УДК 004.9

И.М. Шаляева, В.В. Ланин, Л.Н. Лядова

# О ПРОЕКТЕ РАЗРАБОТКИ СИСТЕМЫ МОНИТОРИНГА ГЛОБАЛЬНЫХ ПРОЦЕССОВ НА ОСНОВЕ ИНТЕРНЕТ-НОВОСТЕЙ

***Аннотация***: Описывается подход к анализу процессов на основе извлекаемых из новостных лент данных о событиях. Полученные данные обрабатываются средствами Process Mining, позволяющими построить формальные модели процессов.

***Ключевые слова***: анализ процессов, онтологии, Web Mining, Fact Mining, Process Mining.

## *Введение*

В настоящее время сложилась парадоксальная ситуация: с одной стороны, имеется множество источников информации в Интернет, представляющих различные точки зрения на происходящие события, факты и на роли в них государств, компаний или конкретных людей, а с другой – эксперты, аналитики испытывают недостаток данных, на основе которых можно было бы выявлять скрытые закономерности и принимать обоснованные решения.

Огромный объем информации, доступной из различных источников, не позволяет осуществить её анализ без применения специальных средств поддержки работы аналитиков.

Существует множество специализированных систем, позволяющих решать задачи анализа данных и процессов, BI-платформ, на основе которых могут быть созданы собственные системы бизнес-анализа. Эти системы используют в качестве источников открытые данные, публикуемые в Интернет, данные специализированных порталов и пр. Обычно эти данные представляют показатели экономического развития стран и регионов, отраслей промышленности, предприятий и пр. Анализ позволяет выявить динамику показателей, существующие между ними зависимости.

Отдельным направлением в области информационных технологий стала разработка средств конкурентной разведки в Интернет.

Ещё один класс систем решает задачи анализа данных в Интернет, мониторинга событий, в частности, в социальных сетях (например, ИАС «Семантический архив», и др.). Реализуются международные исследовательские проекты, направленные на решение подобных задач (например, международный проект SNAPSHOT: a Social Network Analysis Platform for the Support of European and Homeland Threat Prevention and Strategies). Эти средства позволяют не только выделить события, но и ранжировать их, проследить динамику обсуждений, выстроить связи между их участниками и пр. Результаты помогают решать задачи социологических или маркетинговых исследований, используются политологами и т.п. При реализации проектов широко используются статистические методы, методы компьютерной лингвистики, биоинспирированные методы (генетические алгоритмы, искусственные нейронные сети и пр.).

Существуют мощные поисковые системы, позволяющие не только осуществить интеллектуальный поиск данных, учитывающих их семантику, но и решать задачи аннотирования, кластеризации и классификации текстов и пр. Средства поиска данных в Internet позволяют находить публикации, содержащие информацию о событиях (фактах), об участвующих в них объектах и связях между ними, о времени, когда они произошли и пр. (например, RCO Fact Extractor и др.); эти средства позволяют отслеживать динамику событий, ранжировать их и т.д. [1, 2, 5, 6, 7, 9, 10]. Однако практически нет средств обработки этой информации, которые позволили бы обобщить полученные данные, построить формальные модели процессов на основе разрозненных фактов, установив причинно-следственные связи между соответствующими событиями.

Средства, реализующие методы Process Mining, широко используются для решения

задач углублённого анализа процессов в различных областях. Исходные данные для анализа – журналы событий – строятся на основе информации, получаемой из журналов операционных систем и СУБД, приложений. Данные в этих журналах структурированы и легко преобразуются к нужному формату, пригодному для обработки средствами Process Mining. Результатами работы являются формальные модели процессов, которые могут рассматриваться в качестве моделей «*As-Is*» при разработке информационных систем, автоматизирующих бизнес-процессы, для верификации разработанных системными аналитиками моделей, для решения задач оптимизации и реинжиниринга бизнес-процессов.

Предлагается исследовать возможность применения методов Process Mining для анализа событий, информация о которых публикуется в Internet, для мониторинга глобальных процессов, выявления закономерностей, связывающих отдельные события. Формальные модели процессов, которые должны быть построены в результате анализа, могут помочь в установлении причинно-следственных связей, прогнозировании событий на основе выявленных закономерностей.

### *Архитектура и схема работы системы мониторинга событий*

Общая схема реализации предлагаемого подхода включает следующие шаги:

1. Поиск информации об «исходных событиях», опубликованной в Internet, в соответствии с параметрами, определяемыми пользователями-аналитиками (пользователь задаёт интересующий его тип событий, объекты, с которыми могут быть связаны события, временные интервалы, когда могут произойти события, место, где эти события могут произойти, и пр.). Найденные факты становятся «точкой отсчёта»: именно относительно этих событий должны быть выстроены модели процессов.

2. Поиск информации о событиях, которые могут быть связаны с «исходными событиями»: могут предшествовать им и стать их причиной или могут следовать за ними. Поиск осуществляется на основе параметров запросов, которые определяются пользователями-аналитиками. Параметры могут локализовать место, задать время, конкретизировать типы событий, которые могут интересовать пользователя, и т.п.

