Data-driven Relevancy Estimation for Event Logs Exploration and Preprocessing

Pierre Dagnely, Elena Tsiporkova and Tom Tourwé
Sirris - Elucidata Lab, Belgium

Keywords: Event Relevancy Estimation, Data Reduction, Industrial Event Logs, Data Preprocessing.

Abstract: With the realization of the industrial IoT, more and more industrial assets are continuously monitored by loggers that report events (states, warnings and failures) occurring in or around these devices. Unfortunately, the amount of events in these event logs prevent an efficient exploration, visualization and advanced exploitation of this data. Therefore, a method that could estimate the relevancy of an event is crucial. In this paper, we propose 10 methods, inspired from various research fields, to estimate event relevancy. These methods have been benchmarked on two industrial datasets composed of event logs from two photovoltaic plants. We have demonstrated that a combination of methods can detect irrelevant events (which can correspond to up to 90% of the data). Hence, this is a promising preprocessing step that can help domain experts to explore the logs in a more efficient way and can optimize the performance of analytical methods by reducing the training dataset size without losing information.

1 INTRODUCTION

With the realisation of the industrial IoT, more and more industrial assets are continuously monitored. These assets are not only instrumented with sensors, but are also monitored by loggers that report events occurring within or around the device. This includes events related to the state of the asset (e.g. whether it is started, or stopped), warnings and failures. The event logs represent a valuable source of information for a company, as they can be explored to learn about the (normal and abnormal) behaviour of assets in the field, e.g. to discover the root cause for a failure.

Event logs are currently underexploited by most companies, as they come with several challenges. (Bose et al., 2013) identified five categories of challenges. For instance, 1) The voluminous amount of data, 2) The heterogeneity of the data, i.e. the presence of events from distinct process in the logs or generated by different and incompatible firmwares from different manufacturers, 3) The fine granularity of the data, i.e. events that are too precise and lack generalization, 4) Evolutionary changes, i.e. definitive or momentary changes in the process that affect the data, and 5) Quality issues with data that can be missing, incorrect, imprecise or irrelevant, i.e. related to other process than the one analyzed. In this paper we focus on the ”voluminous”, fine granularity and irrelevance challenges.

The voluminous, and subsequently fine grained character of event log data prevents its efficient exploration, visualisation and advanced exploitation. For example, our industrial partner, 3E, monitors, through its Software-as-a-Service platform SynaptiQ, around 4.000 photovoltaic (PV) plants throughout the world and receives 1.000.000 events per day on average. Manually analysing this huge amount of data is no longer possible and methods that can pinpoint events or periods of interest are required to support domain experts. For instance, domain expert would like to (visually) explore a device's historical data when a relevant failure occurs, in order to understand its behaviour and identify the failure's root cause. In this case, he is interested in a (very limited) subset of the events only and the vast majority of the events can be considered as noise that should/could be removed.

Similarly, advanced analytical methods could predict failures and support a maintenance engineer in pro-actively planning maintenance. However, training and validation for such methods can have very high computation time, as they are directly correlated to the size of the dataset. In addition, their performance suffers significantly because of the fact that the vast majority of the events can actually be considered as noise. As an example, we have applied sequential pattern mining (SPM) to event logs in order to iden-

Dagnely, P., Tsiporkova, E. and Tourwé, T.
Data-driven Relevancy Estimation for Event Logs Exploration and Preprocessing.
DOI: 10.5220/0006579503950404
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 395-404
ISBN: 978-989-758-275-2

tify anomalous behaviour in PV plants and found that the algorithmic performance was rather poor, both in terms of running time and in terms of results. A detailed analysis revealed that this was due exactly to the huge amount of events and hence presence of noise in the log data.

In this paper, we propose 10 methods that automatically estimate event relevancy and evaluate their ability to discriminate relevant from irrelevant events. The definition of "relevant" depends of the objective to achieve and may vary. For example, when exploring normal behaviour, only the events describing or representing the regular/normal behavior of a PV plant (called regular events) are of interest and the events describing or representing the failures or underperformances of a PV plant (called irregular events) should be discarded. On the other hand, when investigating failures, irregular events are the most interesting to explore and regular events can be considered noise. For the rest of the paper, we will address the challenge of labelling the regular events as irrelevant (as this is the most common case) since both cases can be addressed with the same methods and only differ in how the results are interpreted.

By applying such methods, we can significantly reduce the dataset size without losing information. For example, we have reduced the computation time of a SPM algorithm by 83%. The computation time on the initial dataset was 16 hours and 14 minutes while it was 2 hours and 40 minutes on the dataset with only the relevant events. However, the same useful patterns were found.

