EuroVis Workshop on Visual Analytics (2017)
M. Sedlmair and C. Tominski (Editors)

PipeVis: Interactive Visual Exploration of Pipeline Incident Data

Z. Sahaf[1] and M. Marbouti[1] and R. Cabral Mota[1] and H. Alemasoom[1] and F. Maurer[1] and M. Costa Sousa[1]

[1]Computer Science Department, University of Calgary, Calgary, Canada

Abstract

The fatal hazards associated with pipeline incidents as well as their frequent occurrence motivate pipeline analysts to learn from historical events and to use that information to prevent future ones by taking proper action. However, the incredible wealth of information contained in pipeline incidents data sets makes it considerably challenging to explore such data. Our solution comprises a visual exploration prototype that aims to help pipeline analysts overcome these difficulties. In this sense, it applies different visual analytical and exploration techniques over the raw pipeline incident data to uncover hidden patterns, unknown correlations, tendencies, and other meaningful information. In that regard, we implemented a prototype which consists of four different views that user can with: a map view that provides spatial information, a chart view that highlights tendencies, a Parallel Coordinates view that discloses hidden patterns and a decision tree view that extracts crucial rules and relations in data.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Human-centered computing]: Visualization/Visualization Application domain/Visual Analytics—

1 Introduction

Major pipelines across the world transport large quantities of crude oil, natural gas and petroleum products. They play a significant role in modern societies and are crucial in providing needed fuels for sustaining vital functions such as power generation, heating, and transportation. In light of hazardous properties of the products being transmitted through these pipelines, a ruptured pipeline has the potential to do severe environmental damage. This research concerns analysis and visualization of pipeline failures with the aim to help domain experts to learn from past failures for future pipeline maintenance and establishments. Therefore, learning from past pipeline failures could help prevent future failures.

Pipeline and Hazardous Materials Safety Administration (PHMSA) [PHM] recently released information about reported hazardous liquid pipeline accidents from 2010 to present. This report revealed details about more than 1,800 accidents that have happened across the United States. Each pipeline incident is reported in an accident report form. This form gathers information regarding different aspects of a pipeline incident including spilled material specification, environmental impact, leak detection information and obvious causes of the incident. Each of these aspects contains a lot of attributes which makes extracting meaningful information from accident reports complicated. In that regard, our proposed visual exploration prototype (PipeVis) utilizes data in accident reports and visually analyses them to extract critical information such as root incident causes and significant trends. In summary, the contributions of this research are as follows:

- Extract domain expert research and requirement questions and map them to visual exploration tasks.
- Design and develop an interactive visual exploration prototype for pipeline incident data.
- Customize popular visual exploration and analytics techniques like map, line chart, PC, and clustering to leverage the domain expert requirements.
- Design and develop a user-friendly visual representation of decision tree that is used for prediction based on history data.

2 Related works

The related work is represented in two reviews: first, current visualization, data mining and statistical approaches for analyzing incidents and more specifically pipeline incidents. Second, we review application of visual exploration techniques for analyzing spatio-temporal data.

2.1 Analysis of incident related data

There are quite a number of studies conducted mainly by pipeline domain experts and researchers on analysis of historical pipeline incidents to identify the risk of pipeline failures. Most of these studies utilize statistical analysis like regression techniques to perform their analysis [ANA12] [Ikp98] [Ndi98]. Researchers also used more complicated methods to investigate pipeline incident data. In a recent study [KST16], a safety assessment model of oil and gas pipeline failure has been provided by incorporating fuzzy logic [SST12] into Bayesian belief network. Using this model, they could identify that construction defect, overload, me-

DOI: 10.2312/eurova.20171116

chanical damage, bad installation and quality of worker are the most significant causes in the oil and gas pipeline failures. The neural network is also another method used by pipeline researchers to predict pipeline failures [SEE*14].

