A Systematic Approach for Discovering Causal Dependencies Between Observations and Incidents in the Health and Safety Domain*

Artem Polyvyanyy[a,*], Anastasiia Pika[b], Moe T. Wynn[b], Arthur H. M. ter Hofstede[b]

[a]*The University of Melbourne, Parkville, VIC, 3010, Australia*
[b]*Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia*

Abstract

The paper at hand motivates, proposes, demonstrates, and evaluates a novel systematic approach to discovering causal dependencies between events encoded in large arrays of data, called *causality mining*. The approach has emerged in the discussions with our project partner, an Australian public energy company. It was successfully evaluated in a case study with the project partner to extract valuable, and otherwise unknown, information on the causal dependencies between *observations* reported by the company's employees as part of the organizational health and safety management practices and *incidents* that had occurred at the organization's sites. The dependencies were derived based on the notion of *proximity* of the observations and incidents. The setup and results of the evaluation are reported in this paper. The new approach and the delivered insights aim at improving the overall health and safety culture of the project partner practices, as they can be applied to caution and, thus, prevent future incidents.

Keywords: Big data, data mining, process mining, proximity of events, causality, health and safety, cause of incidents

1. Introduction

Health and safety (H&S) refers to regulations and procedures intended to prevent incidents, e.g., accidents or injuries, in workplaces or public environments.[1] The failure to control hazards contributes to over half a million work-related injuries and illnesses (annually in Australia), including more than 125,000 serious injury cases [1]. In 2009, the cost of work-related injuries and illnesses to the Australian economy was estimated at about $60.6 billion (4.8% of GDP) [1]. Safe Work Australia[2] reports total economic cost for the 2012–13 financial year to be $61.8 billion (4.1% of GDP for the same period) [2]. Globally, there are 2.3 million deaths annually for reasons attributed to work [3]. In 2012, the costs of work-related injuries and illnesses for the global economy varied between 1.8 and 6.0% of GDP in country estimates, the average being 4% [3].

H&S regulations and procedures are a concern in many industries. Safe Work Australia reports that in the period between 2011–12 and 2014–15, annually, per 1000 employees, serious H&S claims[3] were estimated in the range of 4.7–6.0 claims in the electricity distribution sector, 9.9–12.2 claims in the mining sector, and 22.9–29.9 claims in pipeline and other transport [4]. In the period between 2011–12 and 2014–15, median compensation paid for serious claims in these industries were estimated in the range of $9,900–$14,000 in the electricity distribution sector, $15,600–$17,100 in pipeline and other transport, and $21,300–$26,400 in the mining sector [5].

Organizations strive to decrease H&S risks. They collect and analyze information about internal H&S *incidents* that have occurred in the past with the ultimate goal of acquiring an understanding of their causes and, consequently, preventing them in the future. This information is often scattered across the mass of Big Data [6] stored in organizations' information systems, i.e., voluminous and complex data sets that often miss explicit semantic relationships between individual data items, written documents, knowledge of employees, and aggregate reports. In this paper, we collectively refer to this information as the *universe of (H&S) data*.

To derive valuable insights from the universe of data, organizations resort to *data mining* [7] and *process mining* [8] techniques. Data mining is the process of discovering patterns, i.e., reoccurring dependencies between entities, in large data sets. Data mining techniques are often domain agnostic and, thus, not tailored for the discovery of process-related entities and constraints, like *events* of incidents and their *causalities*. Several existing approaches use data mining to discover correlations

*Results reported in this paper were obtained in a project that meets the requirements of the National Statement on Ethical Conduct in Human Research (2007) of National Health and Medical Research Council, Australia, and has been granted an approval on behalf of the University Human Research Ethics Committee, Queensland University of Technology, Ref. No.: 1700001025.

*Corresponding author. The reported results were obtained while the corresponding author was with the Queensland University of Technology, while the paper was written while with the University of Melbourne.

Email address: artem.polyvyanyy@unimelb.edu.au (Artem Polyvyanyy)

[1] https://en.oxforddictionaries.com/definition/health_and_safety

[2] Safe Work Australia is an Australian Government agency. Its functions include improvement of work health and safety and regulation of workers' compensation arrangements across Australia.

[3] A serious claim is an accepted workers' compensation claim that involves one or more weeks away from work and excludes all fatalities, and all injuries and diseases experienced while traveling to or from work or while on a break away from the workplace.

between known factors captured in descriptions of past incidents [9–12], e.g., most of the fatalities occur on rainy summer days between 7:00 am and 11:00 am. However, these approaches cannot identify unknown factors and events that have caused or contributed to the incidents. Process mining builds on data mining and process model-driven approaches [8]. Process mining techniques analyze event logs recorded by information systems that capture the history of executed processes to identify reoccurring patterns of events. However, process mining techniques are limited in their abilities to recognize process information, e.g., events and their dependencies, recorded in unstructured arrays of data.

The paper presents the results of a project with an Australian public energy company with its core business in natural gas exploration, electricity generation, and energy retailing, on developing an approach for automatic discovery of causal dependencies between *observations* reported by the company's employees as part of the organizational H&S management practices and *incidents* that occurred at the organization's sites. The company has a strong reporting culture; as a result, thousands of reports are created by employees. Analyzing huge amounts of data associated with these reports presents a challenge to the company. Hence, the business objective of the project was to devise an approach for extracting insights into incident prevention from the company's H&S data. This objective was translated into a more specific data mining task: to devise an approach for mining causalities between events recorded in the H&S data, which we will refer to as *causality mining*. The approach is proposed as an adaptation of Cross-Industry Standard Process for Data Mining (CRISP-DM) [13], which is the de facto standard for developing data mining projects [14].

The core contributions of this paper are summarized below:

- A model for Causality Mining for discovering event causalities based on the notion of proximity.
- An evaluation of the proposed proximity model in a case study with the project partner aimed to discover causalities between H&S observations and incidents.

Section 2 is devoted to the discussion of the approach used. Section 3 describes a developed model of H&S data. Sections 4 and 5 exploit the notion of proximity to propose a causality model for discovering causal dependencies between events in H&S data. Section 6 reports on an evaluation of the proposed approach through a case study. Section 7 summarizes related work, whereas Section 8 summarizes the paper's contributions and discusses avenues for future work.

2. Approach

Data mining refers to "the process of discovering interesting patterns and knowledge from large amounts of data" [7]. CRISP-DM is the de facto standard for conducting data mining projects [14]. Our approach for causality mining from the universe of H&S data is an adaptation of CRISP-DM. We introduce CRISP-DM in Section 2.1 and describe our proposed approach in Section 2.2.

2.1. Cross-Industry Standard Process for Data Mining

CRISP-DM is a non-proprietary, documented data mining approach developed by industry leaders which provides a blueprint for data mining projects [13]. The model encourages best practices and helps organizations to conduct faster projects that can be replicated and lead to better results.

