Лядова Людмила Николаевна
доцент, кандидат физико-математических наук,
профессор кафедры математического обеспечения ВС,
Пермский государственный национальный исследовательский университет,
г. Пермь
E-mail: LNLyadova@mail.ru

Заякин Виктор Сергеевич
студент магистратуры, направление «Бизнес-информатика»,
Национальный исследовательский университет «Высшая школа экономики»,
г. Пермь
E-mail: VSZayakin@yandex.ru

Смирнов Михаил Андреевич
аспирант кафедры математического обеспечения ВС,
Пермский государственный национальный исследовательский университет,
г. Пермь
E-mail: MA_Smirnov2020@mail.ru

ФОРМИРОВАНИЕ СОБЫТИЙНЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ МНОГОАСПЕКТНЫХ ОНТОЛОГИЙ

Lyadova Lyudmila Nikolaevna
Docent, PhD in Computer Science,
Professor at Computing Systems Software Department
Perm State National Research University, Perm City
E-mail: LNLyadova@mail.ru

Zayakin Victor Sergeevich
Master's Student, Business Informatics
National Research University Higher School of Economics, Perm City
E-mail: VSZayakin@yandex.ru

Smirnov Mikhail Andreevich
Graduate Student of the Computing Systems Software Department
Perm State National Research University, Perm City
E-mail: MA_Smirnov2020@mail.ru

FORMATION OF EVENT SERIES USING MULTIFACETED ONTOLOGIES

Аннотация: Цель исследования – определение понятия событийного ряда и разработка подхода к формированию событийных рядов на основе данных, извлекаемых из различных источников в Интернет (новостных лент, социальных сетей и баз данных) с использованием многоаспектных онтологий, описывающих

как источники данных, так и структуру извлекаемой из них информации. Событийные ряды определяются по аналогии с временными рядами как совокупности значений некоторых параметров исследуемых процессов, где вместо времени измерения указывается тип события. При анализе необходимо учитывать не только взаимосвязь измерений со временем (или хронологию событий), но и причинно-следственные связи, которые могут быть выявлены при исследовании процессов. Онтологии используются при работе с неструктурированными данными для построения журнала событий, который формируется на первом шаге при построении событийных рядов.

Abstract: The purpose of the study is to define the event series concept and develop an approach to the event series formation based on data extracted from various Internet-sources (news feeds, social networks, mass media and databases) using multifaceted ontologies describing both data sources and the structure of information extracted from them. Event series are defined via analogy with time series as a collection of values of some parameters of the investigated processes, where the type of event is indicated instead of the measurement time. The analysis should take into account not only the relationship of measurements with time (or the events chronology), but also the causal relationships that can be identified at the study of processes. Ontologies are used when working with unstructured data to build an event log, which is formed in the first step when constructing event series.

Ключевые слова: журнал событий, анализ процессов, семантическое аннотирование, поиск данных, анализ данных, многоаспектная онтология.

Keywords: event log, process analysis, semantic annotation, data retrieval, data analysis, multifaceted ontology.

Введение

Существует множество источников информации в Интернет (около 1,5 млн новостей ежедневно), в базах данных и документах. Однако:

- Нет связи между публикуемыми данными различного типа, различными источниками информации.
- Нет баланса в распространении информации, объём данных не соответствует важности соответствующих событий.

Поиск информации – сложная задача, требующая поддержки. Однако выявление зависимостей между данными, связей между событиями – ещё более сложная задача. Для анализа процессов масштаба предприятия широко используются методы и *средства анализа процессов на основе журналов событий* [3, 4]. Проблема анализа процессов на основе информации, получаемой из Интернет, – отсутствие структурированных данных, сложность построения журналов событий на основе данных, извлекаемых из различных гетерогенных источников, где информация, к тому же, публикуется на разных языках и т.п.

Ранее было дано понятие *глобального процесса* в контексте описанной

проблемы, был предложен подход к решению задачи анализа глобальных процессов на основе использования онтологии для поиска источников данных, извлечения и структурирования информации о событиях, построения журналов событий и их обработки, подготовки к использованию средств Process Mining (ProM), а также для настройки параметров анализа [2, 6, 7].

Обычно в ходе анализа процессов используются журналы событий, построенные в соответствии с протоколами, определяющими структуру журналов. Однако отмечается, что данные, получаемые из различных источников информации дают гораздо больше информации. которую можно было бы использовать в ходе анализа, в частности, обычно отбрасываются количественные параметры, числовая информация.

