Ontology-Driven System for Monitoring Global Processes on Basis of Internet News

Irina Shalyaeva
Business Informatics Department
National Research University Higher School of Economics
Perm, Russian Federation
ishalyaeva@bk.ru

Lyudmila Lyadova
Business Informatics Department
National Research University Higher School of Economics
Perm, Russian Federation
llyadova@hse.ru

Viacheslav Lanin
Business Informatics Department
National Research University Higher School of Economics
Perm, Russian Federation
vlanin@hse.ru

***Abstract* — An approach to development of ontology-driven system for monitoring global processes is presented. Ontology-driven architecture (ODA) intends using a few ontology types for different purposes. Ontologies are used to search Internet news with information retrieval tools. The domain ontology describes research interests of the experts needing analysis of world events and creation of global processes models including events of specified types. The sources ontology defines information sources (news feeds for example) used by experts for analysis. Also, domain ontologies are used to clarify results at facts extracting for forming event logs and to cluster objects in logs. Event logs formats are described with logs ontology. It allows storing logs instances with different settings too. Analysts can use these logs to compare results of global processes monitoring during time periods given by experts. Global process models are constructed with process mining tools on the basis of event logs. An advanced multilevel ontology is the kernel of system. The general system architecture and ontology representation are described. Some examples are included in the paper.**

***Index Terms* — domain ontology, Internet news, event analysis, fact extraction, event logs, process modelling, system architecture.**

I. Introduction

Global processes are the processes initiated by human activities in different spheres and enveloping essentially all space of the planet and society. Existence of social, political, economic and natural interactions in scales of large regions and the whole planet gives to the considered processes (and to problems) the status of globality. Globality characterizes unity of cross impact multiple-factor of communications in social, biological and technical sphere. Global process can include different events and sub-processes as components. Global processes evolve, create new tendencies and problems.

Social and economic public processes need to be analyzed in the context of real global social-natural processes, considering economy and policy as components of the global system "society–nature".

Thus, the task of monitoring of the phenomena, processes in different areas, detections of dependences and tendencies becomes especially urgent.

The approach to development of the system intended for formation of global processes visual models on the basis of searching of events, their possible reasons and consequences, significant for the user, is offered.

Boundaries of processes are defined by user (area of analyst's interests), and "starting point" for information search and monitoring of events which can be included in the investigated processes is the event interesting the user.

The choice of news messages as data source is caused by the fact that any significant event receives very fast response in news. News reports represent the status of events almost in real time, describe them in a natural language, enveloping all aspects of the events and phenomena, being one of the largest information sources about the modern society.

Besides, news texts contain a lot of factual data, being one of the best data sources for the existing methods of text processing. News lines don't specialize in certain event classes, so, news can be data source for any scale and expert's interest areas, given by users (they can select one and more data domains, for example: ecology, finance, healthcare and sport). Therefore, Internet news are an attractive and suitable data source for different analysis systems [1], [2], [3], [4].

The review of information from news reports in the Internet can be useful to different user groups from different points of view. So, for example, in ecological monitoring within the offered approach the user will be able to monitor influence of information about large oil disaster on changes of an ecological situation in the region, of an economic situation around the company which allowed catastrophe, of key persons provisions and even change of a political situation in the region; the analyst can reveal what political or public actions preceded "color revolutions" and that followed them; and so on.

There is no need to analyze complete texts of messages for formation of events information base. The analysis of the automatic summaries (news messages) provided on news sections of search engines and news aggregators revealed existence of all necessary data.

Events of different data domains and attributes characterizing them (dates, places of events, participants) can be derived from texts of news. The modern text analysis systems successfully extract the facts on the basis of machine learning methods or with use of syntactic patterns. Appearance

of information on events in one text, their sequence and some norms of language (for example, forms of words), allow to define relations between events and to build chronological chains of events.

There are different approaches both to extraction of the facts [5], [6], and to detection of relations between events.