CRISP-DM breaks down the data mining approach into six phases [13]:

1. **Business understanding.** This is arguably the most important phase of a data mining project. The main tasks of this phase include the acquiring of understanding of business objectives, the definition of data mining goals aligned with these objectives, the evaluation of resources available for the project and the development of a plan which would allow to achieve the business objectives with the available resources.
2. **Data understanding.** This phase is concerned with the initial data collection, the description of the collected data (e.g., describing data attributes and the number of records), the initial data exploration (e.g., using querying or data visualization to understand properties of the data), and the evaluation of the data quality (e.g., the data completeness).
3. **Data preparation.** During this phase of the data mining project the collected data is transformed into a form suitable for data mining methods. This may include the selection of appropriate data sets, records and attributes, the derivation of new attributes, combining data from different sources, data cleaning (e.g., tackling missing values) and data formatting.
4. **Modeling.** During the modeling phase different data modeling methods are considered (e.g., the use of neural networks or decision trees), calibrated and tested. An appropriate model is selected based on initial assessment results (e.g., a model with the highest prediction accuracy).
5. **Evaluation.** The model built during the modeling phase is evaluated against the business objectives identified during the first phase of the project. The evaluation is usually conducted in collaboration with domain experts. Based on the evaluation results, the project manager decides whether to proceed to the deployment phase.
6. **Deployment.** The knowledge acquired through the application of data mining during previous phases of the project has to be presented in a form that can be used by the organization. This may involve the preparation of reports or implementation of the developed model in the organization.

The data mining process or any of its phases can be repeated if required. For example, there may be a need to return to the data preparation phase after the modeling phase if a selected model requires a different data format.

2.2. Approach used in this study

We adapted the CRISP-DM methodology to explicitly take into account the viewpoint of H&S processes involved (e.g., reporting of observations, monitoring of incidents, review of incidents) throughout the various phases. We then designed a

novel causality mining approach for H&S data informed by the interviews with the H&S domain experts. We now detail each phase of the proposed approach depicted in Figure 1.

1. CRISP-DM1: Business and *process* understanding.
   The main tasks of this phase include (i) developing an understanding of business objectives, and (ii) the definition of data and process mining goals aligned with these objectives. A detailed understanding of business processes involved can be achieved through multiple interactions with stakeholders.
   We conducted multiple meetings and unstructured interviews with H&S experts from the case study company who have the knowledge of the company's H&S processes and information recorded about these processes by information systems. This allowed us to identify a major challenge the company faces, i.e., how to extract insights into incident prevention from the company's H&S data, and to formulate the project objective, i.e., devising an approach for discovering causal dependencies between observations and incidents recorded in the data.
2. CRISP-DM2: *Process* data understanding.
   Identifying the domain of discourse, i.e., objects, relations between objects, and constraints that relate to the events in the domain is an essential first step. This is followed by the initial data collection, data exploration and evaluation of data quality. In order to facilitate an initial understanding of data, we also propose to construct a conceptual data model of the domain.
   The conceptual model proposed in this project extends the model proposed in [15], which describes the basic objects , e.g., *events*, and their relations, e.g., *causality*. In collaboration with the case study partner, we learned what information about observations and incidents is reported by employees and recorded by the company's information systems. We created a conceptual model of this information using the Object Role Modeling notation [16]. H&S information recorded by the company and the developed conceptual model are discussed in Section 3.
3. CRISP-DM3: *Process* data preparation.
   The main steps in the process data preparation stage include populating the conceptual model from the universe of data and discovering the missing elements in the population. We propose to refine the conceptual data model developed as part of the process data understanding phase based on a better understanding of the process through a careful analysis of observed instances. Hence, we advocate multiple iterations between these two phases to ensure the quality of the resulting conceptual data model. The data was collected by the case study partner from the company's information systems which record H&S information. We selected relevant data attributes (described in Section 3) and organized them into three data sets containing information about observations (1), incidents (2) and the organizational structure (3), i.e., relationships between departments, sub-groups, groups, business units and divisions.
4. CRISP-DM4: Modeling.
   This modeling phase was adapted to include not only the application of existing data modeling methods or process mining methods but also the development of new methods should they be necessary. For this project, there is a requirement to formulate hypotheses about causal dependencies between events in the domain and to confirm causal dependencies between events in the population by testing the hypotheses. We adapted a method to discover missing pieces of information in the conceptual schema and propose a model based on the notion of proximity in Section 5. We also recognized the need to iterate between the modeling phase and the evaluation phase to improve the outcomes from the proposed technique.
5. CRISP-DM5: Evaluation.
   The findings from the hypotheses in Phase 4 are then confirmed in collaboration with domain experts. This phase is expected to be repeated multiple times after the parameters of the proposed model in Phase 4 are adjusted and the hypothesis testing is repeated. The results of this evaluation are further explained in Section 6.
6. CRISP-DM6: Deployment.
   The findings are then presented in a form that can be used by the organization. This may include the implementation of the developed model and preparation of reports. The deployment informs the learning and enhances the business and process understanding.

3. Data Modeling

Organizations store information about H&S incidents such as incident descriptions, incident investigation results and known factors that contributed to the incidents (e.g., a lack of training or poor weather conditions) [9–12]. Many organizations also use Occupational Health and Safety Management Systems (OHS MS) which record vast amounts of H&S data including "safety observations . . . and near miss reporting" [17].

In order to better understand the nature of such H&S data, we advocate creating a conceptual model of the data, for example using the Object Role Modeling (ORM) notation [16]. ORM allows a modeler to create a graphical representation of object types in a universe of discourse (e.g., H&S events), define relationships between them as well as any related constraints. We used ORM to model H&S data recorded by the case study company.

In order to minimise H&S risks, the case study company encourages employees to report observations about any events that could help prevent incidents. For example, such observations may include information about equipment malfunctions, unusual site conditions or employee malpractices (e.g., not using personal protective equipment (PPE)). Another type of reporting is concerned with incidents that occurred, e.g., an equipment breakdown that resulted in downtime or an employee injury.

