УДК 004.9

И.М. Шаляева

МОНИТОРИНГ ЭКОЛОГИЧЕСКИХ КАТАСТРОФ И ИХ ПОСЛЕДСТВИЙ НА ОСНОВЕ INTERNET-НОВОСТЕЙ

Аннотация: Описываются результаты исследования возможности применения средств Process Mining для мониторинга глобальных процессов, связанных с экологическими катастрофами. Исходные данные для анализа извлекаются из новостных лент.

Ключевые слова: поиск информации, интеллектуальный анализ процессов, Fact Mining, Process Mining.

Введение

В результате роста и развития сети Internet все больше важных для решения различных задач данных можно почерпнуть именно оттуда. В данной работе для анализа глобальных процессов, связанных с экологическими катастрофами, информация о которых представлена в Internet, предлагается применять те же подходы, которые на сегодняшний день успешно применяются для работы с большими массивами данных о бизнес-процессах и событиях, представленных в виде журналов в различных информационных системах.

И задачи информационного поиска, и моделирование процессов на основе данных об операциях и их исполнителях из журналов событий активно применяются и при моделировании социальных сетей. Исследователи в этой области занимаются анализом и прогнозированием во всех сферах человеческой активности от изучения влияния погоды на изменение настроения [1] до предсказания курса акций на фондовой бирже [2], проведения маркетингового анализа Internet-среды (например, измерение удовлетворённости клиентов на основе анализа текстов [3] и пр.). Однако, хотя данные работы направлены на оценку влияния определённых факторов на развитие процессов, они не уделяют достаточного внимания вопросам моделирования процессов и их дальнейшего анализа: имитационного, ресурсного, функционального и пр.

Наиболее выгодным способом абсорбирования данных о процессах является их визуализация посредством построения графических моделей. Автоматизация процесса создания моделей на основе журналов событий – исследовательская область дисциплины Process Mining. Разработанные и реализованные в рамках данной методологии методы позволяют построить модель любого процесса, данные о котором представлены в виде логов определённых форматов.

Предлагаемый подход помимо возможности отслеживать и измерять корреляционную зависимость событий, описанных в Internet, с помощью средств интеллектуального анализа процессов позволяет к тому же моделировать и учитывать ранее не определённые пользователем факторы за счет возможности расширения границ анализируемой предметной области. Методы Process Mining позволяют извлекать информацию из журналов событий и строить на её основе формальные модели процессов. Кроме того, можно использовать методы интеллектуального анализа процессов для мониторинга отклонений (например, сравнивая наблюдаемое в журналах событий поведение с предопределёнными или регламентированными моделями и бизнес-правилами) [4].

Для формирования журнала событий необходимо найти нужную информацию о событиях в Internet, извлечь факты. При этом возникают проблемы, связанные с необходимостью обрабатывать тексты сообщений на разных языках, решать задачи, порождаемые различным именованием одних и тех же объектов, операций и пр. Одной из потенциальных возможностей преодоления проблем используемых моделей информационного поиска является встраивание в разрабатываемые программные средства знаний, описанных с помощью онтологических ресурсов [5, 6].

При разработке исследовательского прототипа системы мониторинга глобальных

процессов использованы существующие средства поиска (RapidMiner) и обработки данных о процессах (ProM). Программная платформа RapidMiner обеспечивает интегрированную среду для анализа текстов, интеллектуального анализа данных и бизнес-аналитики. RapidMiner используется в различных областях для быстрого прототипирования и разработки приложений [7], обеспечивает изучение схем, моделей и алгоритмов и может быть расширен с помощью R- и Python-скриптов [8]. Функциональность RapidMiner может быть расширена также с помощью дополнительных плагинов, которые доступны через RapidMiner Marketplace. ProM – это свободно распространяемая расширяемая платформа, которая содержит более 600 плагинов, охватывающих все возможности интеллектуального анализа процессов. Существует также и возможность разработки языков моделирования [9].

Исследования проводятся на примере мониторинга процессов, связанных с экологическими катастрофами, вызвавшими их причинами и последствиями. Для построения моделей процессов анализируется информация из новостных лент наиболее популярных и авторитетных поисковых систем, таких как Google, Yandex, а так же новостных порталов известных российских СМИ.

