STREAM MINING ON ENVRONMENTAL DATA

Maja Škrjanc, Dunja Mladenić
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773900
e-mail: {firstname.secondname}@ijs.si

ABSTRACT

This paper addresses a problem of environmental data analysis, where the data is obtained from different data sources including sensor measurements. We propose a framework for stream mining on environmental data and illustrate its applicability on real-world datasets. Moreover, in collaboration with domain experts two scenarios of data analysis are identified as potentially useful in environmental stream mining.

1 INTRODUCTION

Environmental phenomena share a dynamic nature of chance with the nature itself. With modern technologies we can observe, trace and analyse more and more of dynamic data describing the environmental changes. A sub-field of Knowledge Discovery called Stream Mining [2, 3] addresses the issue of rapidly changing data. The idea is to be able to deal with the stream of incoming data quickly enough to be able to simultaneously update the corresponding models, as the amount of data is too large to be stored: new evidence from the incoming data is incorporated into the model without storing the data. The underlying methods are based on the machine learning methods of on-line learning, where the model is built from the initially available data and updated regularly as more data becomes available.

This paper proposes a framework for stream mining on environmental data based on the identified needs of domain experts working on analysis of environmental data. Two specific scenarios are addressed, one supporting definition of alarm triggers and the other supporting analytic browsing in the context of historic data.

The paper is structured as follows: Section 2 presents the proposed framework to environmental stream mining. Application of the framework on real-world datasets is described in Section 3. Sections 4 gives conclusions and some directions for future work.

2 PROPOSED FRAMEWORK

The proposed framework for environmental stream mining named *EnStreaM* enables the users to analyze structural, i.e., more static data and stream data obtained from sensors. In addition, it supports usage of semantic annotations of the original data if available as part of the input data.

We assume that most of the environmental domain related tasks which include observation and monitoring of environmental phenomena use three main data inputs:

- structured data, which is more or less static during the monitored period of time
- data streams, which are usually sensor data related to the observed phenomena. Sensor data in general is collected from different locations and sensor platforms. They can have different representations of the sensor measurements and can be compliant to different standards (e.g., [4, 7])
- data annotations, annotating the structured data or providing semantic annotations of the sensor preferably using some standard form, (e.g., [5])

The data is describing environmental phenomena and each event is analyzed and monitored through time. To be able to calibrate and re-use appropriate prediction models, end-users need adequate analytical platform to combine structural data and data streams provided as in time series. The proposed framework enables the users to browse through history of observed phenomena by different categories, analyse and define data patterns and have the ability to identify rules for alarm triggering.

We use a general scenario of environmental stream mining for developing the framework. Figure 1 shows the proposed data model for the general scenario. On the top of that based on the discussions with domain experts, two specific scenarios are proposed: definition of alarm trigger and analysis of observed environmental phenomena in a historical context. Architecture of the proposed framework is shown in Figure 2. The stream mining is designed as a Web service that can be easily incorporated in a larger portal for environmental services [6]. The rest of this Section provides a more detailed description of the scenarios.

2.1 Definition of Alarm Trigger

Stream mining methods can be applied in the phase of monitoring and validation of a prediction model, used in a process of monitoring the environmental phenomena. The results enable the user to upgrade the monitoring process with alarm triggering rules. The EnStreaM framework enables monitoring of the model performance and possibly auto-refresh the model when necessary.

When the specified environmental phenomena i.e., event occurs, the expert user runs the existing process of analysis, which, among other results, also provides prediction of how the observed event will evolve. Based on prediction, the experts can select the appropriate response to the phenomena. There are special situations where the expert user needs to re-run the modelling phase of the prediction model. There are two typical situations when the user should refresh the prediction model [1]: the phenomenon is in reality evolving significantly different then the prediction model anticipated (changed event status); the input data for prediction model has significantly changed (e.g. weather forecast is significantly differs from the reality).

In both cases the framework enables the expert user to analyse and monitor the observed event through time and define appropriate actions accordingly. When some of the observed data used for prediction or event status is significantly changed, the user can re-create or update the alarm triggering rules accordingly. The definition of “significant changes” is defined by the user. We anticipated two different settings, which are both captured in proposed framework.