H&S data collected by the case study company included information about 53,094 observations and 3,187 incidents that spanned 428 days (observation and incident records were stored

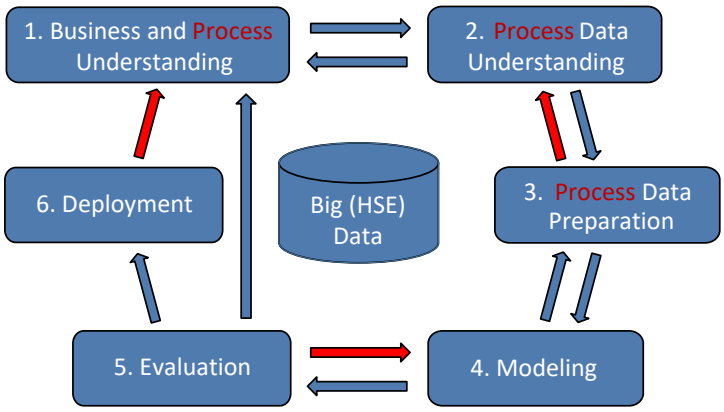

Figure 1: Relationships between the different phases of the proposed approach to Causality Mining.

in separate data sets). Each observation and each incident is described by a set of attributes. In the study, we used the following incident attributes: *ID*, *date of occurrence*, *time of occurrence*, *sign off date*, *department*, *physical location*, *specific location*, *summary*, *description*, *lessons learned* and *findings*; and the following observation attributes: *ID*, *date of occurrence*, *sign off date*, *department*, *physical location*, *specific location*, *summary*, *safe activity description* and *unsafe activity description* (relevant observation and incident attributes were selected in discussions with the company H&S experts). In addition, a separate data set provided information about relationships between different organizational units: *departments*, *sub-groups*, *groups*, *business units* and *divisions*. The values of attributes *date of occurrence*, *time of occurrence* and *sign off date* as well as the values of organizational units are structured; the values of all other observation and incident attributes are unstructured text descriptions entered by employees.

We created a conceptual model of this information using ORM. Figure 2 depicts a fragment of the developed ORM model of the H&S data which was subsequently used to discover causal dependencies between observations and incidents. The model describes two types of events recorded in the H&S data: each event is either an observation or an incident (the exclusion constraint is depicted by a circled "x" symbol in Figure 2). Each event is associated with a department in the organization and has a summary recorded as text, the date of occurrence and the location in which the event occurred (these mandatory roles are indicated by dots in Figure 2). Moreover, each event only has one summary, location, department and date of occurrence (the uniqueness constraints are indicated by bars placed above the roles in Figure 2). An event may also have a sign off date, i.e., the date when it was reviewed and acted upon if required. Each location is specified by a physical location and may have a specific location which is a further specification of the physical location. Each department is related to a sub-group which is a part of a group. A group is a part of a business unit which is a part of a division. Each incident also has the time of occurrence and a description recorded as text. An incident may also have findings and lessons learnt (recorded as text): if one value is recorded for an incident then another value is also recorded (the constraint is depicted by a circled equality symbol in Figure 2). Observations may also have safe or unsafe activity descriptions which provide further details about observed activities.

The model allowed us to unambiguously specify H&S objects (e.g., events, observations and incidents) and relationships between them (e.g., observations can cause incidents). Causal dependencies between observations and incidents are, however, not known. Sections 4 and 5 describe our proposed approach to the discovery of such dependencies based on the analysis of available H&S data.

4. Proximity Model for Causality Mining

In our approach, *proximity*, which is defined as "nearness in space, time, or relationship"[4], is used as a measure for causality. Causality is an important notion that has been examined in quite a few disciplines [18]. As pointed out by Karimi [18], "understanding causal relations" can allow one to "predict the future", and in case one can modify certain variables in the present one may even be able to "exert control on the behavior of the system". While there are different views on causality, our focus is on the prediction of future events, and as such it is natural to think of causes and effects as having temporal and spatial connections. In [19] (p. 9), for example, this is expressed through the causality "attributes" of *antecedence* and *contiguity*, where "[a]ntecedence postulates that the cause must be prior to, or at least simultaneous with, the effects" and "[c]ontiguity postulates that cause and effect must be in spatial contact or connected by a chain of intermediate things in contact".

In the context of our approach, we aim to determine what the *likelihood* is of observations contributing to incidents based on past experience. Predictions are thus probabilistic calculations derived from historic data. These calculations take into account the level of confidence in causal relationships between observations and incidents (or between observations). The level of confidence is a measure of the strength of the causal relationship and is determined by the *duration* of the interval between observation and associated incident and by the *distance* of the locations. The notion of distance is complex as multiple aspects of distance can play a role, both for *virtual* locations (which can

[4] https://en.oxforddictionaries.com/definition/proximity

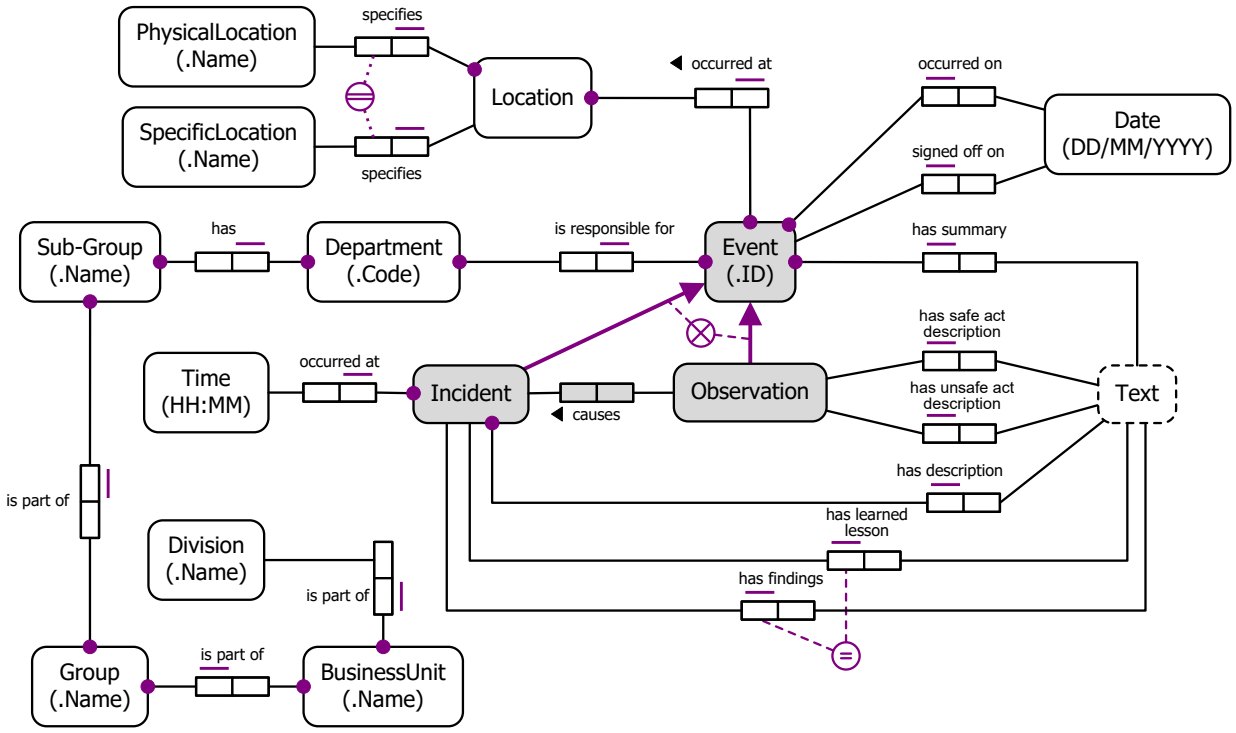

Figure 2: A fragment of the conceptual model of the H&S information collected at the case study company (captured using ORM notation).

be, for example, a position or role on an organizational chart) and for *physical* locations (a distance notion in this context could be influenced by the layout of a building). Temporal relations can be determined by periodicity (e.g., certain events may occur every first Tuesday of the month or every night); however, periodicity is not considered in our model.