2.2 Visual exploration methods for spatio-temporal data

Studies mentioned above are all among computational or statistical techniques. However, visual exploration techniques can ease the analyzing, representing and understanding of the results. Various visualization approaches have been applied to help analysts explore different data types. Researchers explored visualization techniques on time series data [AMST11] and spatiotemporal data [AA06] [WFR*06]. Andrienko et.al. [AAG03] study provides guidelines for selection of proper exploratory geo-visualization techniques. Chae et al. [CTB*12] propose a spatio-temporal design using visual analysis techniques using social media data. Their approach allows users extract abnormal topic and events using multiple social media resources. AIVis [PBB12] enhances situation awareness by allowing users to explore live and historic videos on a spatio-temporal basis. Some studies combine simulation techniques with visual analysis techniques to help users predict the possible outcomes and make informed decisions [WFR*10]. There are also some studies that use visualization and visual analysis techniques for investigating pipeline related data in emergency situations. For instance, [KKY*06] [MKKS*08] proposes visual analysis frameworks that can visualize and analyze multivariate time-series acquired by pipeline inline inspection instruments. Visualizing decision trees has also been suggested to help make better decisions [TM03] [KWS*14].

Additionally, in this research we deal with the typical concepts for interactive visualization and computational analysis techniques including multiple coordinated views [Rob07], and (semi) automated data analysis approaches [BL10]. The ones that are often combined with InfoVis techniques [NSB04] include: data reduction via sampling or algorithmic feature extraction, clustering [NH06] where data items groups according to their similarity and dimensionality reduction that aims to reduce the data dimensionality. Moreover, Ma et al. [VDEvW11] suggests going a step beyond visual data mining by integrating machine learning into the visual process. Such methods could learn from historical data (or scenarios) and perform prediction for the new data (or scenarios).

3 Objectives

Our main objective in this study is to help analysts to better understand a large set of high dimensional pipeline incident data. More specifically we are following four main goals as explained in the following subsections.

3.1 Interactive visualizations

One of the preliminary objectives of pipeline analysts is to be able to explore, filter and analyze incident related datasets and get more details on demand. According to our regular and iterative discussions with domain experts, the following research questions are elaborated in that regard:

RQ1) which spatial regions have a higher occurrence of incidents?

RQ2) how to have a better understanding of incident properties? - e.g., which properties are more correlated with the incident causes?

RQ3) what are the total costs associated with the incidents?

RQ4) how the spatial disposition of the accidents changes over the years?

3.2 Showing trends

The other fundamental user requirement in pipeline industry is to understand how different aspects of data changes over time. Similar to the previous section, the following research questions are elaborated according to our discussions with domain experts:

RQ5) which substances caused the greatest number of accidents over time?

RQ6) what is the trend of critical parameters over time, and how the trends of different parameters can be compared - e.g., are failure rates going up, going down, or staying about the same? Or how to compare the trends of different failure causes?

3.3 Extracting patterns

Due to the large volume of high-dimensional incident records, gaining insight over characteristics of the dataset is challenging. Analysts need to find similarities and differences of incident records. That′s why one of the primary goals of this study is to help analysts extract patterns and correlations between different attributes. To achieve that, the following research questions have been reached out with the aid of domain experts:

RQ7) how the incident properties are correlated with each other - e.g., are there more failures in larger pipe diameters?

RQ8) what are the common phenomena or hidden information in the pipeline incident datasets?- e.g., what are the common features in all the incidents?

3.4 Find root cause

The last major requirement for pipeline analysts is to be able to discover the root cause of an incident. Therefore, the ultimate goal would be to help analysts analyze history data to predict root cause of future incidents. For this aim, the following research question is specified:

RQ9) how the current incidents data (history data) can help to prevent future incidents? (i.e. how to determine parameters affecting the incident causes and use them to predict future incident causes.)

4 Design rationales and Results

In order to visually answer the research questions in the previous section, a number of design goals is identified to be achieved in our proposed prototype. Our web based visual exploration prototype includes four important views: map view, line chart view, parallel coordinates view, and decision tree view. It also has a control panel that handles the content of views (Figure 1).