3. Структурирование информации и найденных событиях, фактах.

4. Подготовка данных – предобработка информации о событиях для удаления избыточной информации и настройки данных для решения конкретных задач пользователей-аналитиков (удаляются записи, которые дублируют информацию об одних и тех же событиях, происходит классификация и «кластеризация» данных и пр.).

5. Преобразование данных о событиях к формату, пригодному для обработки средствами Process Mining (в формат журналов событий).

6. Построение моделей процессов на основе анализа журналов событий с помощью средств Process Mining.

Построенные модели – это информация для аналитиков, позволяющая им выявить связи между отдельными событиями (типами событий), оценить различные варианты развития процессов на основе выявленных закономерностей, найденных прецедентов.

Архитектура системы, реализующей представленный подход, показана на рис. 1.

На данном этапе (реализация исследовательского прототипа системы) для поиска информации используются доступные средства поиска информации в Internet. В результате анализа доступных средств выбрана система RapidMiner, которая удовлетворяет большинству функциональных требований (решает задачу извлечения фактов с выделением информации об объектах, датах и пр., позволяет выполнить классификацию и кластеризацию и пр.) и допускает возможность расширения. Кроме того, RapidMiner легко интегрируется с системой анализа процессов ProM, так как позволяет экспортировать результаты поиска данных в формат, пригодный для передачи в ProM. Таким образом, после доработки (реализации алгоритма автоматического пополнения базы новостей на основе параметров, задаваемых пользователями, с использованием базы знаний о предметных областях и источниках информации) данная система позволяет выполнить три первых шага схемы.

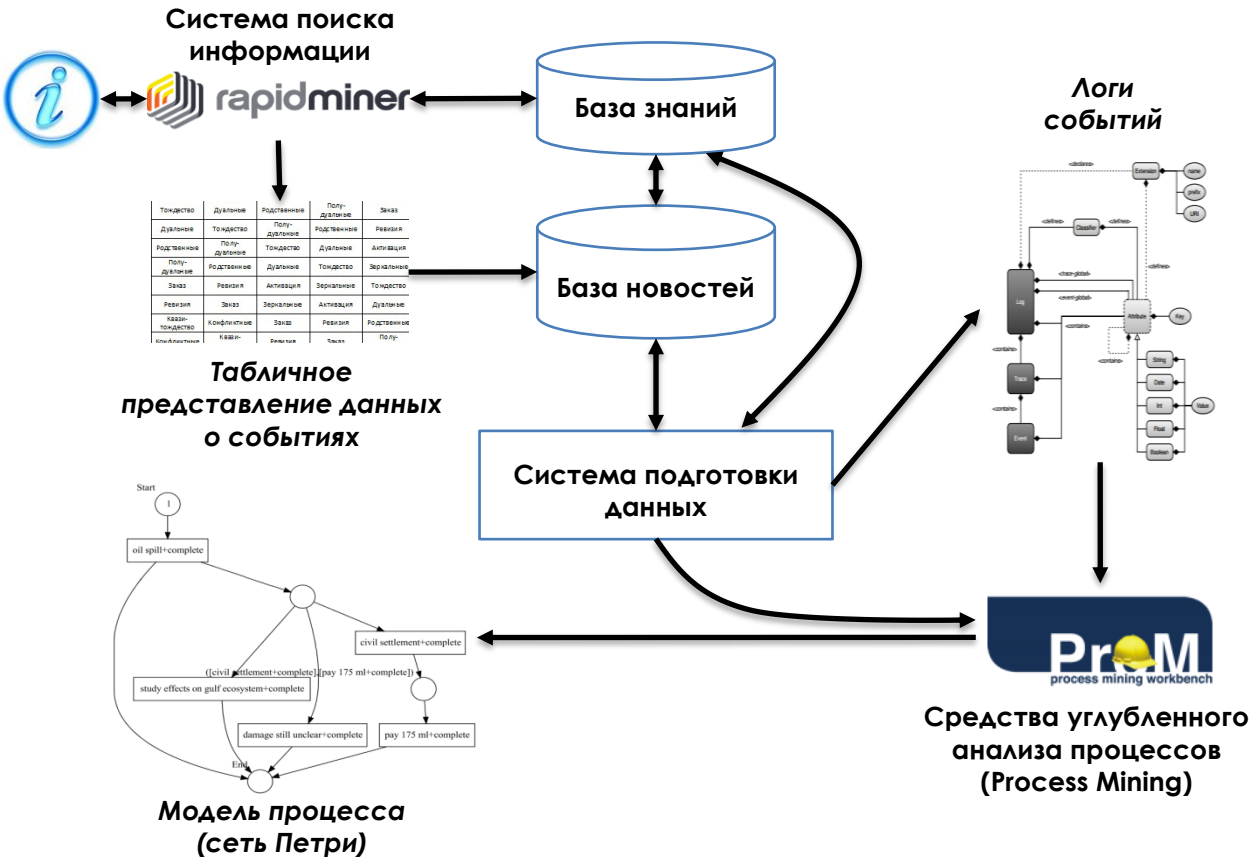


*Рис. 1.Схема анализа глобальных процессов на основе Internet-новостей и средств Process Mining*