The 10 methods, inspired by similar techniques proposed in other domains, are applied to event logs from 2 PV plants for which one year of historical data is available (provided by our industrial partner). They are evaluated using a thorough benchmarking framework, including an evaluation by domain experts. Some of our methods are data-driven, i.e. they are domain-independent and can be applied to any event log, while others rely on domain knowledge related to PV event behaviour, and hence would require tailoring to be applicable in other domains.

The remainder of the paper is structured as follows. We first cover the state of the art of the fields from which we draw inspiration for the methods. Second, we describe the 10 methods that have been developed. Then, in Section 4, we explain and motivate the experimental setup used for the benchmarking. Section 5 discusses the results of this benchmarking. Finally, Section 6 concludes the paper.

2 RELEVANT LITERATURE

The challenges related to the voluminous character of event log data have received minimal attention in literature. The current strategies usually let the process mining methods deal with it, such as (Bonchi et al., 2003) who defined Ex-ante, a constrained pattern mining method adapted to large data by adding a pruning of the search space during the traditional pruning of the frequent itemsets of the APriori method. Nonetheless, we can draw inspiration from five domains that have encountered similar challenges and defined techniques for dealing with them: 1) The process mining domain, 2) The outlier detection domain, 3) The web log cleaning domain, 4) The state index pruning domain from the information retrieval field and 5) The diversity measures in the biological domain.

For the process mining field, we can mention the research of (Fu et al., 2012), where they applied two filtering steps on their events logs: 1) Removed the frequent events, i.e. if the period between two similar events is below a user defined threshold, they removed the second event, 2) Removed periodic events, e.g. events produced by a daemon regularly. For each type of event, they assessed different time intervals and computed, for each, the number of occurrences of that event with this time interval between each consecutive occurrence. If that number (expressed as a percentage on the total number of occurrences of that event) is above a user-defined threshold, they considered that they had a cycle and only kept the first occurrence. One drawback of this method is its inability to find cycle with various interval gaps. (Hassan et al., 2008) used a compression algorithm to detect relevant periods. They divided the logs in small sequences and compressed each of them and subsequently, computed their compression rate CR.

$$\mathrm{CR} = 1 - \frac{\text{size after compression}}{\text{size before compression}}$$

Sequences with low CR should contains sequences with distinct events making it hard to compress. Therefore, they only kept these sequences that should be the most relevant ones.

Outlier detection focusses on the detection of non-frequent events or patterns. Therefore, some outliers detection methods can be adapted to this purpose, i.e. to build datasets of outliers. (Gupta et al., 2014) used decision tree and clustering methods to learn a model of the data and removed the events which differ from it. Similarly, (Conforti et al., 2016) used an automaton to model the system. Every event not filling the usual automaton process was considered as outliers

and discarded. However, these methods require training a model on a voluminous dataset which is computationally expensive and may not lead to optimal results due to the amount of noises in the data.

Data pruning is a long-standing topic of the web log cleaning/pre-processing since the seminal paper of (Cooley et al., 1999), where they removed jpeg and other multimedia files from the event logs. Nevertheless, there has been little progress in this area, as stated in the review of (Jayaprakash and Balamurugan, 2015). The main cleaning procedures concern the removal of queries other than GET, multimedia resources, logs created by robots, errors and some HTTP status codes known as irrelevant. In addition, methods focusing on user/session identification, path completion, i.e. correction of incomplete URL or transaction interaction identification, i.e. cluster of meaningful transaction of a user extracted from the event logs, are now part of the everyday web log preprocessing. (Srivastava et al., 2015) decreased the size of the log data by 80% by applying such techniques. However, most of these methods are web centric and can not be applied to other domains.

Static index pruning is a field of the information retrieval domain. The goal is to reduce the size of the index of one or several texts by removing the entries that will probably not be needed by the users, e.g. the index entries of the word "the" or "me", to reduce the memory size of the index. This field exists since the seminal papers of (Carmel et al., 2001). Even though some methods are domain specific, others can be adapted to industrial event logs. (Billerbeck and Zobel, 2004) used TF-IDF to compute the frequency score of the words by combining the overall frequency of the word in all texts and the number of text in which this word occurs. Given a corpus of texts, e.g. a set of technical documents, this method will provide a ranking of the word for each document, e.g. for document A, the word "inverter" will have a high score as this word is frequent in this document but not in the others, which means that this word is probably discriminative of the document topic. TF-IDF combines two metrics: 1) The term frequency (TF) that measures the frequency of the term, i.e. word, in the document. 2) The inverse document frequency (IDF) that measure the (inverse of the) frequency of the term in the corpus (see formulas below).

$$TF_{w_i,d_i} = \frac{\text{\# occ. of word } w_i \text{ in document } d_i}{\text{\# of word in document } d_i}$$

$$IDF_{w_i} = \log \frac{\text{\# of documents in the corpus}}{\text{\# of documents containing the word } w_i}$$

$$TF-IDF_{w_i,d_i} = TF_{w_i,d_i} * IDF_{w_i}$$

The intuition behind that is to attribute a low score to terms that occur in most or all of the documents as they are probably less relevant (words such as "the" or "of"). Billerbeck et al. used TF-IDF to remove from the index the words with low scores, i.e. frequent words.