Within suggested approach the most successful (from the point of view of similar determination of the concept "event") and suitable (specialization on news texts) is the approach to extraction of events and to detection of relations between them, described in papers of researcher group of laboratory of databases and information systems (DAIS) from Illinois University [7], [8], [9]. However, these methods are intended to detection of responses of social networks users to a chain of the interdependent events covered within one news. The graph, displaying dependences and frequency of entities appearance (the facts describing an event) in responses of users to a certain news, or the table, where events and responses to them are compared, is result of implementation of approach [9].

The existing approaches don't focus attention to possibilities of formation and visualization of event chains belonging to different data domains (global processes), and, as a result, don't assume implementation of any methods of the global processes visual model analysis.

Process Mining tools allow creating visual models of analyzed processes but researchers focused on processes of enterprise scale in different fields of activities [10], [11].

The goal of the presented project is development of ontology-driven global processes monitoring system available for users which are experts in different research domains collaborating at their researches.

The process model in one of the standard notations supported in the ProM system (Petri nets, eEPC, BPMN, etc.) is created as a result of Internet news analysis within suggested approach. Experts can analyze created models with any process mining tools.

II. General Scheme of Monitoring Global Processes on Basis of Internet News

The designed system is the analytical tool which provides uniform access point to the Internet sources set by users. The system is oriented to the data domains described by experts (analysts). It is capable to accumulate historical data and to independently request up-to-date information in the specified sources. Unlike the aggregators of news operating now, the created approach suggests revealing automatically dependences between the events described in the different Internet sources and to provide reports in the form of visual models of processes.

Implementation of system includes several stages of event search and data processing:

1. Domain ontology development. User can add new concepts and relations, event types he is interested in.
2. Sources ontology setting. User can describe information sources for monitoring news in research domain he is interested in.
3. Starting request forming. User define event to start information retrieval according to domain model and sources list set in ontology.
4. News search with information retrieval tools.
5. Extracting events and facts from results of information retrieval returned with search tools according to domain model.
6. Advanced search in the Internet sources of information on events which can be related to the starting event set by the user through formation of additional inquiries on the basis of the concepts and relations which are contained in domain ontology.
7. Results of information retrieval and fact extracting are represented in a XES format. It is a standard format for work with the systems realizing methods of Process Mining.
8. The prepared event logs in XES format are transferred to the ProM system which offers the user a set of various options for construction and the analysis of process models. Support of model creation with use of various visual languages of modeling is realized in this system (eEPC, BPMN, etc.).

The ProM system allows to analyze data from various points of view. It allows to detect not only chronology of events, but also to make assumptions of existence of regularities, relations of cause and effect between events. Besides, the system allows revealing communications between objects, information on which is provided in news.

The obtained data are used for replenishment of the base of events (expanded ontology) and can be used for the subsequent search and the analysis of news.

As this approach is focused on data processing from Internet news sources, it is obvious that such characteristics as reliability and relevance of information entirely depend on the Internet sources set by users. Therefore, and the assessment of quality of knowledge about events and processes presented as a result of work of approach means only check of compliance to the data published in set sources. So, if in model of process the event which hasn't taken place in reality is displayed it can't be considered as a mistake which can be found within the described approach if the relevant information has been provided by the chosen source.

III. Architecture of Ontology-Based System for Monitoring Global Processes

The ontology-based system for monitoring global processes on basis of Internet News includes following components:

- Web and text mining (data collection and text structuring component).
- User's requests constructor.
- Ontologies editor.
- Event logs preparing component.
- Process model analysis component.
- Advanced multilevel ontology including domain problems ontologies, sources ontology and event logs ontology.

General architecture of system is shown in Figure 1.

This system is developed to realize steps described above and includes components realizing information retrieval and process mining.