Следующий шаг – подготовка журналов событий для анализа средствами Process Mining с учётом потребностей конкретного пользователя-аналитика и решаемых им задач. Пользователь может задать свои требования по выбору интересующих его событий (типов событий), которые могут относиться к различным предметным областям (например, если он решает задачу выявления зависимостей между экологическими катастрофами и ростом заболеваемости или изменением экономических показателей и т.п.), указать периоды времени для которых должен выполняться анализ, задать условия «кластеризации» фактов и пр. Подготовка данных осуществляется с использованием базы знаний системы. Основа базы знаний – онтологии рассматриваемых пользователями предметных областей. Кроме того, в базе знаний содержится информация об источниках данных (сведений о событиях) – онтология Internet-источников, которая позволяет настроить средства поиска на особенности извлечения информации из конкретных доступных источников (новостных лент и т.п.) [3, 4].

Выделенные в соответствии с параметрами пользователей сведения преобразуются к формату журналов событий, пригодному для анализа средствами интеллектуального анализа процессов Process Mining [8]. Примерами систем, реализующих различные возможности углублённого интеллектуального анализа процессов, являются ARIS Process Performance Manager (Software AG), Comprehend (Open Connect), Discovery Analyst (StereoLOGIC), Flow (Fourspark), Futura Reflect (Futura Process Intelligence) и множество других. Однако только академическая платформа ProM является свободно-распространяемой, расширяемой, содержит более 600 плагинов, охватывающих все возможности интеллектуального анализа процессов, поэтому именно она была выбрана для реализации исследовательского прототипа системы мониторинга глобальных событий.

С помощью ProM на последнем этапе описанной схемы построены формальные модели процессов в формате сетей Петри, отражающие выявленные связи между событиями. Подготовка данных с использованием базы знаний позволила уточнить построенные модели, решить задачи установления причинно-следственных связей.

### *Заключение*

При реализации исследовательского прототипа показана возможность реализации предложенного подхода. На следующем этапе необходимо оптимизировать представление знаний и реализовать алгоритм наполнения базы новостей. Необходима также доработка средств подготовки данных. Ещё одно направление развития системы – использование предметно-ориентированных языков (DSL) для построения моделей предметных областей и визуализации построенных моделей процессов.

### *Библиографический список*


1. *Ермаков А.Е.* Поиск фактов в тексте / *А.Е. Ермаков* // Мир ПК, № 02, 2005. (URL: http://www.osp.ru/pcworld/2005/02/169703/).
2. *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.
3. *Ланин В.* Интеллектуальный сервис анализа Интернет-контента на основе описания предметной области / *В. Ланин, А. Печенежский* // Математика программных систем: межвузовский сборник научных трудов / Вып. 10. Пермь : Пермский государственный национальный исследовательский университет, 2013. С. 10-19.
4. *Ланин В.В.* Онтология структуры веб-страниц / *В.В. Ланин, Р.А. Нестеров* // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ-2015). Т. 1. Новосибирск : Институт математики им. С.Л. Соболева СО РАН, 2015. С. 176-183.
5. *Balboni A., Colajanni M., Marchetti M., Melegari A.* Supporting Sense-Making and decision-Making through time Evolution Analysis of open Sources. In: Proceedings of the 7th International Conference on Cyber Conflict. 2015. P.185-202.
6. *Peña-Araya V.* Galean: Visualization of Geolocated News Events from Social Media / *V. Peña-Araya, M. Quezada, B. Poblete* // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). ACM New York. 2015. P. 1041-1042.
7. *Schuhmacher M.* Finding Relevant Relations in Relevant Documents / *M. Schuhmacher, B. Roth, S.P. Ponzetto, L. Dietz* // Advances in Information Retrieval: Proceedings of 38th European Conference on IR Research, ECIR. Padua, Italy, March 20-23, 2016. P. 654-660.
8. *van der Aalst W.M.P.* Process Mining Manifesto / *W.M.P. van der Aalst, A. Adriansyah, A.K. Alves de Medeiros* // BPM 2011 Workshops, Part I. Vol. 99. Springer-Verlag, 2012. P. 169-194.
9. *Vokhmintsev A.* The Knowledge on the Basis of Fact Analysis in Business Intelligence / *A. Vokhmintsev, A. Melnikov* // Digital Product- and Process Development Systems.– 2013.– P.134-141.
10. *Wadkar Sh.U.* A Review on Extracting Top-k Lists from the Web / *Sh. U. Wadkar, N.G. Pardeshi* // International Journal of Advanced Research in Computer and Communication Engineering. Vol. 4, Issue 12, 2015. P. 327-329.



I.M. Shalyaeva, V.V. Lanin, L.N. Lyadova


**On the Project of Development of Global Processes Monitoring System Based on Internet News**


***Abstract***: An approach to the processes analysis on the basis of the data on events mined from newsfeeds is described. Retrieved data are processed with the means of Process Mining allowing constructing the formal models of processes.



***Keywords***: process analysis, ontologies, Web Mining, Fact Mining, Process Mining.