There is another aspect that may strengthen the link between observations and incidents and this may be abstracted in the concept of *similarity*. Similarity may relate to textual descriptions of observations and incidents or may concern locations where observations and incidents took place. In the latter case, locations may share certain environmental features or organizational characteristics and thus be considered to be more similar.

All together, the notions of distance in time and space and the notion of similarity act as determinants for causality. In our approach, the *closer* events are in time, space, and the more similar they are, the higher the expected degree of casuality. Hence, *proximity* becomes a measure for causality. For each of the constituent dimensions (time, space, and similarity), the notion of proximity is subjective but needs to be precisely defined. A proximity dimension may be captured through a number of different definitions, each focussing on different aspects of the dimension, noting that such aspects may take characteristics of the application domain into account. All these definitions across the three dimensions need to be weighted to get an overall proximity score. In the next section, a concrete proximity model is proposed in the context of the domain of health and safety.

5. Proximity Modeling in the H&S Domain

In discussions with the project partner and based on the constructed conceptual model of the H&S data (described in Section 3) we formulated several hypotheses about causal dependencies between observations and incidents based on their similarity and proximity in time and space. It was an iterative process which involved several brainstorming sessions with the company H&S experts during which a number of hypotheses were considered. As a result of these discussions, six hypotheses were selected, implemented and evaluated. Let E be the set of all H&S events, $O \subset E$ be a set of observations, and $I \subset E$ be a set of incidents. Let e_a denote the value of attribute a of event $e \in E$; for example, $i_{summary}$, where $i \in I$, denotes the value of attribute *summary* of incident i. We describe the formulated hypotheses and a way to operationalize them in Sections 5.1–5.3. Note that all the subsequently presented sample instantiations of the proposed proximity measures emerged in the discussions with the stakeholders.

5.1. Proximity in time

As discussed in Section 4, causes must precede or happen simultaneously with the effects. In the context of H&S data, for example, an observation which reports a leak from a piece of equipment would precede the equipment breakdown incident. Moreover, such events would likely happen close in time. Hence, our first hypothesis captures this relationship.

> *Hypothesis 1 "Event Ordering":* An incident that occurred shortly after an observation may be related to the observation: the closer the observation date to the incident date the stronger the relationship.

Proximity measure: Each observation and each incident has the date of occurrence. For a given observation $o \in O$ and a given incident $i \in I$, the proximity (denoted by $TP_{eo}(o,i)$) is measured as an exponential function of the number of days between the observation date and the incident date. Let $days(o,i)$ be a function that returns the number of days between $o_{date\_of\_occurrence}$ and $i_{date\_of\_occurrence}$:

$$TP_{eo}(o,i) = \begin{cases} e^{-1 \times days(o,i)/10}, & \text{if } days(o,i) \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

It was observed in practice that observation reports can be created at later dates. For example, an employee who observes an H&S issue may need to attend to other urgent tasks and creates a report of the observation a few days later. Thus, an observation may be reported after the date of a related incident. Our second hypothesis reflects this issue of unreliable event timestamps.

Hypothesis 2 "Fuzzy Event Ordering": An observation whose date is after the date of an incident may be related to the incident; the closer the incident date to the observation date the stronger the relationship.

Proximity measure: Each observation and each incident has the date of occurrence. For a given observation $o \in O$ and a given incident $i \in I$, the proximity (denoted by $TP_{feo}(o,i)$) is measured as an exponential function of the number of days between the incident date and the observation date.

$$TP_{feo}(o,i) = \begin{cases} 0.5 \times e^{-0.5 \times days(i,o)/5}, & \text{if } -10 \leq days(o,i) < 0 \\ 0, & \text{otherwise.} \end{cases}$$

H&S observations reported by employees are regularly reviewed by the company H&S specialists and acted upon if required. When an issue reported in an observation is resolved the observation is signed off and the sign off date is recorded. Hence, our third hypothesis reflects the idea that incidents that happened after an observation sign off date are unlikely to be caused by the observation.

Hypothesis 3 "Event Sign off": If the date of occurrence of an incident is after the sign off date of an observation, then the observation and the incident are not related.

Proximity measure: Each observation and each incident has the date of occurrence. An observation may also have the sign off date. For a given observation $o \in O$ and a given incident $i \in I$, if the incident date is after the observation sign off date the proximity (denoted by $TP_{eso}(o,i)$) is *zero*, otherwise it is *one*.

$$TP_{eso}(o,i) = \begin{cases} 0, & \text{if } days(o_{sign\_off\_date}, i_{date\_of\_occurrance}) > 0 \\ 1, & \text{otherwise.} \end{cases}$$

5.2. Proximity in space

Another property of causality discussed in Section 4 is contiguity which refers to the spatial connection between causes and effects. Our fourth hypothesis considers physical locations of H&S events (e.g., sites or buildings where observations and incidents happened).

Hypothesis 4 "Location proximity": Observations and incidents that occurred at the same or nearby locations may be related.

Proximity measure: Each observation and incident is associated with a location. As described in Section 3, each location is specified by a physical location (e.g., a street address) and may have a specific location (e.g., a level or a room). Physical and specific locations are specified as text and employees reporting observations and incidents may use slightly different descriptions when referring to the same location. Hence, we use string similarity to compare event locations. For a given observation $o \in O$ and a given incident $i \in I$, the location proximity (denoted by LP) is measured as the weighted average string similarity of their respective physical and specific locations. Let $sim(text_1, text_2)$ be a function that returns string similarity of two text values (e.g., cosine similarity can be used or other methods discussed later in this section) and let w_{ph_l} and w_{sp_l} be given weights to use for the physical locations and specific locations, respectively:

$$LP(o,i,w_{ph\_l},w_{sp\_l}) = w_{ph\_l} \times sim(o_{physical\_location}, i_{physical\_location}) + w_{sp\_l} \times sim(o_{specific\_location}, i_{specific\_location}).$$

While the previous hypothesis considers physical and specific locations of observations and incidents, our fifth hypothesis targets their virtual locations, more specifically, locations in the organizational structure. The likelihood of a causal dependency between an observation and an incident increases if the events are associated with the same organizational unit or with units which are close in the organizational structure.

Hypothesis 5 "Organizational proximity": Observations and incidents reported by the same unit or nearby units in the organizational structure may be related.

