# Analytics of industrial operational data inspired by natural language processing

Mariusz Kamola
NASK – Research and Academic Computer Network
Warsaw, Poland
e-mail: Mariusz.Kamola@nask.pl

Mariusz Kamola
Institute of Control and Computation Engineering
Warsaw University of Technology
Warsaw, Poland
e-mail: M.Kamola@ia.pw.edu.pl

*Abstract*—**Industrial processes provide lots of operational data on different timescales. Those data are well-structured and used now for daily control, longer-term management and forensics. We propose to pre-process that data and treat them the way the natural language processing is done - first, in order to find common ways the process is controlled. Such knowledge can then be used in prediction or early detection of faults, or necessary manufacturing shifts. Gas transmission operational data are considered here as the live example.**

***Keywords – predictive analytics, smart historian***

## I. Introduction

Retrieval of information from big data was historically performed for data generated directly by humans, e.g. on social sites. Consequently, analytic tools focused on natural language processing. Now, with the increasing volume of sensor generated data, adequate algorithms are needed that will profit most from the structured nature of the data being processed. Otherwise, tons of valuable information will be lost in the coming era of the internet of things, plainly because there will be no adequate storage for them.

Analysis of structured industrial data usually calls for expert knowledge – at least for validation, filtering and presentation phases. The tools used there differ totally from those applied to natural language processing. However, the final meaning of the data is similar: both humans and industrial installations communicate their internal condition to the outer world.

Here we ask a question: is it possible to approach industrial records the same way we treat documents created in natural language? Does it give any advantage over traditional approaches?

This paper proposes three ways a summary of industrial process operation on some horizon may be created. They are then compared for gas transmission operational data. The results are also compared with the key performance indicator of the process.

## II. Related Work

The necessity of novel treatment of process data is clearly stated by big companies [1-3], which have already come up with appropriate products for gas, oil, energy and manufacturing industries [4]. Their core functionality is advanced monitoring [5] of historical data, provided by the module called usually a historian. This enables more advanced functionalities: diagnosis (e.g. of faults), prediction and, finally, optimization of operations. The raw industrial data are processed before they get stored for further use, which requires adequate computational architecture [6].

In the related work, authors claim that technical knowledge of the nature of the installation providing data is necessary – and they apply different processing rules to each type of data collected. In [7], various filtering and averaging is performed on inputs; then rules are established to infer about specific events that take place in a modern house with a score of sensors collecting data. Another work [8] reports big data approach in modeling glider behavior, and [9] uses large amounts of sensor data to model the rolling process in order to predict product quality and discover factors that influence it most.

Data for predictive analytics, as the ones mentioned above, may be supplied by any other non-industrial sources – e.g. outputs from external systems (weather forecasts) or even by someone's subjective opinions. The work [10] proposes a methodology to combine them into a coherent and useful input for a model, by establishing ontologies and proper mappings.

## III. Proposed Approach

While expert knowledge of the nature of industrial big data being collected is definitely an advantage, here we consider processing them without any prior familiarity with the process that generates them. Such agnostic approach may be justified due to many reasons: the huge number of industrial process outputs, the need to trace relationships in data that lie beyond engineering intuition, the aim to select the minimum set of measurements defining states of the industrial process.

In our approach we propose combining two concepts: uniform data preprocessing and information retrieval traditionally applied for big data expressed in natural languages. The preprocessing should therefore capture the essence of most industrial processes, and create bits of data that could be next treated like words in a human language. This essence is related to two terms: the state of the process and the dynamics of the state. They both are fundamental for control theory; together they compose state and output equations, necessary in control design and execution.

Control theory has well-established apparatuses for modeling dynamical systems of various kinds: be it a $\mathcal{Z}$-transform, auto regression models with external inputs etc. Their common feature is they model dynamics of the observed outputs from the object. Let us then observe industrial process outputs on some horizon, calculate selected statistics and call them *words* the process speaks to the outer world.