(De Moura et al., 2005) adapted this method by considering the context. They analyzed the non-frequent sentences, i.e. the sentences containing non-frequent words and kept the occurrences of the frequents words occurring in these sentences. All the other occurrences were removed from the index. (Jangid et al., 2014) reviewed these methods and compared TF-IDF to BM25 (Jangid et al., 2014), BB2 (Jangid et al., 2014) and other methods. TF-IDF still had (one of) the best accuracy although its computation time was usually superior.

A new promising approach defined by (Chen et al., 2015) uses the Rnyi divergence (a divergence measure from the information theory domain). They redefined the problem as a model induction problem and looked for the pruned index that minimize its (Rnyi) divergence with the full index (in terms of retrieval performance). However, this method lacks a proper comparison with other methods.

Diversity measure is a long-standing problem in the biological and genomic domain. This domain tries to assess the diversity in, e.g. a DNA sequence or an environment like a lake. For example, (Fuhrman et al., 2000) used it to detect DNA subsequences with high diversity that are worth exploring (the other subsequences are discarded). Popular methods are shannon index (also called Shannon-Wiener or Shannon-Weaver indexes), Simpson index or Berger-Parker index. Although Shannon index has been criticized for the difficulty to understand and interpret its meaning (Hill et al., 2003), it's one of the most used method.

The index is based on the Shannon entropy used in the information theory and simply computes the entropy of a sequence. Sequences with the highest entropy are sequences with the highest diversity (as the entropy measure the difficulty to predict the identity of the next individual in the sample, a high entropy means that there is a high diversity of potential individuals that could occur next, i.e. that there is a high diversity of individuals in the sequence). The Shannon index is computed using the formula below where p_i is the proportion of individuals belonging to the ith species, gene, ... in the sequence.

$$\text{Shannon score (H)} = -\sum p_i \log_2 p_i$$

3 METHODS DEFINITION

In this section, we present the definition of 10 methods for estimating event relevancy based on our literature study and the characteristics of our data. These methods have been selected for their effectiveness in their respective fields. However, their applicability to the industrial event log field still needs to be assessed. In addition, these methods, often, can not be directly applied to industrial event logs and need to be adapted and tailored.

These methods are either based purely on data characteristics, hence they are generally applicable, or they are based on specific domain knowledge, and thus require adaptation before they can be applied to other domains.

In addition, the methods differ in the way events are ranked. *Score-based* methods assign a relevancy score to each event type while *boolean-based* methods only classify the events as relevant or not. This distinction is crucial as boolean-based methods are simpler to interpret and apply. Score-based methods offer a finer-grained ranking of the events and provide more information to the domain experts at the cost of a higher complexity, as they imply to define a threshold that will discriminate between relevant and irrelevant events, based on domain knowledge.

Finally, some methods encounter the cold start problem, as they require the suitability of historical data that could span years in order to train a model that can discriminate between the events. Such data is not always available on fresh installations with new devices and/or conditions which prevents the use of such methods in these situations. A summary of the methods characteristics can be found in table 1.

As some methods are based on PV plants characteristics, we first explain the PV plant infrastructure. PV plants are composed of several PV modules (that convert the irradiation into direct current) connected to one or several inverter(s) (that convert the direct current to alternative current) which send the current to the grid. These systems are now continuously monitored (in addition with various on and off-site sensors). Therefore, a PV plant reports statuses, i.e. its current state like start, stop or running, but also events, i.e. specific events that can represent an outage (such as grid fail or string disconnected) or other phenomena (such as over-temperature or DC current under threshold). In this context, an outage, i.e. a failure, is considered as the period where there is enough irradiation reaching the modules to have electricity production but there is no yield. Unfortunately, events can not be trusted to detect all the outages due to some noises in the logs. Therefore, outages are detected by combining the data of the irradiation and yield sensors. Periods with irradiation values above 30kwh/m^2 but yield value null are labelled as outages.