(1) Automatic monitoring the quality of the prediction model. One of the possibilities is that the quality of the prediction models is automatically monitored by comparing the predicted event status with the event status in reality (e.g., satellite image) and if the prediction quality is significantly dropped, the prediction model is updated – re-created with updated input data (e.g., from sensor stream of weather data).

(2) Automatic monitoring the input data streams. The input weather data is monitored and compared to weather forecast data. If the current weather conditions are significantly changed from the weather forecast data used for modelling phase, the prediction model is automatically updated with the fresh input weather data. The threshold for updates of the model depends on the input weather parameters.

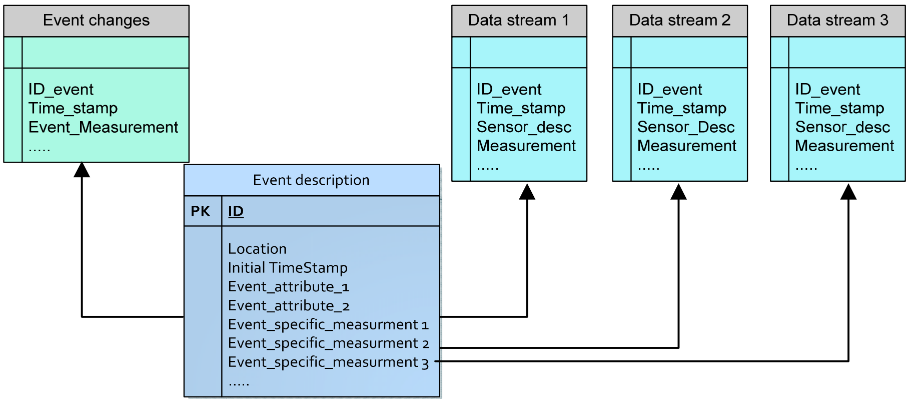

Figure 1: *Data model of general environmental data for the proposed stream mining framework.*

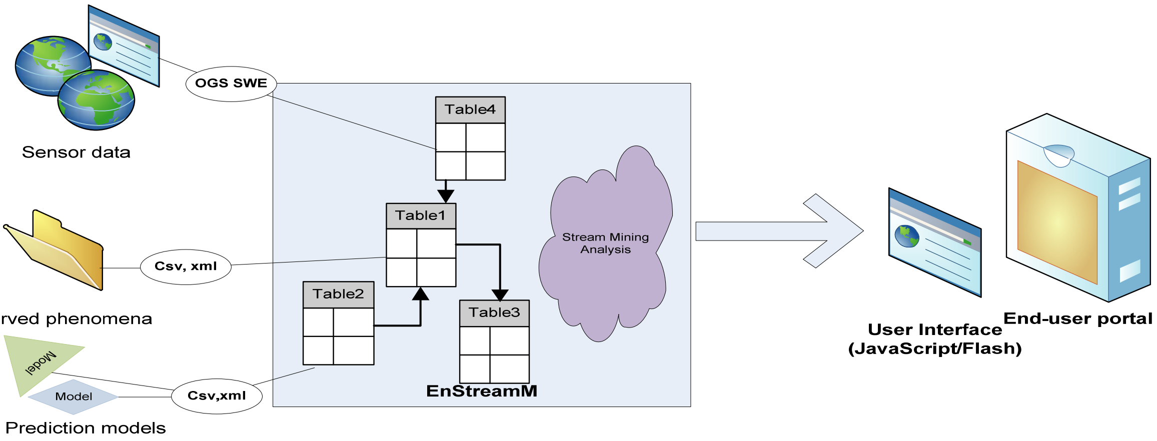

Figure 3: *Architecture of EnStreaM framework.*

2.2 Observed phenomena in a historical context

The proposed second scenario is focused on the empowerment of the user with additional knowledge extracted from advanced analytical browsing of past situations. EnStreaM enables complex data analysis, combining structural data and data streams from different sensors, locations and time. This enables the users to analyse, compare and search for certain phenomena from the past, while at the same time the data related to the observed phenomena will be updated online with incoming sensor data, e.g., from available web services.