More formally, let $\boldsymbol{y}(t) = (y_1(t), \ldots, y_N(t))$ be a vector of observed outputs from industrial installation at discrete time $t$. The outputs can take discrete or real values of any magnitude. By processing outputs $\boldsymbol{y}(t), \ldots, \boldsymbol{y}(t-H)$ observed on time horizon $H$, we prepare a *document* $\boldsymbol{d}(t)$, i.e. a vector summarizing observed behavior of the installation. Document $\boldsymbol{d}(t)$ can be created in many ways; consider for the moment the following propositions:

A. $\boldsymbol{d}$ contains statistics of values and their increments over horizon $H$:

$$\boldsymbol{d}(t)=(y_1^{\text{mean}}, \Delta y_1^{\text{mean}}, y_1^{\text{std}}, \Delta y_1^{\text{std}},\\ y_1^{\text{skewness}}, \Delta y_1^{\text{skewness}}, y_1^{\text{kurtosis}}, \Delta y_1^{\text{kurtosis}}\\ y_2^{\text{mean}}, \ldots, \Delta y_N^{\text{kurtosis}} \quad ). \tag{1}$$

Those statistics capture both the nature of the process outputs (average values and most important properties of their distributions) as well as process dynamics, denoted here with deltas, i.e. increments of outputs between measurements.

B. $\boldsymbol{d}$ contains statistics as in (1), with values rounded to closest typical values (called here *synonyms*):

$$\boldsymbol{d}(t)=(\tilde{y}_1^{\text{mean}}, \Delta\tilde{y}_1^{\text{mean}}, \tilde{y}_1^{\text{std}}, \Delta\tilde{y}_1^{\text{std}}), \tag{2}$$

where

$$\tilde{y}_1^{\text{mean}} = \arg \min_{\bar{y}_{1\in} \in \bar{Y}_1^{\text{mean}}} (\bar{y}_1 - \tilde{y}_1^{\text{mean}}) \ldots \tag{3}$$

– and so on, for subsequent $y$'s. $\bar{Y}_1^{\text{mean}}$ is a set of typical values $y_1^{\text{mean}}$ takes throughout the whole system lifetime.

C. Instead of statistics on outputs, $\boldsymbol{d}$ contains frequencies of $y$'s and $\Delta y$'s, rounded to their synonyms:

$$\boldsymbol{d}(t)=(f_{1,1}, \ldots, f_{1,M_1}, f_{\Delta 1,1}, \ldots, f_{\Delta 1,M_{\Delta 1}},\\ f_{2,1}, \ldots, f_{2,M_2}, f_{\Delta 2,1}, \ldots, f_{\Delta 2,M_{\Delta 2}}, \cdots \quad ), \tag{4}$$

where $f_{i,j}$ is the number of $y_i$ values that have been rounded to $j$-th synonym, calculated analogously to (3).

The concept of synonym is central to propositions B and C. It has been introduced to reduce the number of possible "words", i.e. values that subsequent project output can take before. This is analogous to the way the natural language is working: we have a limited number of words we use to tell an unlimited number of stories. The choice of synonyms for each variable $y$ is based on the distribution of values $y$ takes; in general, synonym values are chosen to represent distributions most efficiently.

## IV. The Industrial Process

High pressure gas network has been selected as the primary object for testing of the proposed methodology. The installation, composed of pipelines, valves and compressor stations, is delivering over 1,000 different measurements (pressure, flow, rpm, humidity and temperature) with varied dynamics. Successful application of the methodology may lead, first, to reduction of collected historical data to a subset large enough to describe the industrial process of gas pumping fully. Using such data, one might 1) classify operation modes of the installation, 2) look up similar scenarios in history, in order to e.g. trace back failure reasons or to predict near future behavior. Consequently, one may finally optimize system operation, either internally, through improved predictions, or externally, by providing operation conditions (flows and pressures on system edges) such that the system operates optimally, although it runs still under its autonomous control algorithms.

## V. Modeling Results

The three proposed methods of "documents" creation (A,B,C) have been applied to data from the installation, collected in winter season 2009/2010. There were 43,000 samples in total. First, it was checked if created documents showed any mutual similarity, and could be clustered into sets corresponding to typical ways the installation was operated. If so, the documents could be used for any predictive analytics task mentioned earlier. We assumed horizon $H$=6 hours, which covers most transitive effects in the installation.