The event logs consist of textual messages containing the ID of the events detected, e.g. 1013 for the event grid fail, associated with a timestamp. Usually, an inverter reports 14 events per days (mainly during the mornings and evenings when the inverter start or stop) and a plant contains in average 18 inverters, but it can go up to 600. The number of events (as well as the type of events monitored and their ID) reported by an inverter depends of the manufacturers of the logger-inverter pair, e.g. the inverters of Fronius combined with the logger system of Solarlog report in average 19 events per days. The number of events also increase when outages and events occurs in the plant. In some case, an inverter can report thousands of events per days with a granularity of a few seconds.

3.1 Status Method

Some events are status **events**, i.e. an indication of the current state of the system such as ”start” or ”running”. As we focus on labelling regular events as irrelevant, and these status events describe regular behaviour, they can be discarded. However, the status of the system can also be a valuable information to understand some failures.

3.2 Simultaneous Status Method

In some circumstances, e.g. during start and stop sequences, the inverter status can change rapidly, i.e. in a few seconds. This method labels the status that lasted less than 60 seconds as irrelevant, as the phenomenon that generated it did not have time to impact the system.

3.3 Consecutive Event Method

Some events are continuously reported every few minutes by the logger until the problem is fixed. This artificially increases the amount of events while only the first and the last occurrences of the events are actually relevant. All the other occurrences can be considered as irrelevant. Note that the id of the last occurrence should be changed to indicate that it corresponds to the end of the event. In addition, if the duration or the length of repetition is random or irrelevant in some industrial cases, it may represent a valuable information in some other cases, preventing the use of this method that would remove this information.

Table 1: Summary of methods' characteristics.

Method	Data-driven based	Domain based	Score based	Boolean based	Cold start issue
Status method		X		X	
Simultaneous status method		X		X	
Consecutive event method		X		X	
Morning and evening method		X		X	
TF-IDF method	X		X		
Contextual TF-IDF method	X		X		
Pattern method	X			X	X
Compression method	X			X	
Statistical method	X		X		X
Shannon index method	X			X	

3.4 Morning and Evening Method

In the PV domain, most of the events occur during the start and stop sequences that occur in the morning and evening. Those sequences contain therefore events that can be considered as irrelevant as they indicate the usual behavior of the system. We have defined the morning and evening as the operational periods with an irradiation below 30 kwh/m2. Note that this means that true failure events occurring during those periods would be lost, except if they are reported again later.

3.5 TF-IDF Method

TF-IDF can be adapted to industrial event logs, by considering the event log of one inverter for one day as one document, called inverter-day document. A corpus of documents is the inverter-day documents of a day. In practice, for each day, the event logs of each inverter are considered as a text and TF is computed for each (see formula below). IDF is then computed for the inverter-day documents of that day and TF-IDF scores are computed by multiplying TF and IDF scores of each inverter-day documents (the scores are computed days by days). Hence, a relevancy (TF-IDF) score is computed for each event type of each inverter-day document, i.e. each type of event has a specific score for each day of each inverter log.

$$TF_{e_i,i_i,d_i} = \frac{\text{\# occ. of event } e_i \text{ in inverter } i_i \text{ for day } d_i}{\text{\# of events in inverter } i_i \text{ for day } d_i}$$

$$IDF_{e_i,d_i} = \log \frac{\text{\# of inverter-days for day } d_i}{\text{\# of inverter days containing } e_i \text{ for day } d_i}$$

$$TF-IDF_{e_i,i_i,d_i} = TF_{e_i,i_i,d_i} * IDF_{e_i,d_i}$$

A threshold can then be puts to discriminates irrelevant events, i.e. events with low scores as they are the frequent and probably meaningless events (in the same way as the words "the" is irrelevant for text processing) . Note that, as the scores are inverter-day specific, the list of events types labelled as irrelevant may vary for one day to another as their score may also vary. In this way, we have extended the traditional TF-IDF to time dependent one or, in other words, we have defined a dynamic TF-IDF score.

3.6 Contextual TF-IDF Method

De Moura et al. (De Moura et al., 2005) extended the TF-IDF method to take into account the context of each word. They only removed irrelevant words if there were no relevant word in their surroundings, i.e. in the same sentence. We have adapted this method by retaining irrelevant events that have a relevant event in the period of 30 minutes before or after their occurrence. This time window of one hour have been selected based on domain knowledge and is therefore domain specific.

3.7 Pattern Method

Based on (Gupta et al., 2014) or (Conforti et al., 2016), another method is to find the usual behavior of the system by e.g. defining an automaton and removing all events that correspond to it. Unfortunately, methods such as automaton would have been complex to use due to the high heterogeneity of the PV events logs (as events are often reported at various conceptual level by different manufacturers, for more information, we refer to our papers (Dagnely et al., 2015)). Therefore, we have decided to use Multi-level sequential pattern mining (MLSPM) methods to find the patterns representing the usual behavior, i.e. the patterns with high support thresholds in our dataset, as MLSPM is well suited to address the granularity issue of the PV events. The events that occur in these patterns can be classified as irrelevant.