Анализ и представление знаний о предметной области

Прежде, чем переходить к решению задач поиска и извлечения информации, необходимо определить, с какими данными нужно работать, какую информацию извлекать и в какой формат преобразовывать. Для корректной постановки задачи поиска информации необходимо систематизировать знания о предметной области и выделить факты и понятия, используемые для извлечения и дальнейшей обработки методами Process Mining.

В качестве «ядра» предметной области выступают экологические опасные ситуации, антропогенные катастрофы. Как пример рассматриваются только *техногенные катастрофы*, связанные с нефтяной промышленностью, – чрезвычайные ситуации на всех этапах производства: разведки, добычи, переработки, транспортировки и продажи нефтепродуктов. Данный сектор выбран также и потому, что довольно просто отследить влияние катастроф на как экономику отдельных компаний, так и на регионы и даже страны. Известно множество случаев, когда масштабные нефтяные катастрофы повлияли на торговые взаимоотношения между государствами и даже повлекли изменения законодательства. Таким образом, работа в данной области может дать хорошие результаты в выявлении связей между различными сферами жизнедеятельности и, следовательно, предоставить широкий спектр возможностей для построения и дальнейшего анализа моделей.

Для того чтобы определить структуру базы для хранения данных, определить все атрибуты и характеристики событий и способы их взаимосвязи, были проанализированы результаты информационных запросов в поисковых системах и новостных агрегаторах. В результате все найденные события были классифицированы, для классов были определены ключевые атрибуты, по которым связываются события и новости. В одной новости могут присутствовать события, относящиеся к различным классам. Чтобы связать их, используются атрибуты-маркеры.

В ходе анализа новостной ленты нефтяных катастроф были выделены следующие *основные типы событий и ключевые атрибуты* для них: *Disaster* (*date, oil company, place*) – непосредственно сами катастрофы (пожар, разлив и пр.); *Financial implication* (*organization*) – оценки финансового ущерба (сюда относятся как затраты по устранению последствий, так и другие экономические показатели предприятий, населения, стран); *Industry news* (*oil company, publication date*) – возможные инновационные открытия, достижения, инновации, либо информация, связанная с функционированием компаний (расширение, закрытие, банкротство и т.п.); *Sanction* (*Date*) – информация о санкциях, штрафах; *Socio-environmental implication* (*publication date*) – влияние на население, жертвы, ущерб сельскому хозяйству, влияние на общество, возможные реакции (митинги, волнения и пр.); *Socio-political* (*Date, Place*) – влияние на политику государства, изменение взаимоотношений, влияние на внешнюю торговлю, морские пути; *Noise* – шумы.

Помимо ключевых атрибутов для каждого класса выделен также ряд *обязательных* и

необязательных атрибутов (дата и источник публикации, объёмы разлива, количество пострадавших и др.). Однако на данном этапе работа ведётся только в рамках *стандартных*, доступных *в формате XES расширений для атрибутов*: имя события, этап жизненного цикла, исполнитель, дата, ID, цена и семантическая связь (позволяет связывать события с элементами онтологий).

Процесс обработки новостных сообщений

Система RapidMiner *принимает на вход ссылку на источник* в Internet и *возвращает html-код страницы*. Далее идёт процесс *обработки текстового документа*: с помощью регулярных выражений вычленяются и разделяются на трассы новостные сообщения, из каждого новостного сообщения так можно выделить определённые разделы структуры документа. Для каждого источника разрабатывается собственное регулярное выражение. Кроме того, в полученных текстах необходимо *убрать теги разметки*. После получения текста на естественном языке и удаления тегов разметки применяется оператор *извлечения фактов* из текста, однако в системе не реализована возможность добавления пользовательских словарей, поэтому после извлечения фактов применяется *оператор фильтрации*. *Извлечение событий* в заданной области так же является проблемным этапом, поскольку система не обладает возможностью распознавать нефтяные катастрофы и их возможные последствия. Уровень данных, являющихся *шумом*, составляет огромный объем. Извлечённые события и их атрибуты могут быть сохранены в таблице и импортированы в ProM с помощью компонента XESame, для этого необходимо определить соответствие атрибутов в таблице и стандартных расширений лога.