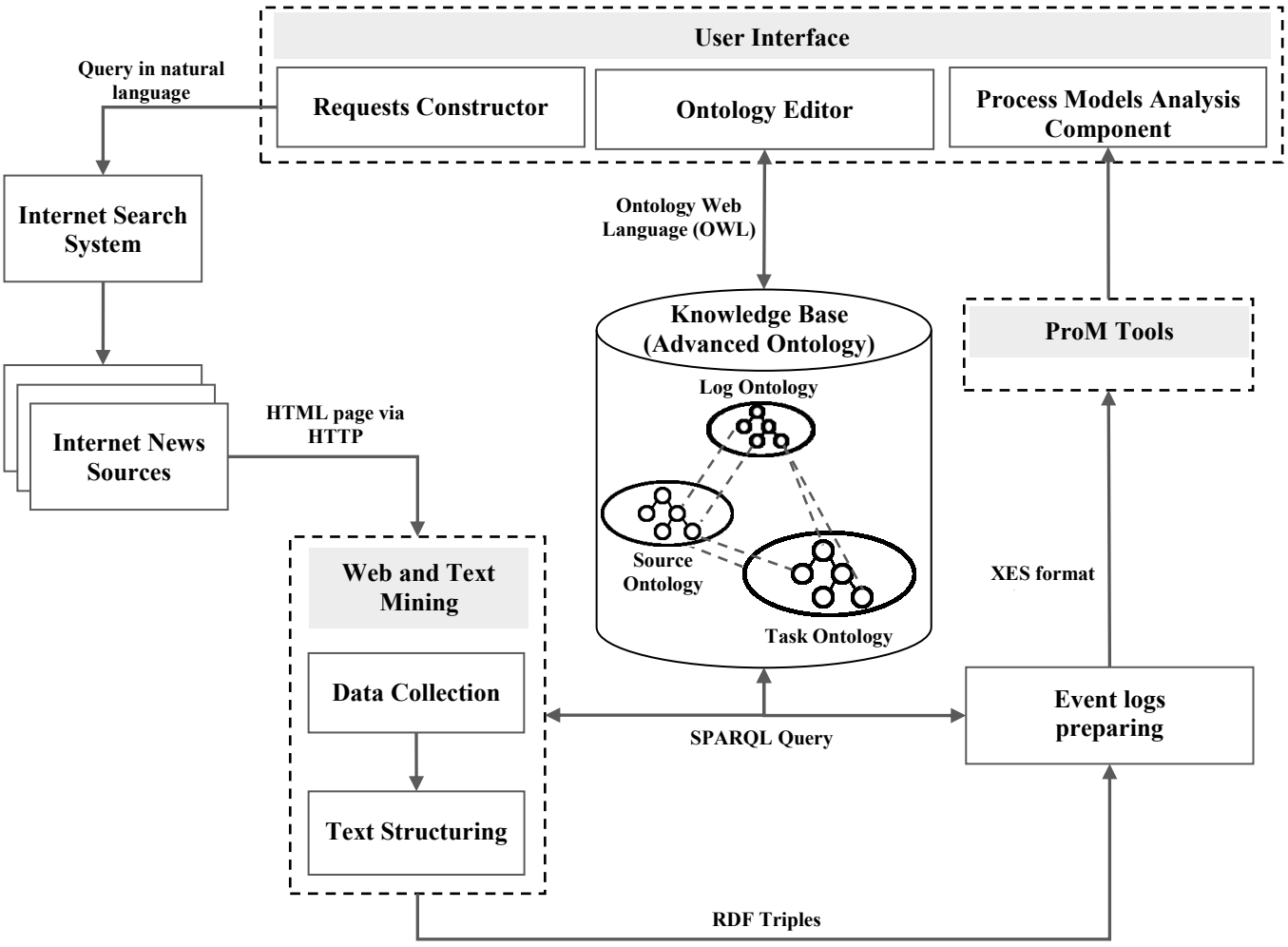

Fig. 1. Main components of the ontology-driven system for monitoring global processes on the basis of Internet news

Data from texts of news have to be extracted according to the patterns and their lexical-semantic description. Instances of the events and the facts representing concrete values (for example presented in the user ontological dictionaries: a name of the owner of the organization, the name of the country, a concrete event or type of event, etc.) are used as patterns.