3.8 Compression Method

Based on Hassan et al. (Hassan et al., 2008), we have defined a method that uses a compression algorithm to detect periods with relevant events. The event logs of each day of each inverter are compressed. The compression rate (CR) is then computed and used as a metric. The intuition is that periods with failures will have many redundant events and will therefore be easier to compress. Therefore, the events occurring in periods with low compression rate are labelled as irrelevant. We have used the same compression algorithm than Hassan et al. (zlib) and the same compression rate formula.

3.9 Statistical Method

The probability that an event is linked to a failure can be computed by analyzing which events occurs in normal and abnormal periods. From the event logs, we have extracted the days preceding the outages and the (same amount of) days that are "the further away" of any outages, i.e. with the more days between them and the next or the previous outage. The first ones have been labeled as regular while the later ones have been labelled as irregular. For each event present in these datasets, its frequency in both datasets has been computed and normalized (per dataset) using the min-max normalization. The probability of an event to be linked to an outage has then be computed using the formula:

$$P(\text{event } e_i) = \frac{\text{\# occ. of event } e_i \text{in the abnormal dataset}}{\text{\# occ. of event } e_i \text{ in both datasets}}$$

Domain experts have selected a threshold of 0.3 as the more adapted to discriminate the events as relevant or irrelevant, i.e. event with a probability lower than 0.3 are considered irrelevant. Another threshold could have been 1, to only classify as relevant the events that never occurred during the normal behavior. However, this may be too strict and can remove harmless events that have an impact on abnormal behavior in combination with "failure" events. One drawback of this method is the need of sufficient historical data to compute a statistically significant probability. We have used 1 year of historical data to compute the probabilities.

3.10 Shannon Index Method

The Shannon index has been adapted by using the formula below applied to the log file of each day of each inverter. If the Shannon index score of a day is below a certain threshold, the day is considered as having a low diversity and therefore is discarded as irrelevant. A threshold of 2.5 has been selected by domain experts as an appropriate threshold.

$$\text{Shannon score (H)} = -\sum p_i \log_2 p_i$$

$$\text{With } p_i = \frac{\text{\# occurences of event } e_i \text{ during day } d_i}{\text{\# events during day } d_i}$$

3.11 Combination of Methods

Methods can also be combined. Two good candidates are the status and TF-IDF methods. One of the drawback of TF-IDF is its computation time which is directly correlated to the size of the dataset. Therefore, applying first the status method to quickly label some of the events as irrelevant and therefore decreasing the size of the (relevant) dataset on which applying TF-IDF may dramatically decrease the computation time. The status events will therefore be removed from the TF-IDF ranking but, as the status method don't interferes with the distribution of the warning and failure events, their TF-IDF ranking will not be impacted.

The three others domain-based methods only remove some occurrences of the events and therefore may decrease the accuracy of the data-driven methods by creating biased datasets. For data-driven methods, the pattern method needs the state events as they play an important role in the regular patterns. The Shannon index method has already a low computation time and do not need preprocessing. Finally, the compression method could benefit from the status method to reduce its computation time. Therefore, the two only valid combinations are the TF-IDF and compression methods combined with the status method.

4 EXPERIMENTAL SETUP

The event relevancy estimation methods have been evaluated in the PV domain. One of the challenges of events logs in that domain is the lack of consistency (in addition to the voluminous challenge). Most of the events represent the usual start and stop sequences of an inverter. However, the composition of these sequences is not fixed. Each manufacturer has defined its specific start and stop sequences but even within one manufacturer device, the sequences may vary with distinct events and different time gaps, depending on the external conditions (such as the weather). In addition, outages can be preceded by none, a few or even more than thousands of events, with continuous repetition of the related event(s) and start and stop sequences, as the inverter tries to restart. Nevertheless, the number of events is not an indication of an outage, as many events can be reported continuously for simple warnings.

To evaluate these methods, we have benchmarked them on one year (2016) of data from two PV plants, one with regular and one with irregular behaviour. The regular plant is composed of 16 inverters and contains

84.961 events. This plant is known to have a regular behavior with only a few outages due to external factors such as the weather. The irregular plant is composed of 26 inverters and contains 133089 events. Over the year, this plant has suffered of several outages due to various causes such as Riso low or Ramp fail. Both plants use the same inverter type and are located in Belgium (hence they share the same climate conditions). Those two plants will allow to compare the methods accuracy in two distinct situations: one with only a few outages and almost only irrelevant (regular) events and one with more outages and more relevant (irregular) events (up to 30% of the events). The methods need to be accurate in both situations which may be especially complex for methods with cold start issues, i.e. for methods that need to learn from the data. As the regular plant only contains a few outages and relevant events, it may be not enough to properly train these methods. The distribution of the amount of events occurring in one day can be found in Figure 1. The usual daily amount of event for both plants is around 14 but the irregular plant encounters more days with around 80 events reported.