На рис. 1 представлена модель, построенная в системе ProM на основе извлечённых из Internet-новостей данных о событиях, связанных с деятельностью нефтяных компаний.

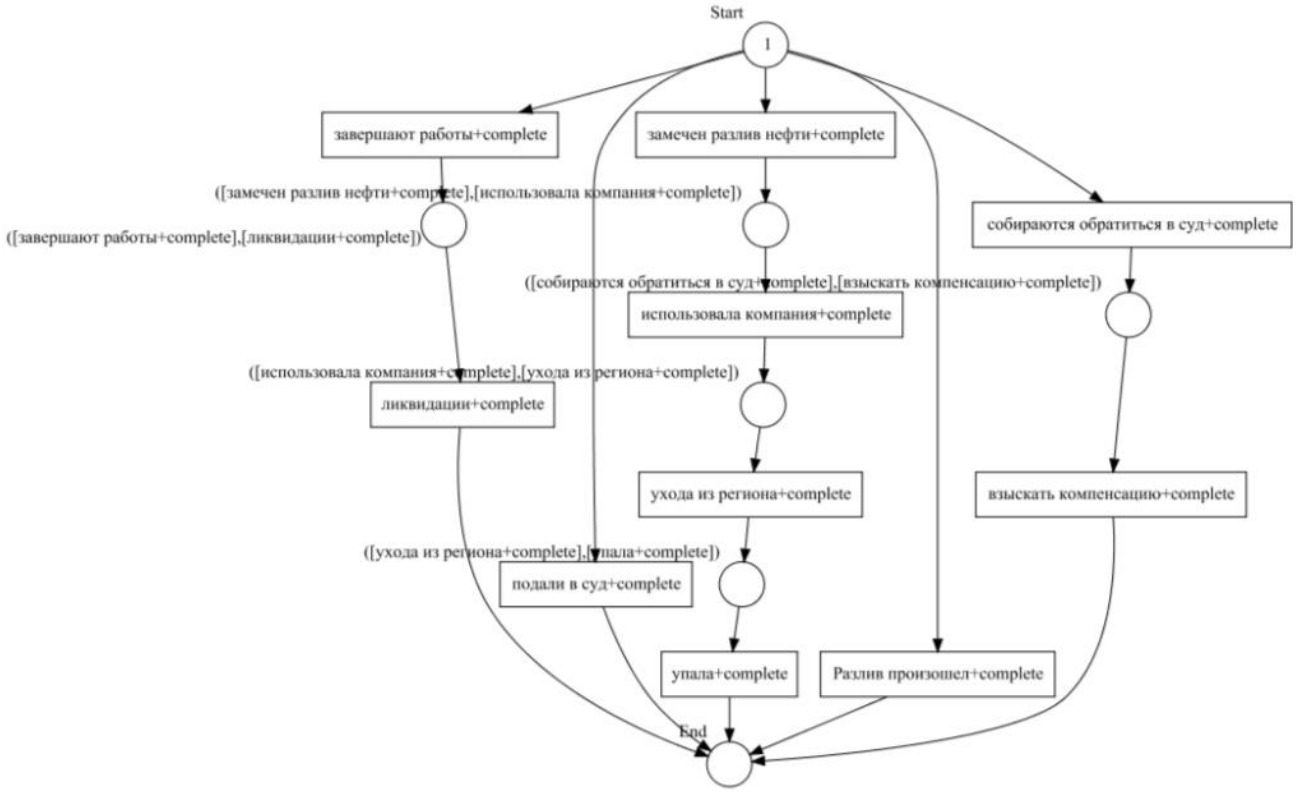

Рис. 1. Модель процесса, полученная в системе ProM

Можно заметить, что каждая трасса представлена как отдельный вариант развития событий, синонимичные понятия не были сгруппированы и приведены к единому виду, поэтому алгоритмы интеллектуального анализа процессов не смогли структурировать и выявить связи в трассах и последовательностях событий.