Continuous expansion of the constructed model of problem area (expanded ontology) is planned. The found events and the facts have to be kept also in the base of events expanding basic ontology. Inquiries to search engines and news aggregators are also carried out by means of ontology. At formation of a data set for the module of automatic creation of processes models, it is also necessary to consider complex structures of relations between the facts and events. If as a result of model formation for concrete process the new facts or events are elicited, they are also added to model of the description of problem area.

To realize these opportunities, the decision to store data in the RDF format has been made. Each RDF file represents one trace of process (i.e. one transaction or the scenario), the route is formed on the basis of the text documents received as a result of one query to search engines the Internet.

For formation of connections between RDF documents time intervals and identifiers of the events described in them are compared. Thus, an opportunity to reveal communications between the events described in different texts and even results of different queries (Figure 2) is implemented.

Thus, all collected data will be submitted in language of the ontologies description and connected with each other through the ontological dictionary of problem domain (Figure 3).

Facts and events extraction is also carried out with means of ontology concepts. All entities which are described in the ontological dictionary can be taken from the text.

The lexical-semantic description of the concepts “event” and “fact”, allows to filter data from the grammatical point of view (in case the user has made mistakes when forming the ontological dictionary of subject domain). Besides, it allows to carry out expansion of ontology: in case the fact or an event which aren't described in the ontological dictionary is found in the text, the system will suggest the user to add the revealed data to model of his problem area.

Thus, the module of processing and extraction of the text interacts with the knowledge base, fills it and expands the

model of subject domain described in ontology. The inquiries formed by the user are also transferred to the module of processing and extraction of the text through the knowledge base for the purpose of formation of templates of inquiries and expansion of field of search due to automatic formation of queries on the basis of the connected concepts described in ontology of subject domain. After filling of the knowledge base data in the XES format can be transferred to the ProM system for automatic modeling of processes on their basis.

IV. An Example of Process Analysis: an Experiment on Monitoring of Oil Accidents with Using the Existing Tools

The experiment on realization of all stages of the developed approach is described in this section. Each stage is implemented with using one of the existing tools.

Main steps of an experiment are:

1) execution of user's query and extraction of data from a web page (this step is carried out with means of Google and RapidMiner);

2) analysis of data structure and formation of an event template (the system of automatic text processing is used to perform execution);

3) extraction of events and facts from the text (on this step Tomita-parser is applied);

4) construction and analysis of a process model with the ProM system.

At the first stage in the news section of the Google search engine a search is performed on the request of an "oil spill". As a result, a process trace is generated based on news messages collected from two pages of results. For the subsequent formation of traces it is necessary either to address to other information sources or to generate a specifying extended request.

According to the data obtained by a general request, the system can build a very simple process models within the same news. Obviously, this is not enough. So, user is prompted to specify new requests. Let's choose the information about Shell Spills Oil in the Gulf as example. Formal model generated with ProM on the base of this event log is given below (Figure 2).

The separate routes, presented in Petri net model, show different cases of the described process (different scenarios) received on the basis of the analysis with ProM of event logs, where information on events is obtained from variant sources in the Internet.

This model isn't informative:

- Each trace is provided as separate option of succession of events.
- Synonymous concepts are not grouped and brought to a uniform look therefore algorithms of the intellectual analysis of processes couldn't structure and reveal communication in routes and the sequences of events.

It is possible to mark the following shortcomings:

- Absence of a possibility of detection of event causes and effects, relationships between events.
- A lot of errors and losses in case of events extraction.

The revealed problems are solved when constructed ontologies [12] are used for information retrieval and event log preparing.