4.1 Map visualization

One of the major themes within our research questions is the need for spatiotemporal exploration of incidents (RQ1 to RQ4).

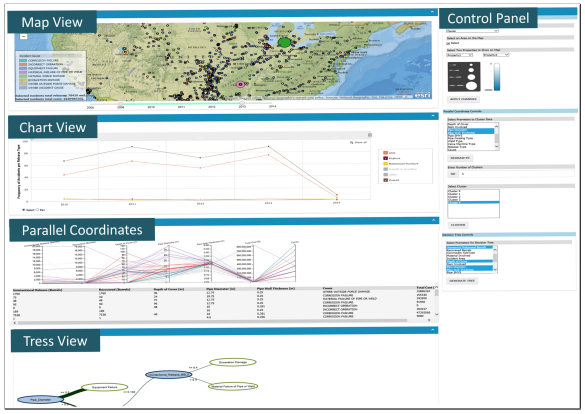

Figure 1: *An overview of PipeVis prototype.*

The interactive map is a common method for representing the spatiotemporal information. It helps analysts gain a geospatial awareness of the incident data. Latitude and longitude are the key elements for showing an incident location on the map. However, different visual channels such as color, shape or size can be used to represent different attributes. PipeVis is designed to use color for representing nominal attributes. For example, various commodity types are shown in different colors. Furthermore, for comparison of two attributes, circle color intensity and size are mapped to a pair of attributes simultaneously (Figure 2). For example, user can compare the cost of an incident (mapped to circle color intensity) along with the amount of spilled commodity (mapped to circle size) for each incident location on the map.

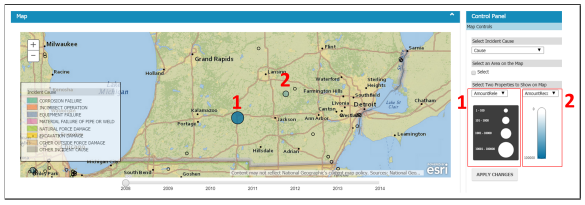

Figure 2: *Visualize pipeline incidents using two attributes.*

4.2 Line chart

In order to get an insight on the trends in the dataset (RQ5, RQ6), an appropriate chart would help analysts to see significant trends in the dataset. Line chart is a useful diagram to visualize trends in time series data. This chart is designed in PipeVis to represent the trend of different attributes over time. Visualized incident points on the map view are designed to be synchronized with the ones used in the line chart view (Figure 3). It means that a region of interest can be selected on the map and a trend for that specific selected region is shown in the line chart view subsequently. Reversely, an area of interest can be selected on the line chart, and then the corresponding incidents are visualized on the map. Among all the available attributes, and based on our discussions with domain experts, the four following different trends are considered in PipeVis:

- *Costs:* costs of incidents over time.
- *Volume:* estimated commodity volume released for an incident over time.
- *Failure Cause:* number of incidents with a specific failure cause over time (Incidents have different failure causes).
- *Part of pipe involved:* number of incidents corresponding to a specific part of pipe over time (in each incident, there is a different part of pipe which is involved).