Proximity measure: As described in Section 3, each event is associated with a department which is related to a sub-group. Sub-groups are contained in groups that are contained in business units that are contained in divisions, i.e., the organizational model is a tree-like structure. For a given observation $o \in O$ and a given incident $i \in I$, the organizational proximity (denoted by $OP(o,i)$) is a function of the position of the Lowest Common Ancestor (LCA) in the organizational hierarchy. A deeper LCA suggests a stronger relationship, i.e., the strongest relationship is between two events associated with the same department. Let $lca(o,i)$ be a function that returns the LCA of the organizational units that reported o and i in the organizational structure of the project partner:

$$OP(o,i) = \begin{cases} 1, & \text{if } lca(o,i) = \textit{department} \\ 0.5, & \text{if } lca(o,i) = \textit{sub-group} \\ 0.25, & \text{if } lca(o,i) = \textit{group} \\ 0.125, & \text{if } lca(o,i) = \textit{business-unit} \\ 0.0625, & \text{if } lca(o,i) = \textit{division} \\ 0.03125, & \text{otherwise.} \end{cases}$$

5.3. Event similarity

Finally, our sixth hypothesis considers semantic similarity of H&S events. If an observation and an incident share common vocabulary in their descriptions (e.g., refer to the same type of activity or equipment), then the probability that these events are related increases.

Hypothesis 6 "Event similarity": Observations and incidents with similar descriptions may be related.

Proximity measure: As discussed in Section 3, each observation and incident has a summary specified as text. Additional information about events may be provided in other text attributes, e.g., in descriptions, findings or lessons learned of incidents or in safe or unsafe activity descriptions of observations. Similarity of a given observation $o \in O$ and a given incident $i \in I$ (denoted by $ES(o,i)$) is measured as the maximum similarity of their text descriptions provided in any of the attributes specified above, e.g., in summaries or findings (we calculate pairwise similarities and select the maximum value). This can be achieved using text similarity techniques. Let $desc(e)$ denote the set of all attributes of event $e \in E$ which constitute textual descriptions of the event. For example, according to Figure 2, for an incident $i \in I$, it holds that $desc(i) = \{summary,\ description,\ findings,\ lessons\ learned\}$, and for an observation $o \in O$, it holds that $desc(o) = \{summary,\ safe\ activity\ description,\ unsafe\ activity\ description\}$. Then, the event similarity proximity measure can be captured as:

$$ES(o,i) = \max_{d_o \in desc(o), d_i \in desc(i)} sim(d_o, d_i).$$

A number of text similarity approaches have been proposed in the literature [20] that can be used to compare text descriptions of events. Our approach currently supports cosine similarity, Jaro-Winkler similarity and an ontology-based text similarity approach [21], while other approaches can be easily integrated.

5.4. Overall proximity

The overall proximity (denoted by Pr) of a given observation $o \in O$ and a given incident $i \in I$ is a function of specific proximities described in Sections 5.1-5.3. Let w_{ph_l} and w_{sp_l} be weights of the physical and specific locations, and let c_{op}, c_{lp}, c_{tp} and c_{es} be coefficients of the organizational proximity, location proximity, time proximity and event similarity, respectively. Then, the overall proximity can be defined as follows:

$$\begin{aligned} Pr(o,i,w_{ph\_l},w_{sp\_l},c_{tp},c_{lp},c_{op},c_{es}) = \\ c_{tp} \times (TP_{eso}(o,i) \times TP_{eo}(o,i) + TP_{feo}(o,i)) + \\ c_{lp} \times LP(o,i,w_{ph\_l},w_{sp\_l}) + \\ c_{op} \times OP(o,i) + c_{es} \times ES(o,i). \end{aligned}$$

Thus, a high value of the overall proximity between two events should signify a high likelihood of a causal dependency between them. For example, there may be a high probability of a causal dependency between an observation and an incident with similar summaries that were reported by the same department, happened on the same day at the same location.

In the case study, the values of the coefficients were defined in discussions with the project partner (for details see Section 6.1). A direction for future work is the use of machine learning to learn these values from data.

6. Case Study

This section reports the results of the case study with the project partner and discusses several examples of identified causalities between observations and incidents.

6.1. Configuration and results

We applied our approach to the collected H&S data and discussed discovered causal dependencies between observations and incidents with H&S experts from the case study company. The process was repeated using different coefficients in the overall proximity formula and different text similarity methods.

We first executed our approach using all observations (53,094) and incidents (3,187). The following coefficients and weights were defined in discussions with the industry partner: $c_{tp} = 1$, $c_{lp} = c_{es} = 0.8$, $c_{op} = 0.4$, and $w_{ph\_l} = w_{sp\_l} = 0.5$. In the first iteration, we considered observation and incident summaries (i.e., $desc(o) = \{summary\}$ and $desc(i) = \{summary\}$) and used cosine similarity to measure event similarity.

More than 169 million observation-incident combinations were checked and ranked based on their overall proximity scores (53,094 observations and 3,187 incidents yield 169,210,578 combinations). We then inspected 200 observation-incident combinations with the highest proximity scores (ranging from 2.05 to 2.71): 20 out of 200 combinations were found interesting or somewhat interesting. The company H&S experts commented that the approach works well for process safety incidents (examples 2 and 3 described in Section 6.2 were identified during the first iteration of the analysis); however, it does not identify interesting insights for incidents that resulted in injuries. They asked us to modify the approach configuration and repeat the analysis for a subset of injury-related incidents.

The second round of analysis was performed for all observations (53,094) and 582 injury-related incidents (identified by the H&S experts). To measure event similarity of the observations and the incidents, we considered all attributes which provide text descriptions of the events (i.e., for a given incident $i \in I$, $desc(i) = \{summary,\ description,\ findings,\ lessons\ learned\}$, and for a given observation $o \in O$, $desc(o) = \{summary,\ safe\ activity\ description,\ unsafe\ activity\ description\}$). Text similarity was measured using an ontology-based text similarity approach [21, 22]. The following values of the coefficients and weights were used during this round of analysis: $c_{tp} = 0.5$, $c_{lp} = 0.8$, $c_{op} = 0.2$, $c_{es} = 1.5$, $w_{ph\_l} = 7/8$ and $w_{sp\_l} = 1/8$; the values were adjusted after the first iteration in the discussions with the domain experts.

In total, 30,900,708 combinations of observations and injury-related incidents were ranked based on their overall proximity scores. We checked 100 observation-incident combinations with the highest proximity scores (ranging from 1.89 to 2.99): 19 combinations were identified by us as potentially interesting. An H&S expert from the partner company noted that the analysis (with the modified configuration) helped to identify some interesting causalities for injury-related incidents (e.g., examples 1

and 4 described in Section 6.2) and that some other identified causalities describe the same events; for example, when an employee records an event as an observation and then creates an incident for the same event (although such events are technically related, they do not provide interesting H&S insights). The use of an ontology-based method for measuring text similarity, in particular the one reported in [21, 22], contributed to a better result for the injury-related incidents. It was noted by the company H&S expert that process safety incidents are often described using common terms and phrases shared by employees (hence, it was possible to identify process safety causalities using a simple text similarity measure such as cosine similarity), while there is no shared vocabulary used to describe injury-related incidents (hence, an ontology-based method performed much better than cosine similarity).

