33–34. Available at: https://moluch.ru/archive/145/40562/ (accessed Jan. 12, 2020). [In Russian]
20. Kendel M. *Vremennye ryady* [Time series]. Moscow: Finansy i statistika, 2015, 200 p. [In Russian]
21. Brillindzher D. *Vremennye ryady. Obrabotka dannykh i teoriya* [Time series. Data processing and theory]. Moscow, 2017, 653 p. [In Russian]
22. Afanas'ev V. N., Yuzbashev M. M. *Analiz vremennykh ryadov i prognozirovanie* [Time series analysis and forecasting]. Moscow: Finansy i statistika: Infra-M, 2015, 320 p. [In Russian]
23. Kolesnikov I. N. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii* [Modeling, optimization and information technology]. 2019, no. 7 (4), pp. 12–21. DOI 10.26102/2310-6018/2019.27.4.039. Available at: https://moit.vivt.ru/wp-content/uploads/2019/11/Kolesnikov_4_19_1.pdf [In Russian]
24. Lukashin Yu. P. *Adaptivnye metody kratkosrochnogo prognozirovaniya vremennykh ryadov* [Adaptive methods for short-term time series forecasting]. Moscow: Finansy i statistika, 2015, 416 p. [In Russian]
25. Neurohive. Word2Vec. *Kak rabotat' s vektornymi predstavleniyami slov* [Neurohive. Word2Vec. How to work with vector representations of words]. Available at: https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/ (accessed Jan. 16, 2020) [In Russian]

Колесников Илья Николаевич
младший разработчик,
ООО «КСК ТЕХНОЛОГИИ»
(Россия, г. Пенза, ул. Суворова, 64)
E-mail: iljakolesnikoff@yandex.ru

Kolesnikov Ilja Nikolaevich
junior java developer,
LLC «KSK TECHNOLOGIES»
(64 Suvorova street, Penza, Russia)

Финогеев Алексей Германович
доктор технических наук, профессор,
кафедра систем автоматизированного проектирования,
Пензенский государственный университет
(Россия, г. Пенза, ул. Красная, 40)
E-mail:alexeyfinogeev@gmail.com

Finogeev Alexey Germanovich
doctor of technical sciences, professor,
sub-department of computer-aided design systems,
Penza State University
(40 Krasnaya street, Penza, Russia)

Образец цитирования:
Колесников, И. Н. Проактивный мониторинг событий на основе предиктивного анализа временных рядов / И. Н. Колесников, А. Г. Финогеев // Модели, системы, сети в экономике, технике, природе и обществе. – 2020. – № 1 (33). – С. 111–125. – DOI 10.21685/2227-8486-2020-1-9.