В ходе работы были выявлены следующие проблемы:

– Отсутствие методов и алгоритмов выявления причинно-следственных связей, как между, так и внутри трасс в случае непрямого порядка следования событий.

– Потеря информации и большое количество шумов при извлечении событий является результатом того, что не все события описываются глагольными словосочетаниями, а извлекаемые сочетания зачастую не несут относящейся к предметной области смысловой нагрузки.

– Отсутствие возможностей унифицирования синонимичных форм слов. Для того чтобы система извлекала корректные модели процессов, на вход необходимо подавать журналы событий, данные в которых приведены к единому формату, а все факты, обозначающие одно событие, должны быть выражены единообразно.

– Отсутствие возможностей создания перечня атрибутов и их наборов без необходимости программной разработки дополнительных операторов.

Совершенствование подхода на основе онтологий

Для устранения недостатков предлагается: использовать онтологии для поиска данных и обработки логов; создать визуальный язык (DSL) для разработки регулярных выражений; применить средства синтаксического разбора текстов новостей на основе синтаксических шаблонов; реализовать алгоритм автоматического наполнения базы новостей на основе параметров, заданных пользователем; для представления построенных моделей разработать и использовать DSL.

На рис. 2 представлен граф фрагмента *онтологии событий*, связанных с разливами нефти и их последствий (выделены три класса: события (их типы соответствуют классификации), маркеры или атрибуты событий, материалы). Понятия онтологии предметной области отображаются на *онтологию логов*. Результирующий документ (журнал) создаётся в формате XES.