In certain situations, the end-user define or calibrate additional input parameters, which are crucial for the quality of the prediction model. To be successful at the modelling phase, the end-user needs to gain expertise by analysing the past situation and appropriate solutions. Data about past situations could include information about predicted models and used user defined parameters. EnStreaM framework enables advanced analytical browsing of similar past situations, which empower the user with additional expertise. Moreover, the user can define what type of similarity is the most important for a given case in a specific situation. When determining the similarity factors, the user is able to search the database of past phenomena and retrieve a set of similar situations from the past. One additional possibility is that at the same time the stream mining capabilities enable the comparison of weather conditions from the online web services with historical values from the similar situations. The available database also includes different types of available predictive models and their parameters.

3 APPLICATION TO REAL-WORLD DATA

The proposed framework was developed by taking into account real-world scenarios as described in Section 2 and two case studies [1]. In this Section we describe the current status of the real-world case studies from two viewpoints:

- process of usage - description of how the user is currently exploiting the workflow process and,
- data availability- data sources that are currently used in case studies' process.

3.1 Oil spill case study

The main goal in the oil spill case study as described in [1] is to predict how the oil spill drifts and its consequences affect the environment. This prediction is the basis for decision making strategies and tactics of responding to oil spills in the sea. Historical data on past situations including weather conditions represents one of the key factors for a successful learning process for experts to create prediction models.

The most common process of predicting oil spills drifts is as follows. Domain expert gets request to create prediction model with the following data: location of the spill (latitude, longitude), amount or rate, oil type, type of spill (on the water, bellow he water), prediction window (in days) – for how many days ahead the customer needs prediction – time period. Domain expert collects the available data: bathymetry - grid of the see ground, grid of the coast line, weather forecast for selected time period (winds, current,etc) (see Figure 2). If the weather data is not available, experts can replace missing data with historical data from the same period of the year (one or several years ago) for the specific location or by a summarized data - calculated averages (e.g., based on sparse historical data) for the specified area. Alternatively, the expert can run the prediction model, which provides among other results 3D time-stamped representation of how the oil is expected to drift and changes of mass balance.

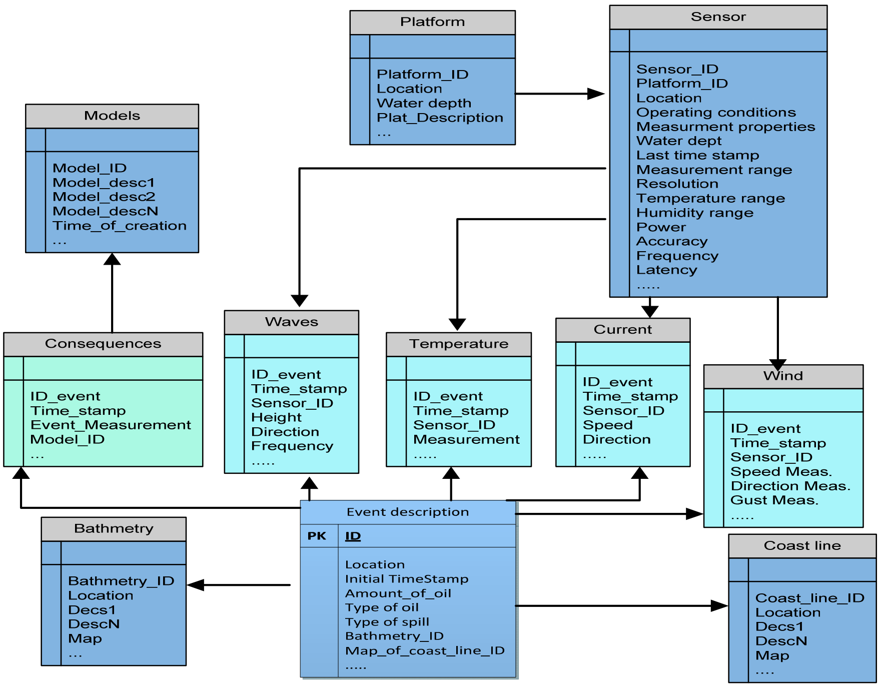

Figure 2: *Oil spill Data Model - adaptation of the general data model (Figure 1).*

Observed events and the corresponding data include initial characteristics of the oils spills, weather forecast data and to some extent dynamic data stream. Location(longitude, latitude), user input (amount, type of spill, oil type), bathymetry - map of the see-ground, coast line map,weather forecast for selected time frame, model outputs. Historical data about past oil spills was not systematically collected and stored, so therefore this data has limited availability. Currently available data does also not include information about the progress of the spill, e.g. data about the drift of the oil spill (at least not in a form required for stream mining analysis, where the data about prediction and "real" situation must be available for several selected time-stamps). This means that it is not possible to compare the "real situation" to predicted situations (predicted oil drift) in a form required for stream analysis.