Существующие средства анализа процессов позволяют расширять протоколы, доопределять структуру журналов событий с целью расширения возможностей анализа.

Ставится задача *создания средств анализа процессов на основе информации, извлекаемой из гетерогенных источников данных*, неструктурированных и слабо структурированных документов, с целью выявления закономерностей, связывающих события различных типов и изменения количественных параметров, характеризующих объекты, изменение их состояния, вызванного этими событиями.

Подобные средства актуальны при проведении междисциплинарных исследований, в которых участвуют специалисты, эксперты из различных предметных областей. В качестве примера можно рассмотреть, например, события, связанные с распространением COVID-19, выявление закономерностей, связей между вводимыми или снимаемыми ограничениями и ростом заболеваемости, изменениями экономических показателей и политическими событиями и т.д.

Проектирование и реализация этих средств основаны на *понятии событийного ряда*, определение которого приводится в данной работе.

Понятие и построение событийного ряда

Понятие событийного ряда вводится по аналогии с понятием *временного ряда* (или *ряда динамики*) – собранного в разные моменты времени статистического материала о значении каких-либо параметров (в простейшем случае одного параметра) исследуемого процесса. Каждая единица статистического материала называется измерением или отсчётом, также допустимо называть его уровнем на указанный с ним момент времени. Во временном ряду для каждого отсчёта должно быть указано время измерения или номер измерения по порядку.

Событийный ряд определим в *два этапа*:

- Определим понятие событийно-временного ряда.
- Определим понятие событийного ряда.

Предположим, что состояние системы описывается m различными показателями, объединенных в k групп по обобщающим признакам, которые могут быть выбраны, исходя из нужд анализа (экономические, политические

группы показателей и пр.).

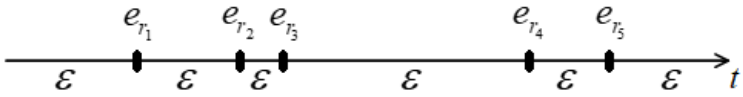
Для определенности допустим, что множеством значений каждого показателя является поле вещественных чисел R. Пусть имеется n наблюдений (показатели разных групп обозначим разными буквами):

$$x_i = [(a_{i1}, a_{i2}, \ldots, a_{ig_1}), (b_{i1}, b_{i2}, \ldots, b_{ig_2}), \ldots, (c_{i1}, c_{i2}, \ldots, c_{ig_k})] \in R^m, i = 1, 2, \ldots, n; \sum_{s=1}^{k} g_s = m.$$

Введем в рассмотрение множество $E = \{e_j \mid j = 1, 2, ..., l\}$, состоящее из l элементов, определяющих *типы событий*, которые могут наступать в системе.

Событийным пространством $S(E)$ назовем тройку: $S(E) = <R^m, T, E \cup \{\varepsilon\}>$, где T – компонент, определяющий время, а ε – специальный неопределённый тип события. Каждой точке q_x событийного пространства соответствует *состояние* системы, которое определяется *вектором наблюдений* x.

Событийное пространство представляет собой модель динамической системы. В любой момент времени система характеризуется значениями показателей, а события, возникающие в ней, могут принадлежать определенному типу из множества E. Интервалы времени между регистрацией событий характеризуются неопределённостью типа:



Наблюдением, зарегистрированным в момент времени t, является точка событийного пространства с неопределенным типом события: $e_x^- = (x, t, \varepsilon)$.

Событие, зарегистрированное в момент времени t, определим как наблюдение с присвоенным ему по определенному правилу типом события:

$$e_x = (x, t, e_x^*), e_x^* \in E.$$

Событийно-временным рядом $Ser(E)$ назовем множество, состоящее из событий, упорядоченных по времени наступления (наблюдения):

$$Ser(E) = \{e_{x_i} \mid i = 1, 2, \ldots, n\}, e_{x_i} = (x_i, t_i, e_{x_i}^*); Ser(E) \subset S(E).$$

Для событийно-временных рядов определены такие характеристики как *плотность последовательности событий*, *коэффициент уплотнения ряда* относительно заданного события. Они используются для выявления событий, оказывающих значительное влияние на частоту появления событий в системе после их наступления, что может быть использовано для нахождения причин перехода системы в определенные состояния.

Событийный ряд формируется на основе событийно-временного ряда. Его элементами являются *события* определенного типа, между которыми устанавливаются причинно-следственные связи. С каждым событием, включённым в ряд, связана последовательность моментов времени, когда были зафиксированы эти события, а также последовательности значений рассматриваемых показателей. При формировании событийно-временных рядов данные могут быть получены из различных источников.