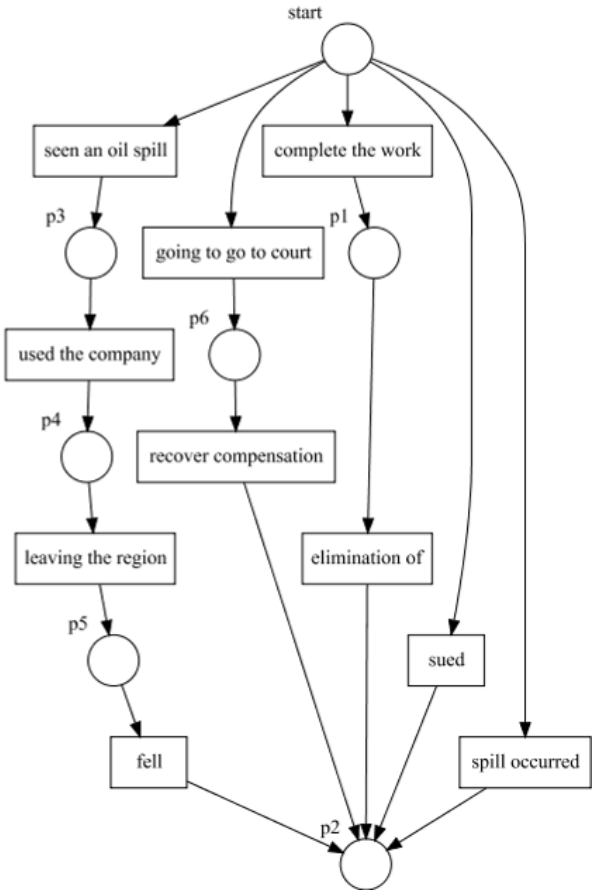

Fig. 2. Formal model generated with ProM on the base of facts table extracted from news

The models constructed with ontology (Figure 3) represent chains of the events found in different news messages and queries, at the same time, the possibility of displaying alternative scenarios isn't lost. Therefore, it is possible to say that the problems revealed at the previous stage of an experiment are solved partially or completely.

The created advanced model is more suitable for the analysis, is evident owing to smaller dimension, "merge" of the different routes. It can be important at the analysis of the difficult processes including a lot of types of events which are taking place on extended period and in various places. Besides, information on events, the facts can be received from various sources using different languages. This effect is received due to the preprocessing of the event log with use of ontology before its transmission to ProM.

In the model (Figure 3) some additional information on the socio-economic consequents related to the assessment of damage and the imposition of a fine is appeared. Thus, the user fills the database with new data and gets more detailed models via clarifying and expanding the information requests.

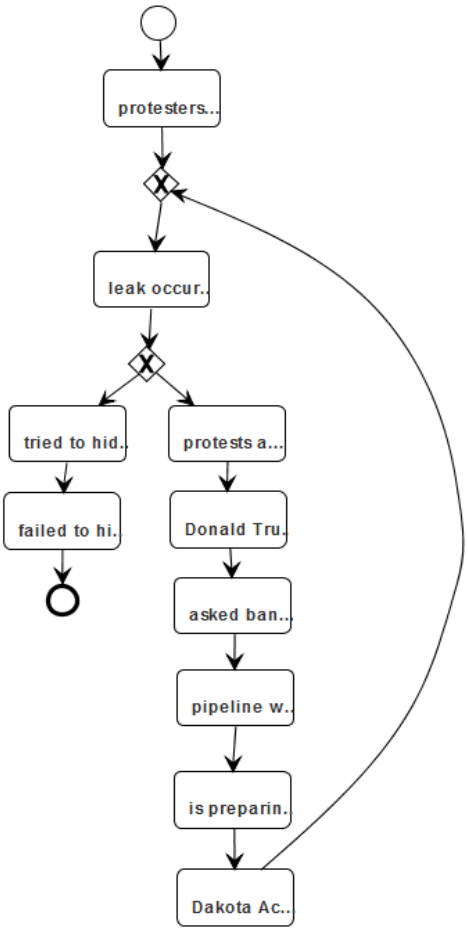

Fig. 3. The process model constructed for the event "oil spill energy transfer partners" in BPMN notation

V. Conclusion

An approbation of the offered approach for the analysis of the processes associated with the environmental technogenic catastrophes caused with oil spills showed prospects of the developed means. Ontologies allow performing flexible tuning for various data domains. All algorithms realized in system are based on advanced multilevel ontology. The ontological kernel of system provides new opportunities for the analysis of global processes, decrease of analysts' labor intensity. The system provides means of the versatile analysis of events and processes.