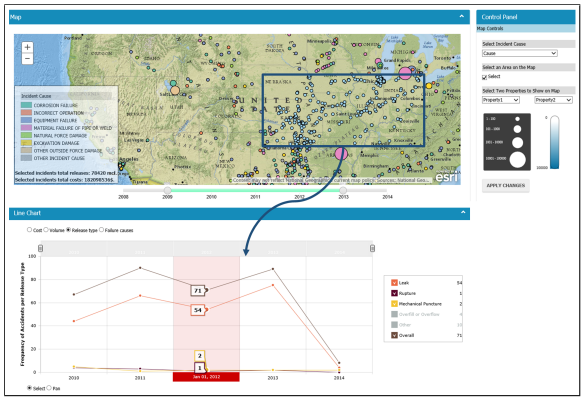

Figure 3: *Synchronization between map and line chart view.*

4.3 Parallel coordinates and clustering

The other major goal of this study is to enable analysts to extract correlation between different attributes (RQ7). In that regard, Parallel Coordinates and clustering have been considered as visual analytics techniques. PC is a promising approach for representing high dimensional data. PC represents each attribute as a vertical line. A record in a dataset is displayed as a series of connected points along the vertical lines (data attributes). Moreover, PC helps analysts determine outliers [MKO*08]. An outlier is a piece of data that is located in an abnormal distance from other points. Furthermore, to enable analysts to extract hidden information (patterns) from data and have a better understanding of the relationship between different attributes, PC is designed to be augmented with clustering algorithm results.

In our developed prototype, clustering can group incidents based on the similarity between their attributes. The incidents in each cluster are shown with the same color on PC. This approach then helps similar incidents recognized among others. The incidents in one cluster are similar in attribute values and therefore have a common behavior known as a pattern. For instance, two main patterns can be seen in Figure 4. High pipe diameter and high pipe wall thickness are related to high released amount (top). However, the average size pipe diameter and pipe wall thickness are related to low release amounts (bottom).

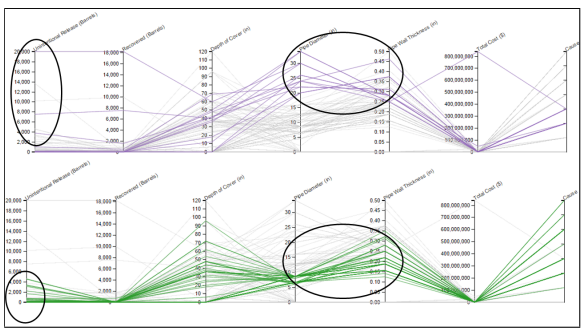

Figure 4: *Visualization of two different clusters on PC.*

4.4 Decision tree

The last but not least benefit of our designed prototype is to utilize historical pipeline incident data for predicting future incidents and more importantly to be able to prevent them (RQ9). For instance, knowing the specification of a pipe and its incident causes in the past, analysts can predict possible causes of failure for a pipe with the similar specification in the future. In this study, we take advantage of the decision tree to do the prediction using historical data. Our decision tree is trained using existent incident data.

Decision tree [Qui86] is a model which predicts the value of one attribute (failure cause in our case) based on the other attributes (pipe diameter, pipe wall thickness, etc.). Each interior node of the tree corresponds to one of the input attributes. Leaf nodes are the possible values of the attribute to be predicted (for example different types of failure cause). Each branch in the tree represents one of the possible alternatives or courses of action available at that node. The tree branches are weighted, where the weights represent the confidence level of an action in that branch. The confidence level is proportional to the number of records satisfying the condition of that branch. For instance, the tree in Figure 5 is generated using the following properties "pipe diameter", "pipe wall thickness", "unintentional released barrels", "depth of cover" and "cause". The important highlighted rule in Figure 5 is:

if *(pipe diameter > 0.5* ***and***

pipe wall thickness > 0.148 ***and***

depth of cover <= 284 ***and***

depth of cover <= 155)

then *by the confidence level of 70% (calculated from the history incident records), the cause is corrosion failure.*

This type of rules could help in the prediction of future causes. For instance, when a new pipe with new specifications (wall thickness and depth of cover) is developed, analysts would use this tree to predict what types of incident causes can occur for that type of pipe. In addition to that, knowing the facts about pipelines with certain specifications, they can maintain existent pipelines with similar specifications before occurring any future incidents.