Пример формирования событийно-временного ряда показан на рис. 1. Пример формирования событийного ряда на основе построенного событийно-временного ряда показан на рис. 2.

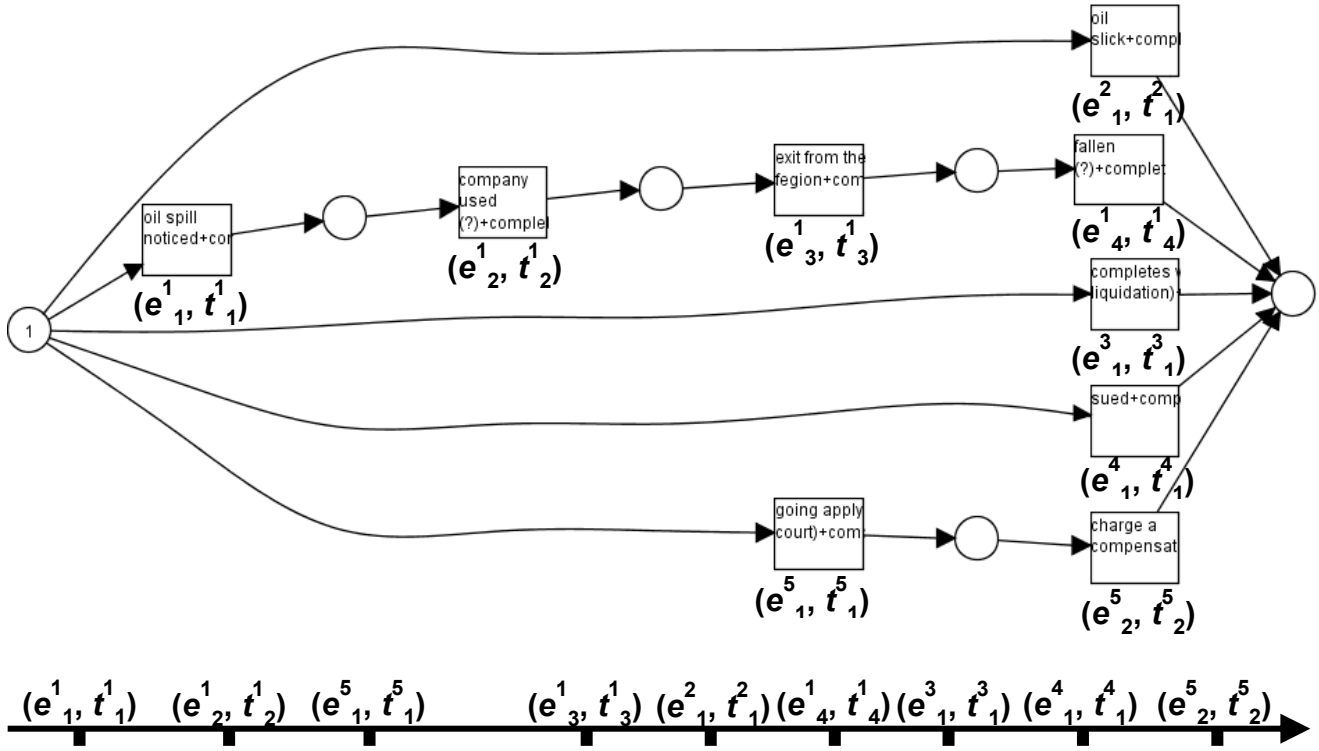


Рис. 1. Пример формирования событийно-временного ряда на основе трасс событий

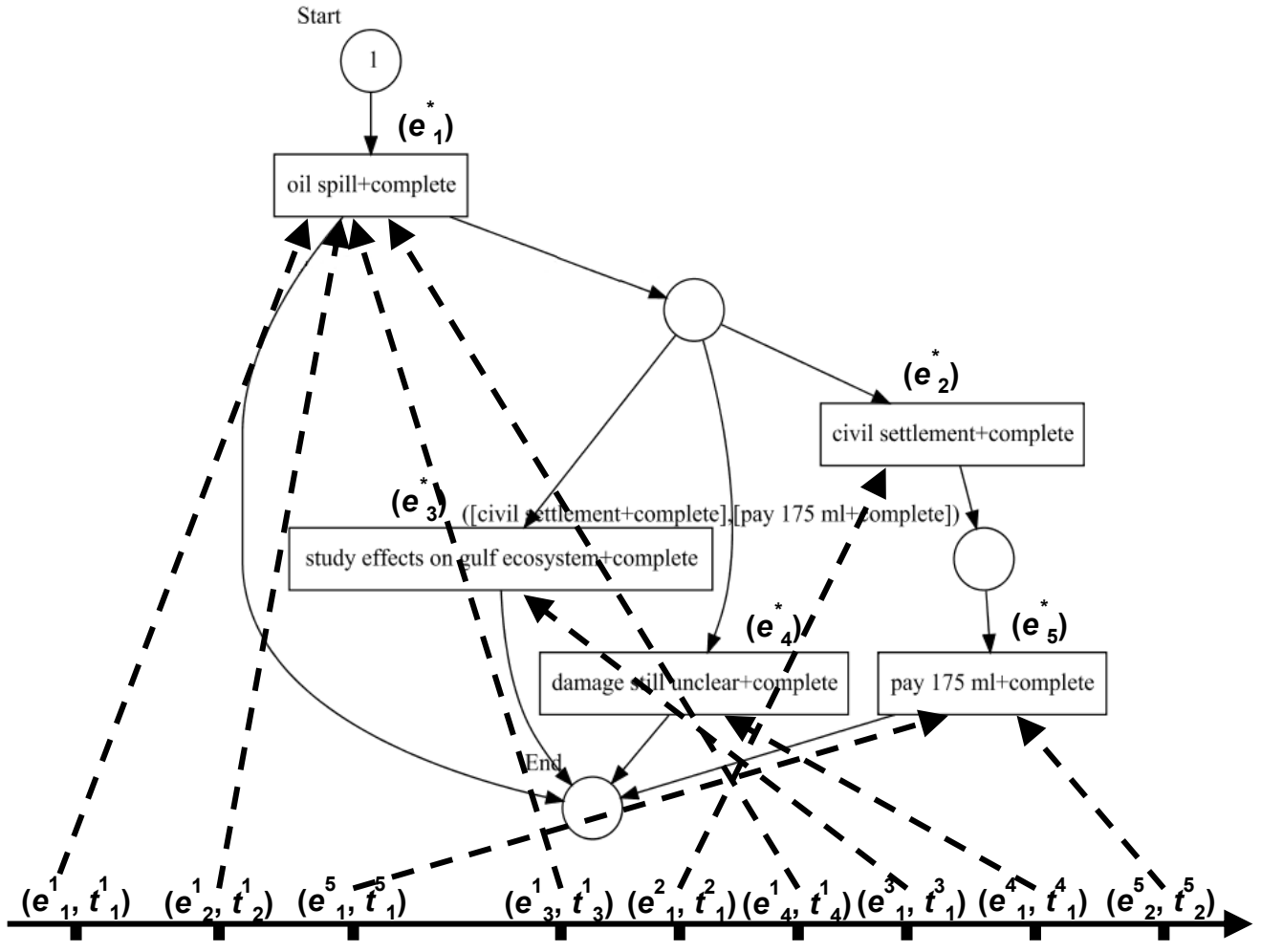


Рис. 2. Пример формирования событийного ряда на основе построенного событийно-временного ряда

Источниками данных для формирования событийных рядов являются новостные ленты, социальные сети, базы данных, аналитические отчёты и т.д. Используются как структурированные данные, так и неструктурированные источники. Поиск источников данных и извлечение информации осуществляется с использованием многоаспектной онтологии, включающей онтологию источников данных, описывающую источники, информация из которых должна использоваться в ходе исследования; онтологию структуры данных, описывающую структуру журнала событий, расширенную дополнительными атрибутами, которые должны быть учтены в ходе анализа; онтологию предметной области (или нескольких областей при проведении междисциплинарных исследований).

Архитектура программной системы

На рис. 3 представлена архитектура разработанного программного средства. Для анализа процессов на основе расширенных журналов событий используется система ProM. В качестве модели журналов выбрана модель XES, описанная онтологией.