The conducted experiments demonstrate the ability of approach to provide assistance in the search for hidden information, facts and connections between them. Of course, the obtained models can serve as a source for generating different hypotheses and assumptions rather than reliable and complete description of the real situation. Attributes of events stored in the advanced ontology contain all necessary references to the sources, and the possibility of a large coverage of information allows to take into account different points of view for analysts.

The next stage of the system development – implementation of event assessment means and tools for integration with the means of data analysis. It is proposed to create the "event series" connecting data changes (for example indicators of social and economic development) with the extracted events. It is able to allow analysts to use more fully information (structured and unstructured), received from various open sources in the Internet.

References

[1] M. Bautin, L. Vijayarenu, S. Skiena, "International Sentiment Analysis for News and Blogs", ICWSM, 2008.

[2] D. Cheney "Text mining newspapers and news content: new trends and research methodologies", in Proceedings of the IFLA World Library and Information Congress, 2013, pp. 24.

[3] M.-A. Mittermayer, G. Knolmayer, "Text Mining Systems for Predicting the Market Response to News: A Survey", Working Paper No. 184, Institute of Information Systems, Univ. of Bern, Bern, 2006.

[4] B. Zhao, S. Vogel, "Adaptive parallel sentences mining from web bilingual news collection", in Proceedings IEEE International Conference on Data Mining, 2002. ICDM 2009, pp. 745-748.

[5] M. Keller, M. Blench, H. Tolentino, C.C. Freifeld, K.D. Mandl, A. Mawudeku, "Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance", in Emerging infectious diseases, vol. 15, №. 5, 2009, pp. 689-695.

[6] K. Radinsky, E. Horvitz, "Mining the Web to Predict Future Events", in Proceedings of the sixth ACM international conference on Web search and data mining, Rome, Italy, 2013. pp. 255-264.

[7] R. Korolov, D. Lu, J. Wang, G. Zhou, C. Bonial, C. Voss, L. Kaplan, W. Wallace, J. Han, H. Ji, "On Predicting Social Unrest Using Social Media", in Proceedings IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, 2016, pp. 89-95.

[8] M. Li, J. Wang, W. Tong, H. Yu, X. Ma, Y. Chen, H. Cai, J. Han, "EKNOT: Event Knowledge from News and Opinions in Twitter", in Proceedings of the AAAI Conf. on Artificial Intelligence, Phoenix, 2016.

[9] J. Wang, T. Wenzhu, Y. Hongkun, L. Min, M. Xiuli, C. Haoyan, T. Hanratty, J. Han, "Mining Multi-Aspect Reflection of News Events in Twitter: Discovery, Linking and Presentation", in Proceedings of the 2015 IEEE International Conference, 2015, pp. 429-438.

[10] W.M.P. van der Aalst, A. Adriansyah, A.K. Medeiros, "Process Mining Manifesto", in BPM 2011 Workshops, Part I. Vol. 99. Springer Verlag, 2012, pp. 169-194.

[11] D. Calvanese, M. Montali, A. Syamsiyah, W.M.P. van der Aalst, "Ontology-Driven Extraction of Event Logs from Relational Databases", in Business Process Management Workshops, 2015.

[12] I. Shalyaeva, L. Lyadova, V. Lanin, "Events Analysis Based on Internet Information Retrieval and Process Mining Tools", in Proceedings of the 10th International Conference on Application of Information and Communication Technologies (AICT). 12-14 Oct. 2016, Baku: IEEE, 2016, pp. 168-172.