6.2. Examples of identified causal dependencies

In this section, we describe four examples of causal dependencies identified with the help of our approach and confirmed by the industry partner. The examples are anonymised. We show text descriptions of events reported by employees; however, words which refer to locations and equipment identifiers are replaced with the text "[Location anonymised]" and "[Equipment ID anonymised]", respectively.

Table 1: Example 1: "Systemic issue"

1	**Summary of observation 1**	"TOILET PAPER LEFT ON THE FLOOR"
	Incident summary	"Employee slipped on water and fell on the floor in the female toilet"
	Time proximity	Observation 1 was reported 4 days before the incident.
	Organizational proximity	Both events were reported by the same department.
	Location proximity	Both events happened at the same physical location, text descriptions of the specific locations are different.
2	**Summary of observation 2**	"WATER ON FLOOR"
	Incident summary	"Employee slipped on water and fell on the floor in the female toilet"
	Time proximity	Observation 2 was reported 3 days after the incident.
	Organizational proximity	Both events were reported by the same business unit.
	Location proximity	Both events happened at the same physical location, text descriptions of the specific locations are different.

Table 1 shows two discovered causalities associated with the same incident. The events report either paper (observation 1) or water on the floor (observation 2 and the incident) identified at the same physical location within a few days. In one case an employee slipped on water and fell (reported in the incident). An H&S expert from the case study company commented that these causalities show that there is a general housekeeping issue at the location which could lead to an employee injury. The example demonstrates that causalities could help to uncover systemic issues. Once such issues are addressed, this could lead to incident prevention.

Table 2: Example 2: "Early warning"

Observation summary	"The compressor bund has a fair amount of oil in it from the compressor oil pump leak and needs to be pumped out"
Incident summary	"[Location anonymised] Field Compressor [Equipment ID anonymised]. Identified Low-Low level shutdown switch for Compressor Oil was installed incorrectly and would never alarm in a real low oil situation, which could potentially cause major engine damage"
Time proximity	The observation was reported 2 days before the incident.
Organizational proximity	Both events were reported by the same department.
Location proximity	Both events happened at the same physical location, text descriptions of the specific locations are different.

Table 2 describes an observation which reports an oil leak from a compressor and an incident at the same physical location reported two days later by the same department which describes a problem with the installation of a switch in a compressor (the switch "would never alarm in a real low oil situation, which could potentially cause major engine damage"). An H&S expert from the partner organization noted that both events describe issues identified in similar pieces of equipment at the same location within a few days and that identification of such problems "could lead to ad hoc maintenance to identify similar issues". This example shows that causalities could help to identify events which provide early warnings and thus could help to prevent serious incidents (e.g., major engine damage).

Table 3: Example 3: "Incident investigation"

Observation summary	"PTA performed prior to isolating compressors and topping up oil."
Incident summary	"[Equipment ID anonymized] oil leaking from drivehead gearbox splattering oil to grade."
Time proximity	The observation was reported 3 days before the incident.
Organizational proximity	Both events were reported by the same department.
Location proximity	Both events happened at the same physical location, text descriptions of the specific locations are different.

Table 3 shows another example of an oil leak report (in the incident). The incident is linked to an observation reported three days earlier with the following summary: "PTA performed prior to isolating compressors and topping up oil". PTA (Personal Task Assessment) refers to a job hazard analysis which is performed before a maintenance activity and aims to identify all possible hazards associated with the maintenance activity. The causality indicates that the oil leak reported in the incident could be related to the equipment maintenance reported in the observation. We learned from the incident's "Lessons learnt" record

that the incident was indeed caused by preventative maintenance: "Vigilance around preventative maintenance programs for drive head maintenance are necessary to help prevent future spills". This example shows that causalities could assist in incident investigations and shed light onto incident causes.

Similarities between events described in Examples 1–3 were measured based on their summaries; hence, in Tables 1–3 we only show summaries and omit other text descriptions of the events. In Example 4 (described below) safe activity descriptions of the observations and the incident's findings were used to measure event similarity; hence, in Table 4 we also show these attributes.

Table 4 describes two causalities related to the same incident. The incident and the observations were reported on the same day. The incident reports an injury sustained by an employee and the observations describe a project safety stand down held after the incident (observation 1) and a safety stand down meeting conducted after the incident (observation 2). As we can see from the incident's findings (Table 4), wet weather was a factor contributing to the injury: "Combination of wet weather; rain and residual oil from broken pipe work present on the skid deck". During the safety stand down meeting employees discussed details of the incident and the importance of assessment of changing weather conditions as well as other safety practices: "... Open discussion on the incident including; oil and water on work surfaces; ...; use of absorbent pads to clean and maintain work surfaces; assess changing weather conditions rain water on skid area ..." (Table 4, safe activity description of observation 2). The example demonstrates that causalities can help to discover how incidents are managed in order to prevent similar incidents from occurring.

In summary, the examples described above show how causalities could help to identify systemic issues (Example 1) or events that provide early warnings (Example 2), could assist with incident investigations (Example 3) and help to understand incident management practices (Example 4). Such discoveries could help to prevent H&S incidents (Examples 1 and 4) as well as process safety incidents (Examples 2 and 3).

7. Related Work

In order to improve the management of H&S risks, many organizations use occupational H&S management systems [17, 23]. Such systems record large volumes of H&S data; however, it remains a challenge to analyse this data and extract insights which could help to prevent H&S incidents [17, 23, 24].

A few recent articles describe the use of Big Data tools in the H&S domain [25–29]. Rashidy et al. [25] present an approach for modeling safety data which is based on the use of graph databases and aims to facilitate the identification of factors associated with SPADs (Signals Passed At Danger, i.e., "events where a train passes a stop signal and proceeds onto a section of track where it does not have authority" [25]). Guo et al. [26] describe a case study in which they developed a Big Data platform for collecting, classifying and storing data about unsafe behaviours of workers involved in a construction project. Walker and Strathie [27] demonstrate how data from On-Train Data

Table 4: Example 4: "Incident management"