3.2 Landslide case study

The main goal in the landslide case study is to predict if the selected road (in Guadeloupe) has to be closed for the traffic due to an upcoming landslide. The event is defined as a question: how high is the probability that the landslide will damage the road? The process of calculating the alarm for road closure is composed of sequential composition of five prediction models [1] – the workflow. For all the predicting models the user must specify special input parameters related to certain conditions, such as type of soil, characteristic of the landslide which determines the speed of possible slide propagation or other influential factors for the predicting model in question. By using the existing modelling services the user can obtain a prediction whether the landslide will damage the road.

From the data perspective, the expert user has to define: input data correlated to location, static data sources (elevation model, the geological map, borehole data – data about geological structure, map of geotechnical formations, map of the selected road), dynamic data source correlated to the location and with respect to a specified period of time (precipitation data from the meteorological sensor systems). The user can then create and calibrate prediction models via additional input parameters. Experts empirically determine the suitable parameters, related to certain conditions such as the type of soil, or other influential factors for the predicting model in question. The workflow also provides additional data as intermediate results (prediction models, which are created during the workflow process and are also input for next prediction model): Geological model, Groundwater map model, Landslide probability model, Landslide hazard map, Risk map.

The data sources are divided into input data and output data – which represents results of prediction models in format of raster or vector maps, available also in txt formats (for each measurement unit, appropriate probability is calculated). Some historical information about landslides data is available. The additional available data includes past landslides in Guadeloupe area (more than 50 landslides), described by time stamp, type of slide, how big is the slide impact, damage, etc. Additional data source include the weather sensor data from Guadeloupe area.

We have adjusted the general data model provided in Figure 1 to the specifics of this case study in a similar way as for the Oil spill (omitted due to space restrictions).

6 CONCLUSION AND FUTURE WORK

We have proposed a general framework for environmental stream mining and describe its adaptation to two real-world environmental case studies. Furthermore, we propose two possible scenarios of the framework usage which can be applied in various environmental situations: definition of alarm trigger and analysis of the observed phenomena in historical context. The domain experts found the scenarios potentially useful for the two case studies.

Future work involves implementation and testing of the proposed framework on publicly available environmental data, as well as on real-world data from the two case studies.

7 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and ENVISION (IST-2009-249120).

References

[1] N.R. Bodsberg, U. Bronner, H. Kobayashi, J. Langlois, D. Roman, G. Athanasopoulos; Envision Deliverable 1.1: Report presenting definition of case studies.

[2] J. Gama, M. M. Gaber (eds.), Learning from Data Streams: Processing Techniques in Sensor Networks, Springer Verlag, 2007.

[3] D. Mladenić, M. Grobelnik. 2005, Knowledge Discovery for Ontology construction, In Davies, Studer, Warren (eds.), Semantic web technologies: trends and research in ontology-based systems, Chichester: John Wiley & Sons, 2006, 9-27.

[4] Open Geospatial Consortium Inc. (OGC): OpenGIS Sensor Observation Service (SOS) Implementation Specification. A. Na, and M. Priest (eds.). Available at http://portal.opengeospatial.org/files/?artifact_id=26667

[5] Open Geospatial Consortium Inc. (OGC): OpenGIS Sensor Model Language (SensorML) Implementation Specification, M. Botts and A. Robin (eds.). Available at http://portal.opengeospatial.org/files/?artifact_id=21273

[6] D. Roman, S. Schade, A. J. Berre, N. R. Bodsberg, J. Langlois, 2009, Environmental Services Infrastructure with Ontologies –A Decision Support Framework, In Proceedings of EnviroInfo Conference 2009, Berlin, Germany.

[7] World Wide Web Consortium (W3C): Web Services Description Language (WSDL). E. Christensen, F. Curbera, G. Meredith, S. Weerawarana (eds.).