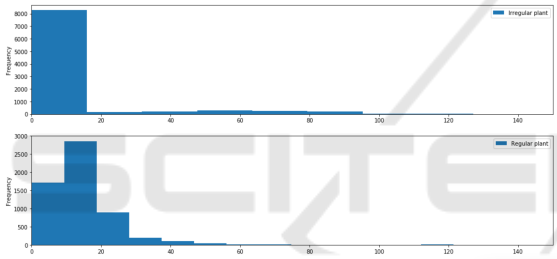

Figure 1: Distribution of the amount of events occurring in one day for regular and irregular plants.

Unfortunately, the data is not labelled and as such we don't know which events are actually relevant or not. This problem arises frequently in industrial domains, where the only way to label event logs is to let domain experts do it manually, which is obviously not a viable solution on such large datasets. Actually, it is this lack of labelled data that motivated this study. Hence, there is no proper metric that could be used to assess the methods' accuracy (other than having domain experts manually evaluate all methods). Any metric that could allow to directly assess this labelling would actually be a perfect method to label the data and solve the issue. Therefore, we have developped metrics to indirectly approximate the methods accuracy based on data characteristics specific to our dataset. As a consequence, the comparison with other dataset is hampered by the fact that these metrics are dataset-centered and can only be applied to datasets with similar behaviour.

First, domain experts have indicated that a regular plant typically reports around 90% of irrelevant (regular) events while an irregular plant typically reports between 70% to 80% of irrelevant events. These numbers are an estimation but can be considered as a correct indication with a margin of error of around 10% depending of the plants considered. Second, the failures occurring in a plant that give rise to the events can be detected by combining two additional datasets: 1) The irradiation sensor values that measure the solar irradiation reaching the plant, and 2) The yield value that measures the electricity production of the plant. An outage is detected when there is enough irradiation to have electricity production but there is no yield.

Based on the above, we can define two metrics: 1) The percentage of events labelled as irrelevant. As we have a rough estimation of the expected percentage based on domain knowledge, we can evaluate if the method seems to over- or under-label the events as irrelevant. 2) The percentage of outages without explanation, i.e. without any relevant events during or (up to 45 minutes) before its occurrence. A wrong labelling would increase the number of outages without explanation, because the over-labelling of relevant events as irrelevant will lead to outage only preceded by irrelevant events.

Note that these two metrics only provide an approximate and should be interpreted carefully as they can not directly assess the labelling accuracy but rather try to approach it by analyzing over- and under-labelling. To compensate this, we have also relied on domain experts who have analyzed the accuracy of the methods on small data subsets or have analyzed the event ranking of score-based methods.

The third metric is the computation time. These methods may need to be applied on datasets on demand when domain experts and technical staff want to explore and exploit them. Therefore, the computation time is crucial to ensure scalability and an industrial applicability.

5 RESULTS AND DISCUSSION

Table 2 contains, for each dataset and each method, the three metrics: the % of events labelled as irrelevant (% events irr.), the computation time (comp. time), and the % of outages whithout explanation (expl. lost).

The observed methods can be split in four characteristics groups based on their performances: 1) The first group contains the methods which fail for time windows issues, i.e. the sequences of events have to be split in windows that are not adapted to the PV event logs. 2) The second group consists of the methods which fails due to the lack of sufficient historical data to train the model. 3) The third group contains the domain-based methods that, in overall, perform well but only label a small part of the events. 4) The last group consists of the methods that perform well.

The first group contains the Shannon index and compression methods. The Shannon index method is clearly not adapted to event logs. For regular event logs, the method has failed to label the events as irrelevant (only 19% of the events have been labelled as such while most of them are probably irrelevant). This result can be explained by the fact that regular days only contain a few

Table 2: Summary of methods' efficiencies.