1	**Summary of observation 1**	"HP Drains Compressor Re-Wire project held a safety stand down after arnlaceration incident."
	Safe activity description of observation 1	"HP Drains Compressor Re-Wire project held a safety stand down after an arm laceration incident. I ran the assembled team and Operations personnel through the incident that occured today and had a session around hazards and controls for safe access; egress and work conditions in a very tight workface. The input was constructive; open and honest with involvement of plant operators as well. The output from the session will be reviewed in another construction risk assessment tomorrow morning."
	Incident summary	"IP was accessing an area of the [Equipment ID anonymised] Compressor. Whilst doing so sustained a laceration to his lower left forearm."
	Incident findings	"Work areas tight and restrictive due to the design of the fixed plant and scope of work required. There was no specific risk assessment of the method to access and egress the work area over and through piping.Grip tape had not been installed on all pipes that were to be footed. Combination of wet weather; rain and residual oil from broken pipe work present on the skid deck.IP was carrying hand tools in one hand and stumbled momentarily loosing three points of contact whilst accessing work area and as a result left arm came in contact with fixed plant."
	Time proximity	Both events happened on the same day.
	Organizational proximity	Both events were reported by the same business unit.
	Location proximity	Both events happened at the same physical location, text descriptions of the specific locations are different.
2	**Summary of observation 2**	"Safety Stand-Down Meeting conducted following FAC incident"
	Safe activity description of observation 2	"Safety Stand-Down Meeting conducted following FAC incident to discuss - Details of the incident Open discussion on the incident including; oil and water on work surfaces; use of grip tape on all steel pipework that must be stepped onto for access; access to tight and awkward work spaces; maintain three points of contact whilst accessing work areas; use of absorbent pads to clean and maintain work surfaces; assess changing weather conditions rain water on skid area; Other issues - Emphasis on health and safety as first priority and then schedule Requirement to maintain work areasVigilance around open grid mesh work zones.Potential heat stress and hydration issues leading into summer."
	Incident summary	"IP was accessing an area of the [Equipment ID anonymised] Compressor. Whilst doing so sustained a laceration to his lower left forearm."
	Incident findings	As in "Incident findings" shown above for causality 1 (not shown here due to space limit).
	Time proximity	Both events happened on the same day.
	Organizational proximity	Both events were reported by the same business unit.
	Location proximity	Both events happened at the same physical location, text descriptions of the specific locations are different.

Recorders (OTDR) can be used to help identify risks faced by rail operators and related to human performance. Tan et al. [29] describe the use of Big Data tools for risk mitigation, e.g., the use of hands-free checklists by workers in the oil & gas industry to decrease the chance of mistakes during equipment assembly. Huang et al. [28] propose an accident investigation paradigm based on the use of safety Big Data and present a case study which illustrates the paradigm. Despite an increasing interest in the use of Big Data tools in the H&S domain, there are a number of challenges associated with such applications which require further research [24]. Moreover, Ouyang et al. [24] argue that "safety data is becoming exponentially un-analyzable with traditional statistics methods" and researchers should "rethink on how to exploit the vast values of safety data efficiently".

Process mining is a research area at the intersection of business process management (BPM) and data analytics which is concerned with analysing process execution history recorded in event logs [8]. Process mining includes process discovery and enhancement tools which can construct or repair process models based on event log data (e.g., Leemans et al. [30] and Polyvyanyy et al. [31]) as well as methods that can analyse and visualise various aspects of process behaviour such as process conformance and performance (e.g., van der Aalst et al. [32]). Process mining methods were recently used to analyse process execution data of a safety process and helped to uncover some process performance and conformance issues [33]. However, such methods can only work with logs which record events that are "clearly defined and refer to precisely one case (i.e., process instance) and one activity (i.e., step in the process)" [34].

If events are not explicitly linked to process instance identifiers, event logs suitable for process mining can be created using an event correlation approach [35–40]. Event correlation is "the process of finding relationships between events that belong to the same process execution instance" [35]. These approaches [35–40] are based on the definition of event correlation conditions, constraints or rules, e.g., "two events can only be causally related, if their activities are also related in the mined process model" [39]. Similar to these event correlation approaches, we aim to find relationships between events and we define a set of hypotheses for linking events; however, our approach aims to find causal dependencies between H&S events rather than link events to process instance identifiers.

In summary, while a number of recent articles report the use of Big Data analytics in the H&S domain [25–29], to the best of our knowledge, the problem of identifying unknown causal dependencies between H&S events has been overlooked. On the other hand, a number of approaches in the process mining area aim to correlate events [35–40]; however, these methods are tailored to process execution data. Unlike the methods described above, we presented an approach for finding causal dependencies between events recorded in H&S data.

8. Conclusion

The paper presented a novel approach for discovering causal dependencies between events recorded in large volumes of H&S data. The approach is based on the notion of proximity of H&S observations and incidents. The proposed approach was evaluated through a case study conducted in an Australian energy company. The evaluation demonstrated that the approach can uncover causal dependencies between observations and incidents reported by the company's employees and that the discovered causalities can shed light onto prevention, investigation and management of H&S and process safety incidents.

We acknowledge that the presented method is prone to several limitations. For example, at this stage, we can not conclude that the proposed model for computing proximity between observations and incidents generalizes to other data sets from the same or a different domain. Moreover, the method requires interventions of experts to model the domain and to fine tune its parameters. Hence, at this stage, the method cannot be fully automated. In addition, one can formulate and test further hypothesis for establishing proximity between events and study other factors, beyond the phenomenon of proximity, that explain the causal relationships between events.

The presented work opens a few avenues for further research. One possibility is the incorporation of machine learning into the approach in order to more accurately evaluate the overall proximity of H&S events. For example, one can apply supervised and semi-supervised learning techniques to use the information on the identified, and confirmed by the domain experts, causalities to incrementally learn "good" coefficients and weights of the computation model. Another direction for future work is the development of a method which can monitor observations and trigger a warning when the likelihood of an incident is high. Finally, the possibility of embedding the approach into the organization's health and safety environment and its impact on the organization's health and safety culture could be investigated.

References

[1] S. O'Neill, N. Martinov-Bennie, A. Cheung, K. Wolfe, Issues in the Measurement and Reporting of Work Health and Safety Performance: A Review, Macquarie Lighthouse Press, 2013.

[2] Safe Work Australia, The Cost of Work-related Injury and Illness for Australian Employers, Workers and the Community: 2012–13, Canberra, URL: https://www.safeworkaustralia.gov.au/system/files/documents/1702/cost-of-work-related-injury-and-disease-2012-13.docx.pdf (accessed 2018-10-8) (2015).

[3] J. Takala, P. Hämäläinen, K. L. Saarela, L. Y. Yun, K. Manickam, T. W. Jin, P. Heng, C. Tjong, L. G. Kheng, S. Lim, G. S. Lin, Global estimates of the burden of injury and illness at work in 2012, Journal of Occupational and Environmental Hygiene 11 (5) (2014) 326–337.

[4] Safe Work Australia, Table 2.1 - number, frequency rate and incidence rate of serious claims by industry (2011–12 to 2015–16p), URL: https://www.safeworkaustralia.gov.au/system/files/documents/1805/number-frequency-incidence-serious-claims-by-industry-2011-12-to-2015-16p.pdf (accessed 2018-10-8).