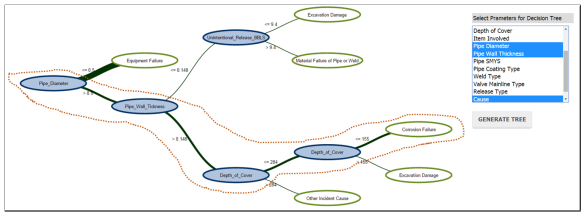

Figure 5: *A representation of decision tree. A sample prediction rule is also highlighted on the tree.*

5 Discussion

Our access to the domain experts was limited, which was not suitable for a complete formal validation. However, we had iterative sessions with a domain expert from a leading industry partner. Since we designed and developed the prototype in an agile process, it allowed for continuous and coherent feedback and also a refinement of the prototype. For example, in the data preparation phase, there were some outliers in the data, some data that were missed, or wrongly measured. Based on their context, the domain expert suggested us to remove missing values by putting a negative value for numerical values and "NONE" for polynomial values. We also received lots of valuable feedback during the development process regarding different views of the system which formed our current design. Finally, we conducted an informal evaluation by demoing the prototype to the same domain expert.

The domain expert provided positive feedback for many of our features. The map visual encodings and interactions were expressive enough so that it is straightforward and understandable without any further explanation. In regards to the trends and line chart view, he didn't want to observe the trends for all combination of properties. For instance, he said "Leak vs. Rupture is one of the important ones that users are usually looking for". Regarding line chart, he also suggested that we connect the line chart with the map using the brushing features to allow seeing trends of the selected spatial points. Initially, PC was not familiar to our domain expert. After introducing PC to him, he found it somewhat useful. He suggested that clustering should be performed on a selective set of attributes instead of all of them. After a set of discussions, we extracted a limited set of attributes for clustering which makes the results more meaningful. The decision tree was less challenging than PC and was much easier for him to understand it.

We also presented our prototype in the IBM Advanced Energy Analytics Competition [†] in which our prototype won the first prize. The judges in the competition were already familiar with PC and decision tree and found them much more interesting and useful than the map view and the line chart. As a comment, they suggested to add the trajectory of the pipeline to the map view, which we do not have access to such data currently.

6 Conclusion

In this paper, we presented PipeVis a visual exploration prototype for analysis of historical pipeline incident data. Our prototype consists of four main views: map view for showing incident locations on the map, line chart which shows incident trends based on different attributes, parallel coordinates view which represents the correlation between properties and clustering results. Finally, we used decision tree to predict future events using the history data. A demo of PipeVis [‡] is prepared which helps to understand our framework functionalities better. Our prototype was iteratively evaluated by pipeline domain experts. Most of the designed tasks are found to be very useful and informative.

7 Acknowledgement

The authors wish to thank Alasdair Clyne from ROSEN Canada Ltd [§] for his feedback on pipeline domain aspects.

† http://www.ucalgary.ca/research/energy-analytics-competition

‡ https://www.youtube.com/watch?v=zuNqMQarZfU

§ http://www.rosen-group.com/

References

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006. 2

[AAG03] ANDRIENKO N., ANDRIENKO G., GATALSKY P.: Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing 14*, 6 (2003), 503–541. 2

[AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of time-oriented data*. Springer Science & Business Media, 2011. 2

[ANA12] ACHEBE C., NNEKE U., ANISIJI O.: Analysis of oil pipeline failures in the oil and gas industries in the niger delta area of nigeria. In *Proceedings of The International Multi Conference of Engineers and Computer Scientists* (2012), pp. 1274–9. 1

[BL10] BERTINI E., LALANNE D.: Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *ACM SIGKDD Explorations Newsletter 11*, 2 (2010), 9–18. 2

[CTB*12] CHAE J., THOM D., BOSCH H., JANG Y., MACIEJEWSKI R., EBERT D. S., ERTL T.: Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 143–152. 2

[Ikp98] IKPORUKPO C.: Environmental impact assessment and human concern in the petroleum industry: NigeriaâĂŹs experience. In *9th International Conference on the Petroleum Industry and the Nigerian Environment* (1998), pp. 766–782. 1