### *A. Document containing raw statistics*

For documents created with method A, the angle between two documents was assumed as the measure of their dissimilarity, or the distance. The matrix of distances for documents calculated every hour was shown in Fig. 1. One can see that the installation is controlled in a number of typical ways (red squares on the diagonal). Some of them are

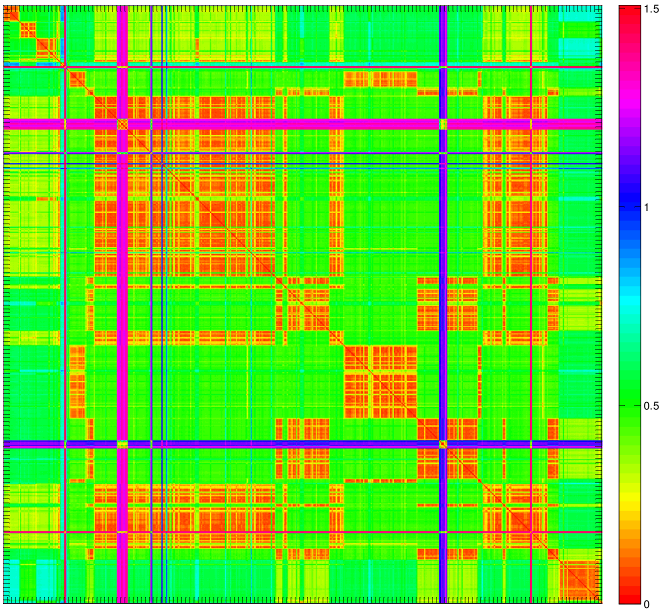

Fig. 1. Distance between documents created with method A (red – most similar; purple – least similar).

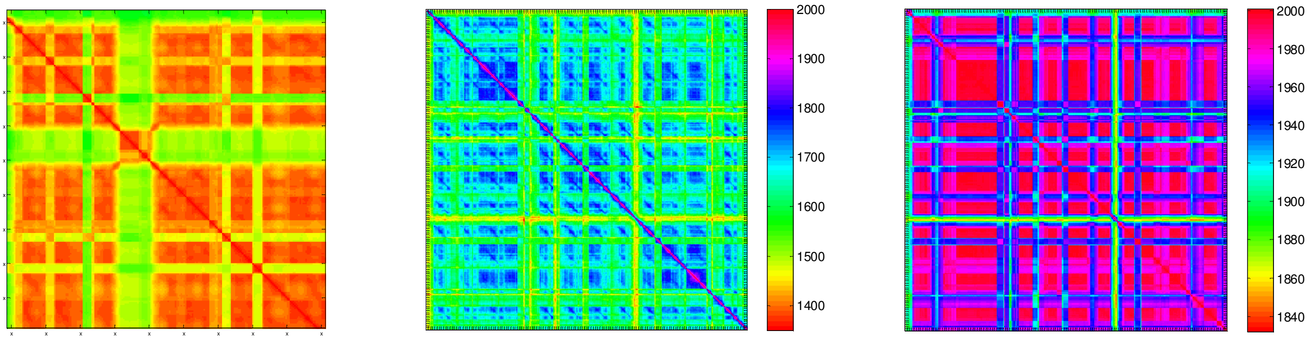

Fig. 2. Left: Close-up of Fig. 1 (shown daily seasonality of control, ticks on axes mark subsequent days). Middle: Similarity of documents created with method B, considering first 2,000 rarest synonyms. Right: Like in the middle, but only synonyms from position 3,000 to 5,000 taken into account.

effectuated a number of times in history: their mutual similarity is indicated by red rectangles away from the diagonal. It is interesting that consecutive modes of operation frequently show no apparent sign of similarity. The explanation is that occasionally the system undergoes a reconfiguration, e.g. the decision is made to divert excess gas volume to caverns, which requires extra pumping and valve adjustment, although other parts of the system (supply and local consumption) stay unaffected.

Moreover, if we examine document similarity on finer timescale (Fig. 2, left), we will also clearly see the daily rhythm of process control. There are shorter periods (usually nights) when the system is controlled the different way – in the figure it is marked with narrow yellow and green stripes. Occasional 1-day reconfigurations, like the one almost in the center are usually due to the operator executing the manual control of the system, or by temporary extraordinary pumping orders received from country-level coordination center.

It should be noted that no normalization of the elements of documents has been done yet; however the model is already capable of performing quite accurate comparisons of historical situations.

### B. *Document containing statistics rounded to synonyms*

In documents created with method B we stored synonyms that describe the statistics of outputs as closely as possible. For document comparison we first sort those synonyms in each document, putting to front the rarest ones, i.e. those appearing least frequently in whole system history. This corresponds looking for rare terms while comparing documents written in natural language, which is the essence of the well-known TF-IDF (term frequency – inverse document frequency) scheme. The difference is we consider the TF part to carry binary (*true*/*false*) information. Eventually, the similarity of any two documents is just the number of synonyms appearing in both documents. Since we give attention to rare terms, we consider only the first $M$ rarest synonyms in each document, for the comparison.