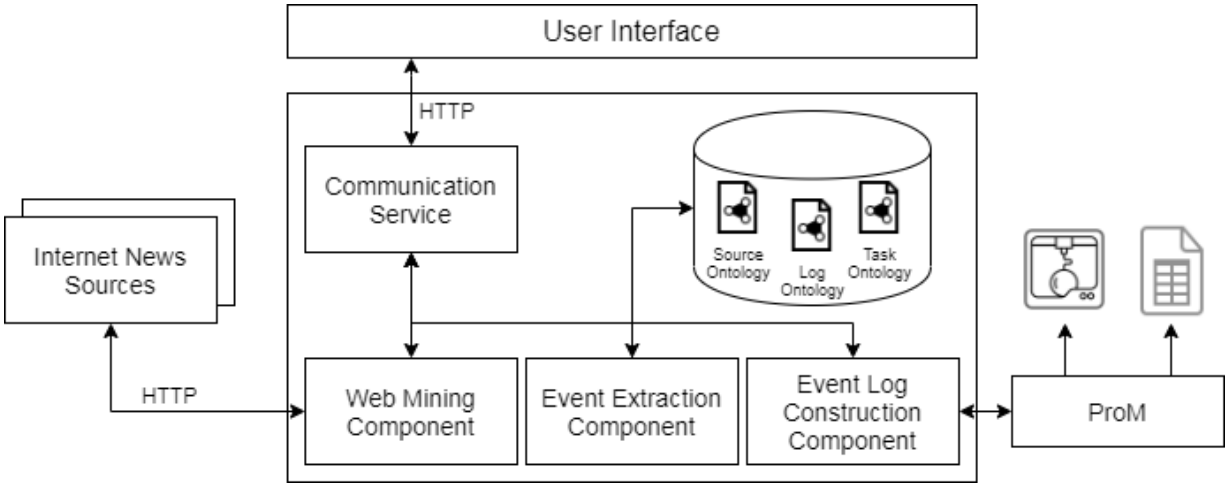


Рис. 3. Архитектура системы анализа событийных рядов

При реализации системы применяются различные подходы к решению задач извлечения информации из Интернет-ресурсов, обработки журналов [1, 3, 4, 5, 8, 9]. Семантическое аннотирование основано на ролевой структуре предложения, а также используется набор библиотек Pullenti SDK поддерживающий выделение основных актантов предиката.

Заключение

Разработанные программные средства прошли апробацию, исследование проведено на основе извлечения и анализе данных о событиях, связанных с эпидемией COVID-19. Анализ полученных результатов показал практическую значимость предложенного подхода.

Список литературы

1. Сигов А.С., Жуков Д.О., Новикова О.А. Моделирование процессов реализации памяти и самоорганизации информации при прогнозировании новостных событий с использованием массивов естественно-языковых текстов // Современные информационные технологии и ИТ-образование. – 2016. – Т. 12, № 1. С. 42-55.

2. Шаляева И.М., Ланин В.В., Лядова Л.Н. О проекте разработки системы мониторинга глобальных процессов на основе Интернет-новостей // Технологии разработки информационных систем – ТРИС-2016: материалы VII Международной научно-технической конференции. Том 1. – Таганрог: Издательство ЮФУ, 2016. С. 166-170.

3. Cook J.E., Wolf A.L. Discovering Models of Software Processes from Event-based Data // ACM Transactions on Software Engineering and Methodology. – 1998.– Vol. 7, № 3. – P. 215-249.

4. De Medeiros A.K.A., Van der Aalst W.M.P., Pedrinaci C. Semantic process mining tools: Core building blocks // Proceedings of the 16th European Conference on Information Systems. – Galway, Ireland, 2008. P. 15-23.

5. Salton G., Wong A., Yang C.S. A vector Space Model for Automatic Indexing // Communications of the ACM. – 1975. – Vol. 18, № 11. – P. 613-620.

6. Shalyaeva I., Lyadova L., Lanin V. Events Analysis Based on Internet Information Retrieval and Process Mining Tools // Proceedings of 10th International Conference on Application of Information and Communication Technologies (AICT2016). – Baku: Publishing Department of Qafqaz University, 2016. – P. 168-172.

7. Shalyaeva I., Lyadova L., Lanin V. Ontology-Driven System for Monitoring Global Processes on Basis of Internet News // 11th IEEE International Conference on Application of Information and Communication Technologies (AICT): Conference Proceedings (Vol.2). – M.: Institute of Electrical and Electronics Engineers Inc., 2017. P. 385-389.

8. Utiu N., Ionescu V. Learning Web Content Extraction with DOM Features // Proceedings of the IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP). – Cluj-Napoca, 2018. – P. 5-11.

9. Zhan L., Jiang X. Survey on Event Extraction Technology in Information Extraction Research Area // Proceedings of the IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). – Chengdu, China, 2019. P. 2121-2126.