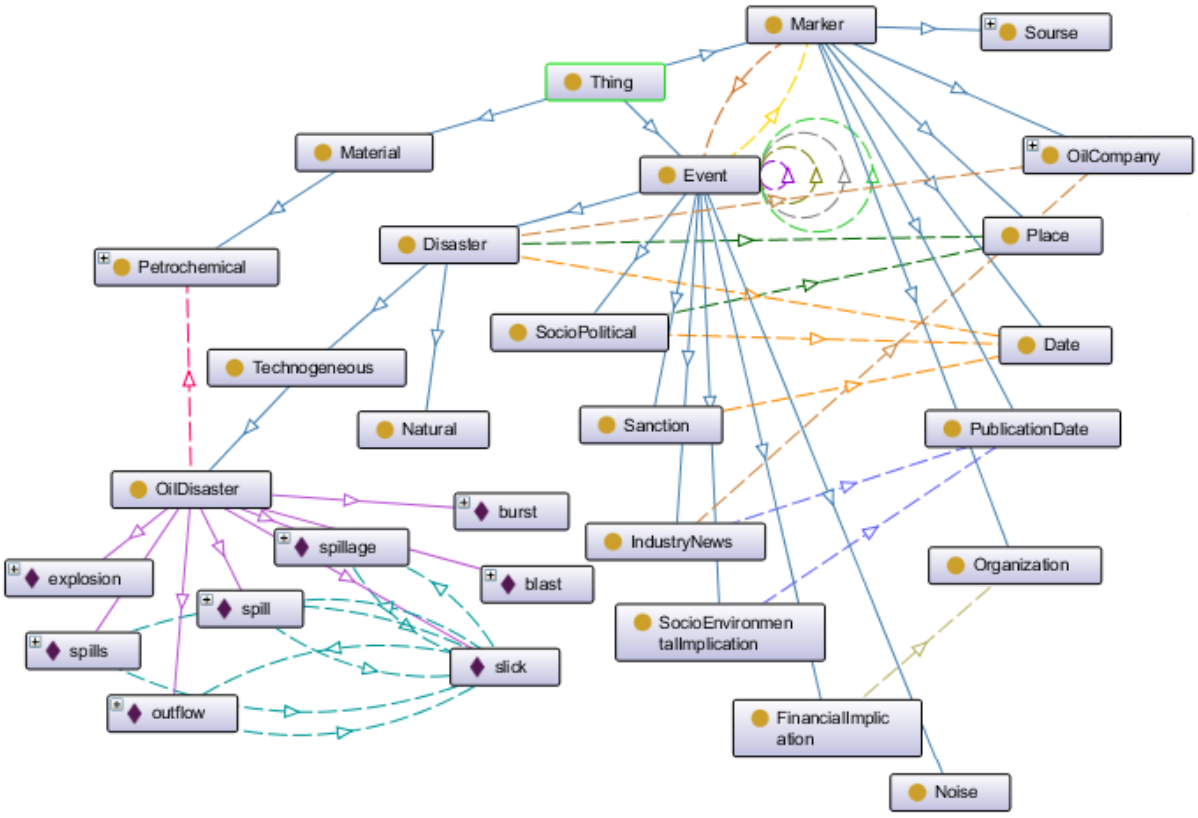

Рис. 2. Фрагмент онтологии предметной области

Журнал обрабатывается в ProM. На рис. 3 и 4 показаны уточнённые модели, построенные на основе логов, сформированных с использованием онтологий.

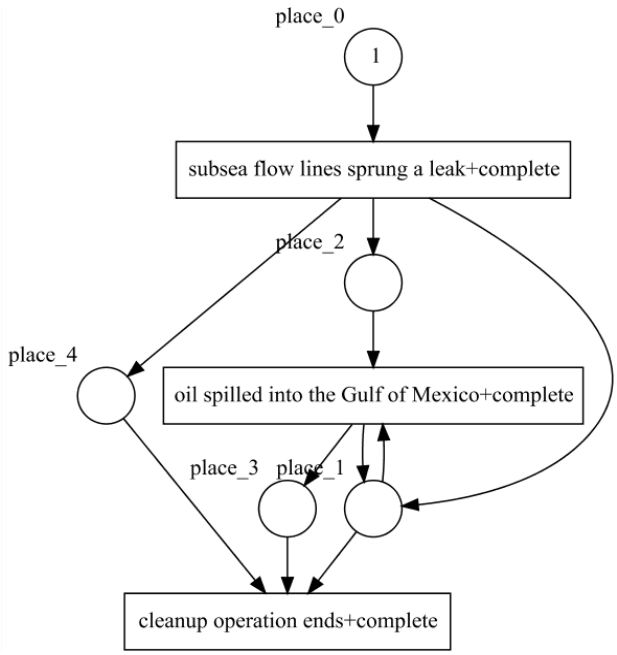

Рис. 3. Модель, построенная для запроса "Shell Spills Oil in the Gulf"