[5] Safe Work Australia, Table 2.2 - number, time lost and compensation paid for serious claims by industry (2011–12 to 2014–15), URL: https://www.safeworkaustralia.gov.au/system/files/documents/1805/number-time-lost-and-compensation-paid-serious-claims-industry-2011-12-2014-15.pdf (accessed 2018-10-8).

[6] H. Chen, R. H. L. Chiang, V. C. Storey, Business intelligence and analytics: From Big data to big impact, MIS Quarterly 36 (4) (2012) 1165–1188.

[7] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011.

[8] W. M. P. van der Aalst, Process Mining: Data Science in Action, 2nd Edition, Springer, 2016.

[9] C.-W. Liao, Y.-H. Perng, Data mining for occupational injuries in the Taiwan construction industry, Safety Science 46 (7) (2008) 1091–1102.
[10] M. Bevilacqua, F. E. Ciarapica, G. Giacchetta, Data mining for occupational injury risk: A case study, International Journal of Reliability, Quality and Safety Engineering 17 (04) (2010) 351–380.
[11] C.-W. Cheng, S.-S. Leu, Y.-M. Cheng, T.-C. Wu, C.-C. Lin, Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry, Accident Analysis & Prevention 48 (2012) 214–222.
[12] J. Ruso, V. Stojanovic, Occupational health and safety using data mining, International Journal for Quality Research 6 (4) (2012) 355–364.
[13] C. Shearer, The CRISP-DM model: the new blueprint for data mining, Journal of Data Warehousing 5 (4) (2000) 13–22.
[14] O. Marbán, G. Mariscal, J. Segovia, A data mining & knowledge discovery process model, in: Data Mining and Knowledge Discovery in Real Life Applications, I-Tech Education and Publishing, 2009.
[15] A. Polyvyanyy, C. Ouyang, A. Barros, W. M. P. van der Aalst, Process querying: Enabling business intelligence through query-based process analytics, Decision Support Systems 100 (2017) 41–56.
[16] T. Halpin, T. Morgan, Information modeling and relational databases, Morgan Kaufmann, 2010.
[17] S. Sinelnikov, J. Inouye, S. Kerper, Using leading indicators to measure occupational health and safety performance, Safety Science 72 (2015) 240–248.
[18] K. Karimi, A brief introduction to temporality and causality-https://arxiv.org/abs/1007.2449 (accessed 2018-10-8).
[19] M. Born, Natural philosophy of cause and chance, Clarendon Press, Oxford, 1949.
[20] W. H. Gomaa, A. A. Fahmy, A survey of text similarity approaches, International Journal of Computer Applications 68 (13) (2013) 13–18.
[21] A. Polyvyanyy, Evaluation of a novel information retrieval model: eTVSM, Master's thesis, Hasso Plattner Institute, Potsdam, Germany (2007).
[22] A. Polyvyanyy, D. Kuropka, A quantitative evaluation of the enhanced topic-based vector space model, Potsdam, Germany, 2009, ISBN: 978-3-939469-95-7; URN: urn:nbn:de:kobv:517-opus-33816.
[23] D. Podgórski, Measuring operational performance of OSH management system — a demonstration of AHP-based selection of leading key performance indicators, Safety Science 73 (2015) 146–166.
[24] Q. Ouyang, C. Wu, L. Huang, Methodologies, principles and prospects of applying big data in safety science research, Safety Science 101 (2018) 60–71.
[25] R. A. H. E. Rashidy, P. Hughes, M. Figueres-Esteban, C. Harrison, C. Van Gulijk, A Big Data modeling approach with graph databases for SPAD risk, Safety Science 110 (2017) 75–79.
[26] S. Guo, L. Ding, H. Luo, X. Jiang, A big-data-based platform of workers behavior: Observations from the field, Accident Analysis & Prevention 93 (2016) 299–309.
[27] G. Walker, A. Strathie, Big data and ergonomics methods: a new paradigm for tackling strategic transport safety risks, Applied ergonomics 53 (2016) 298–311.
[28] L. Huang, C. Wu, B. Wang, Q. Ouyang, A new paradigm for accident investigation and analysis in the era of big data, Process Safety Progress 37 (1) (2018) 42–48.
[29] K. H. Tan, V. G. Ortiz-Gallardo, R. K. Perrons, Using Big Data to manage safety-related risk in the upstream oil & gas industry: A research agenda, Energy Exploration & Exploitation 34 (2) (2016) 282–289.
[30] S. Leemans, D. Fahland, W. van der Aalst, Discovering block-structured process models from event logs – a constructive approach, in: International conference on applications and theory of Petri nets and concurrency, Springer, 2013, pp. 311–329.
[31] A. Polyvyanyy, W. M. P. van der Aalst, A. H. M. ter Hofstede, M. T. Wynn, Impact-driven process model repair, ACM Transactions on Software Engineering and Methodology 25 (4) (2017) 1–60.
[32] W. van der Aalst, A. Adriansyah, B. van Dongen, Replaying history on process models for conformance checking and performance analysis, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2 (2) (2012) 182–192.
[33] A. Pika, A. ter Hofstede, R. K. Perrons, G. Grossmann, M. Stumptner, J. Cooley, Analysing an industrial safety process through process mining: A case study, in: J. Mathew, C. Lim, L. Ma, D. Sands, M. Cholette, P. Borghesani (Eds.), Asset Intelligence through Integration and Interoperability and Contemporary Vibration Engineering Technologies, Lecture Notes in Mechanical Engineering, Springer, 2019, pp. 491–500.
[34] W. van der Aalst, Extracting event data from databases to unleash process mining, in: J. vom Brocke, T. Schmiedel (Eds.), BPM – Driving innovation in a digital world, Springer, 2015, pp. 105–128.
[35] H. R. Motahari-Nezhad, R. Saint-Paul, F. Casati, B. Benatallah, Event correlation for process discovery from web service interaction logs, The VLDB Journal – The International Journal on Very Large Data Bases 20 (3) (2011) 417–444.
[36] R. Pérez-Castillo, B. Weber, I. G. R. de Guzmán, M. Piattini, Improving event correlation for non-process aware information systems., in: International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE), SciTePress, 2012, pp. 33–42.
[37] F. Mannhardt, M. De Leoni, H. A. Reijers, Extending process logs with events from supplementary sources, in: International Conference on Business Process Management, Vol. 202, Springer, 2014, pp. 235–247.
[38] R. Pérez-Castillo, B. Weber, I. G.-R. de Guzmán, M. Piattini, J. Pinggera, Assessing event correlation in non-process-aware information systems, Software & Systems Modeling 13 (3) (2014) 1117–1139.
[39] S. Pourmirza, R. M. Dijkman, P. Grefen, Correlation miner: Mining business process models and event correlations without case identifiers, International Journal of Cooperative Information Systems 26 (2) (2017) 1–32.
[40] L. Cheng, B. F. van Dongen, W. M. van der Aalst, Efficient event correlation over distributed systems, in: Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2017, pp. 1–10.