Methods	Regular plant			Irregular plant		
	% events irr.	comp. time	expl. lost	% events irr.	comp. time	expl. lost
Status method	**90%**	**3 sec**	**0%**	**79%**	**23 sec**	**0%**
Simultaneous status method	48%	10 min	0%	57%	7 min	0%
Consecutive event method	0.5%	60 sec	0%	14%	4 min	0%
Morning and evening method	98%	14 sec	4.9%	81%	25 sec	4%
TF-IDF method	88%	28 min	0.8%	73%	84 min	0%
Contextual TF-IDF method	88%	30 min	0%	73%	98 min	0%
Pattern method	57%	30 min	1%	39%	66 min	17%
Compression method	83%	7 min	6%	41%	4 h	19%
Statistical method	39%	16 sec	1%	**82%**	**58 sec**	**0%**
Shannon index method	19%	28 sec	2.7%	62%	12 sec	19%
status + TF-IDF methods	**90%**	**17 min**	**0%**	**82%**	**19 min**	**0%**
status + compression methods	99%	64 sec	6%	83%	4 min	18%

distinct types of events. Therefore, the entropy of such sequence will be high. This method failed as well for irregular periods as it wrongly labelled events as irrelevant and lost information of some outages (19% of the outages for which there is no explanation).

Similarly, the compression method has failed with the irregular dataset as the number of outages without explanation also rose to 19%. This may be because these two methods analyze periods of 1 day and then label the whole period as irrelevant. These twenty-four hours time windows have been selected as they represent physical periods, i.e. the inverter "reboots" after the night where it has stoped. However, by selecting smaller periods it may be possible to decrease this over-labelling. The difficulty would reside in the choice of the time windows as a too small time windows may artificially split outage periods into many smaller sub-periods that may decrease the method accuracy. The combination of status and compression methods only reduces the computation time but keeps the time-windows problems. Hence, this combination of methods has the same low accuracy than the compression method alone.

The second group consists of the pattern and statistical methods that fails due to the lack of sufficient historical data to train the model on the dataset. The pattern method suffered a lack of accuracy for irregular periods (with an increase of 17% of the outages without explanations). Although this method has performed well for the regular dataset (in all metrics). The patterns were found with high support thresholds, hence they represent the (very) frequent patterns common to regular days. However, some of the less frequent harmless patterns occurring during the less frequent irregular days are missed. It explains the difference of performance on regular and irregular datasets. A drawback of this method is the computation time (1 hour for the irregular dataset). In addition, this computation time does not include the time needed to identify the patterns, which in our experience could take days, as it is manufacturer specific. However, these patterns only need to be found once for a type of inverters and can then be applied to any plants with that type.

The statistical method has performed well for irregular periods but poorly for regular periods. As this method is based on a probabilistic analysis of the outages, it requires a dataset that contains enough outages to have statistically significant results, which is not the case for regular datasets. However, this method is still a perfect method to quickly analyze irregular periods and rank the events. In addition, a larger dataset combining several plants with devices from the same manufacturer and close location (hence same behavior) could be used to compute a ranking of the events specific to that type of device (and location). It could then be applied to any plant with that type of device, without the cold start issue.

The third group contains the domain-based methods, namely the morning and evening, status, simultaneous and consecutive event methods. Status and simultaneous methods have been both really effective in terms of events correctly labelled for both datasets and with low computation time (below 10 minutes). However, they do not provide any help to analyze the remaining events. For example, although the status methods labelled as irrelevant 79% of the events of the irregular dataset, there were still 25.140 relevant events in the dataset which can still be cumbersome to analyze.

The consecutive event method also seems an interesting method for the irregular dataset as 14% of this dataset is composed of the repetition of one event until its fix. Therefore, these 14% could easily be removed without impacting the dataset. Note that this method has obviously no use for regular dataset where such events do not occur.

The morning and evening method seems too strict in its labelling as the amount of events without explanation increased by 4% (which may, however, be acceptable in some situations). It indicates that for some periods the start and stop periods are important and can not be labelled as irrelevant as a whole. It is worth noting that for the regular dataset, 98% of the events occurred within an hour period of the start and stop sequence (and only 81% of the events are in the same situation for the irregular dataset).

Finally, the last group consists of both TF-IDF methods and the combination of the status method followed by TF-IDF that performed well. Both of TF-IDF methods have performed very well for both datasets with respectively 88% and 73% of the events labelled as irrelevant with (almost) no losses of information. Therefore, the contextual TF-IDF method barely improve the accuracy of the traditional TF-IDF on event logs. The only drawback of these methods are their computation times that went up to one hour and a half for the irregular dataset.

A combination of the status method followed by TF-IDF can reduce significantly the computation time of TF-IDF without impacting the accuracy of the method. The TF-IDF scores were also still similar to the one found without the application of the status method as preprocessing step. The computation time on the regular dataset dropped from 28 minutes to 17 minutes and was reduced by 65 minutes (77%) on the regular dataset (from 84 minutes to 19 minutes).