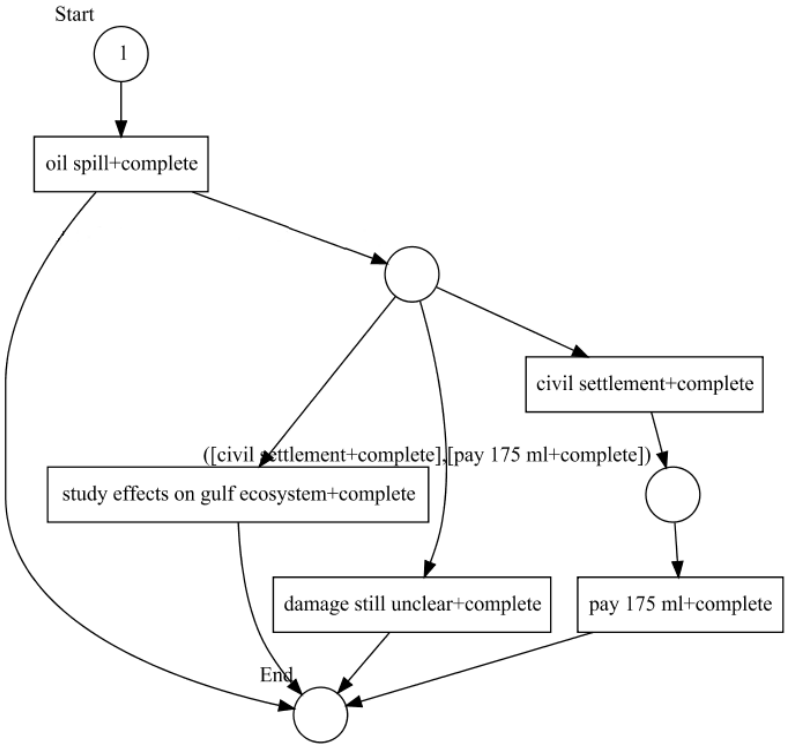

Рис. 4. Модель, построенная для запроса без указания места события

Заключение

Исследован подход к формированию моделей процессов, протекающих в рамках заданной предметной области и связанных с ней областей. Модели представляют процессы, последовательность событий в которых сформирована на основе причинно-следственных связей, выявленных в ходе анализа данных, полученных из Internet и ранее неизвестных для пользователя. В ходе эксперимента по реализации подхода на базе существующих инструментов была продемонстрирована несостоятельность их применения «в чистом виде», сформулированы основные проблемы, уточнены требования к усовершенствованию подхода и пути их реализации с использованием онтологий трёх уровней (Internet-источников, предметных областей, пользовательских запросов).

На следующем этапе необходимо реализовать программное расширение системы

RapidMiner для автоматизации наполнения базы новостей и предобработки данных на основе онтологического подхода и выполнить более тесную её интеграцию с ProM.

Ещё один шаг – расширение множества источников данных (Internet-новостей) через описание их онтологий и разработку соответствующих регулярных выражений. Для разработки выражений создаётся предметно-ориентированный язык (DSL), который позволит пользователям-непрограммистам настраивать систему на интересующие его источники данных.

Данный подход может использоваться в различных предметных областях. Для настройки на новые предметные области необходимо разработать их онтологии. Возможности системы расширяются за счёт использования для разработки онтологий визуального предметно-ориентированного языка. Причём сам язык расширяется и развивается при расширении онтологии. Созданные модели также визуализируются на DSL, который делает модели более наглядными и понятными аналитикам – экспертам в предметных областях.

Библиографический список

1. *Hannak, A.* Tweetin' in the Rain: Exploring Societal-scale Effects of Weather on Mood / *A. Hannak, E. Anderson, L. Barrett, S. Lehmann, A. Mislove* // Proceedings of the International AAAI Conference on Web and Social Media. – 2012.
2. *Bollen, J.* Twitter mood predicts the stock market / *J. Bollen, H. Mao, X. Zeng* // Journal of Computational Science. – 2011. – № 1. – С.1-8.
3. *Бойко, М.В*. Исследование удовлетворенности потребителей в банковской сфере на основе анализа текстовых отзывов / *М.В. Бойко* // Вестник УГАТУ. – 2014. – № 5 (66). – С. 139–145.
4. *van der Aalst, W.M.P.* Process Mining Manifesto / *W.M.P. van der Aalst, A. Adriansyah, A.K. Alves de Medeiros* [и др.] // BPM 2011 Workshops, Part I. T. 99. Springer-Verlag. – 2012. – С. 169–194.
5. *Киселев, С.Л.* Поиск фактов в тексте естественного языка на основе сетевых описаний / *С.Л. Киселев, А.Е. Ермаков, В.В. Плешко* // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. – Москва, Наука. – 2004.
6. *Кормалев, Д.А*. Повышение производительности при распознавании текстовых ситуаций // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием: Труды конференции. – М.: ЛЕНАНД. – 2008. – С. 192-200.
7. *Hofmann, M*. RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) / *M. Hofmann, R. Klinkenberg* // CRC Press. – 2013.
8. *Norris, D*. RapidMiner – a potential game changer / *D. Norris* // Bloor Research. – 2013.
9. *Шершаков, С.А*. DPMine/P: язык построения моделей извлечения и анализа процессов и плагины для ProM / *С.А. Шершаков* // В кн.: Proceedings of the 9th Central & Eastern European Software Engineering Conference in Russia / Науч. ред.: *A. Terekhov, M. Tsepkov*. NY : ACM. – 2013.

I.M. Shalyaeva

Monitoring of Environmental Disasters and Their Consequences on the Basis of Internet-news

Abstract: Results of the research of application possibility of Process Mining means for monitoring of the global processes associated with environmental disasters are described. Source data for the analysis are derived from newsfeeds.

Keywords: information retrieval, intellectual processes analysis, Fact Mining, Process Mining.