[KKY*06] KOO S. O., KWON H. D., YOON C. G., SEO W. S., JUNG S. K.: Visualization for a multi-sensor data analysis. In *Computer Graphics, Imaging and Visualisation, 2006 International Conference on* (2006), IEEE, pp. 57–63. 2

[KST16] KABIR G., SADIQ R., TESFAMARIAM S.: A fuzzy bayesian belief network for safety assessment of oil and gas pipelines. *Structure and Infrastructure Engineering 12*, 8 (2016), 874–889. 1

[KWS*14] KONEV A., WASER J., SADRANSKY B., CORNEL D., PERDIGAO R. A., HORVÁTH Z., GRÖLLER M. E.: Run watchers: Automatic simulation-based decision support in flood management. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1873–1882. 2

[MKKS*08] MACIEJEWSKI R., KIM S., KING-SMITH D., OSTMO K., KLOSTERMAN N., MIKKILINENI A. K., EBERT D. S., DELP E. J., COLLINS T. F.: Situational awareness and visual analytics for emergency response and training. In *Technologies for Homeland Security, 2008 IEEE Conference on* (2008), IEEE, pp. 252–256. 2

[MKO*08] MUIGG P., KEHRER J., OELTZE S., PIRINGER H., DOLEISCH H., PREIM B., HAUSER H.: A four-level focus+ context approach to interactive visual analysis of temporal features in large scientific data. In *Computer Graphics Forum* (2008), vol. 27, Wiley Online Library, pp. 775–782. 3

[Ndi98] NDIFON W.: Health impact of a major oil spill: Case study of mobil oil spill in akwa ibom state. In *9th international conference on the petroleum Industry and the Nigerian Environment, Abuja* (1998), pp. 804–815. 1

[NH06] NOVOTNY M., HAUSER H.: Outlier-preserving focus+ context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 893–900. 2

[NSB04] NOCKE T., SCHUMANN H., BÖHM U.: Methods for the visualization of clustered climate data. *Computational Statistics 19*, 1 (2004), 75–94. 2

[PBB12] PIRINGER H., BUCHETICS M., BENEDIK R.: Alvis: Situation awareness in the surveillance of road tunnels. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 153–162. 2

[PHM] Pipeline and hazardous materials safety administration. http://phmsa.dot.gov/pipeline. 1

[Qui86] QUINLAN J. R.: Induction of decision trees. *Mach. Learn. 1*, 1 (Mar. 1986), 81–106. 4

[Rob07] ROBERTS J. C.: State of the art: Coordinated & multiple views in exploratory visualization. In *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV'07. Fifth International Conference on* (2007), IEEE, pp. 61–71. 2

[SEE*14] SENOUCI A., ELABBASY M., ELWAKIL E., ABDRABOU B., ZAYED T.: A model for predicting failure of oil pipelines. *Structure and Infrastructure Engineering 10*, 3 (2014), 375–387. 2

[SST12] SHAHRIAR A., SADIQ R., TESFAMARIAM S.: Risk analysis for oil & gas pipelines: A sustainability assessment approach using fuzzy based bow-tie analysis. *Journal of Loss Prevention in the Process Industries 25*, 3 (2012), 505–523. 1

[TM03] TEOH S. T., MA K.-L.: Paintingclass: interactive construction, visualization and exploration of decision trees. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (2003), ACM, pp. 667–672. 2

[VDEvW11] VAN DEN ELZEN S., VAN WIJK J. J.: Baobabview: Interactive construction and analysis of decision trees. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 151–160. 2

[WFR*06] WEAVER C., FYFE D., ROBINSON A., HOLDSWORTH D., PEUQUET D., MACEACHREN A. M.: Visual analysis of historic hotel visitation patterns. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On* (2006), IEEE, pp. 35–42. 2

[WFR*10] WASER J., FUCHS R., RIBICIC H., SCHINDLER B., BLOSCHL G., GROLLER E.: World lines. *IEEE transactions on visualization and computer graphics 16*, 6 (2010), 1458–1467. 2