As a conclusion, the methods Shannon index, compression, pattern, contextual TF-IDF and Morning and evening have not been found addapted for industrial events logs. A good way to label the events as relevant or irrelevant is to combine first a domain-based method such as status method to already perform a first quick labelling. Then TF-IDF method can be used to help to label the remaining events. A statistical method can also be used for irregular datasets. Both methods are actually complementary as statistical methods only focus on the correlation between events and outages while TF-IDF method sort the event by their relevancy and uniqueness, without any regard for the outages.

6 CONCLUSION AND FURTHER RESEARCH

In this paper, we have considered and evaluated 10 methods (from various research fields) to estimate the event relevancy in industrial event logs, to detect irrelevant events that could be discarded during the preprocessing of voluminous data. These methods have been benchmarked on two datasets containing real industrial events logs from two PV plants. We have found that a combination of two methods (one removing the state events and one applying TF-IDF) allows to label up to 90% of the events as irrelevant with a reasonable computation time

For further research, we intend to evaluate other score-based methods from the static pruning index field, especially the methods BM25, BB2 or the Rnyi divergence used by (Chen et al., 2015), to benchmark them on industrial events logs. In addition, the statistical method can also be applied on device specific datasets, i.e. on datasets containing the event logs of devices of same type from multiple plants. This may allow to create device specific ranking that could then be applied on all devices of that type without pre-processing of the data. However, a thorough study of these scores would need to be performed to assess e.g. if the location of the device has an impact on these events scores.

ACKNOWLEDGEMENTS

This work was subsidised by the Region of Bruxelles-Capitale - Innoviris.

REFERENCES

Billerbeck, B. and Zobel, J. (2004). Techniques for efficient query expansion. In *International Symposium on String Processing and Information Retrieval*, pages 30–42. Springer.

Bonchi, F., Giannotti, F., Mazzanti, A., and Pedreschi, D. (2003). Exante: Anticipated data reduction in constrained pattern mining. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 59–70. Springer.

Bose, R. J. C., Mans, R. S., and van der Aalst, W. M. (2013). Wanna improve process mining results? In *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*, pages 127–134. IEEE.

Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y. S., and Soffer, A. (2001). Static index pruning for information retrieval systems. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM.

Chen, R.-C., Lee, C.-J., and Croft, W. B. (2015). On divergence measures and static index pruning. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 151–160. ACM.

Conforti, R., La Rosa, M., and ter Hofstede, A. H. (2016). Filtering out infrequent behavior from business process event logs. *IEEE Transactions on Knowledge and Data Engineering*.

Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, 1(1):5–32.

Dagnely, P., Tsiporkova, E., Tourwe, T., Ruette, T., De Brabandere, K., and Assiandi, F. (2015). A semantic model of events for integrating photovoltaic monitoring data. In *Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on*, pages 24–30.

De Moura, E. S., dos Santos, C. F., Fernandes, D. R., Silva, A. S., Calado, P., and Nascimento, M. A. (2005). Improving web search efficiency via a locality based static pruning method. In *Proceedings of the 14th international conference on World Wide Web*, pages 235–244. ACM.

Fu, X., Ren, R., Zhan, J., Zhou, W., Jia, Z., and Lu, G. (2012). LogMaster: mining event correlations in logs of large-scale cluster systems. In *Reliable Distributed Systems (SRDS), 2012 IEEE 31st Symposium on*, pages 71–80. IEEE.

Fuhrman, S., Cunningham, M. J., Wen, X., Zweiger, G., Seilhamer, J. J., and Somogyi, R. (2000). The application of Shannon entropy in the identification of putative drug targets. *Biosystems*, 55(1):5–14.

Gupta, M., Gao, J., Aggarwal, C. C., and Han, J. (2014). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267.

Hassan, A., Martin, D., Flora, P., Mansfield, P., and Dietz, D. (2008). An industrial case study of customizing operational profiles using log compression. In *Software Engineering, 2008. ICSE'08. ACM/IEEE 30th International Conference on*, pages 713–723. IEEE.

Hill, T. C., Walsh, K. A., Harris, J. A., and Moffett, B. F. (2003). Using ecological diversity measures with bacterial communities. *FEMS microbiology ecology*, 43(1):1–11.

Jangid, C. S., Vishwakarma, S. K., and Lakhtaria, K. I. (2014). Ad-hoc Retrieval on FIRE data set with TF-IDF and Probabilistic Models. *International Journal of Computer Applications*, 93(18).

Jayaprakash, S. and Balamurugan, E. (2015). A Comprehensive Survey on Data Preprocessing Methods in Web Usage Minning. *International Journal of Computer Science and Information Technologies*, 6(03):3170–3174.

Srivastava, M., Garg, R., and Mishra, P. K. (2015). Analysis of Data Extraction and Data Cleaning in Web Usage Mining. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, page 13. ACM.