Fig. 2, middle, shows similarities rather than distances of the documents, defined as cardinality of the intersection of two sets, containing $M$ rarest synonyms in each document. Comparing to Fig. 1 we note that method B gives much coarser results: it qualifies documents not distant in temporal terms, which would be considered similar by method A, as considerably different. Those unsatisfactory effects may be due to the naïve analogy made between rare readings from industrial installation and rare words used in human speech. While the latter appear always deliberately and may really solve as landmarks for documents, the earlier may be just noise.

To verify this, we tried to compare documents using neither the rarest synonyms nor the most frequent – but those in the middle of the scale (position 3,000 to 5,000). The results are shown in Fig. 2, right. As we turn toward more frequent synonyms, the documents differ less and less: in this case by at most 170 different synonyms used in them. However, this is sufficient to reveal the most differing modes the gas network operates in.

### C. *Document containing frequencies of outputs rounded to statistics*

Finally, we checked similarity of documents that contain frequencies of installation outputs and output increments. Values subject to those statistics had been previously discretized to their typical values, like in method B. TF-IDF was again the inspiration for document similarity computation; here, we used cosine similarity, and dropping the IDF part. The matix of document distances for the initial forty days of system operation is presented in Fig. 3 (left), against the corresponding close-up of Fig. 1 (Fig. 3, right). Apparently, method C is less prone to qualify documents as

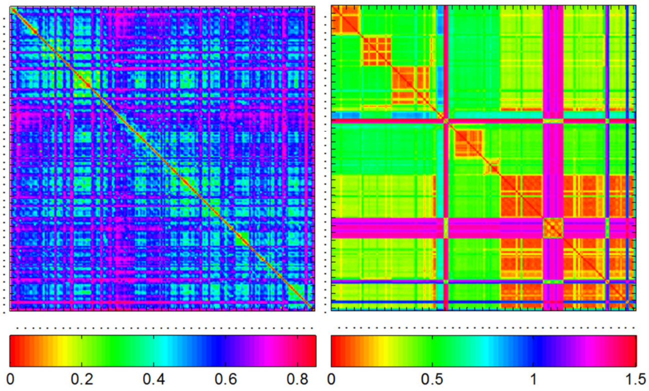

Fig. 3. Distance for each pair of documents created with method C (left) vs. method A (right).

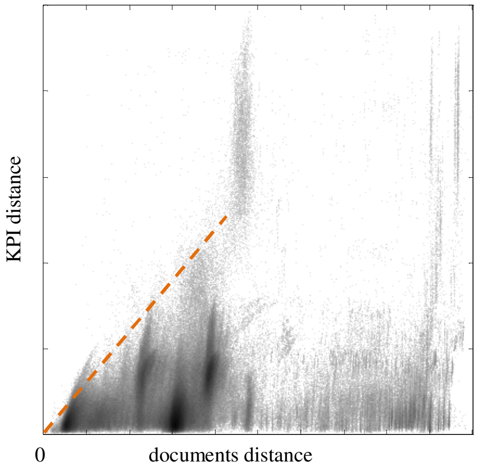

Fig. 4. Correlation of distances between documents (method A) and corresponding operational cost

similar. This conclusion has to be still verified for the whole history of the system. Nevertheless, both figures bear some resemblance: many blocks of red color on right diagonal have vaguely corresponding blocks in green. Some pairs of blocks are quite accurate, as the one in the bottom-right corner.

## VI. Documents vs. KPI

Since the documents created by method A apparently can be clustered, thus describing typical ways the process is operated, we may ask: are the documents naturally related to the key performance indicator (KPI) of the process? Pumping cost is the main cost of the gas grid operation, and the KPI. The nature of the grid makes the total cost on 6 h horizon not much informative for further comparisons, so let us consider relative distance of instantaneous cost trajectories to be the measure of the KPI dissimilarity. This is what a typical historian program would look for in the system history. We calculate such KPI distance for each pair of documents $(\boldsymbol{d}, \boldsymbol{d}')$ and compare it in Fig. 4 with the documents distance calculated as described in Sec. V.A. Small document distance visibly implies similarity of cost trajectories, enveloped quite precisely with the dashed line – but not the reverse. This is because the similar cost trajectories can be obtained by running formally quite different process configurations – e.g. activating another set of compressors, with the same pumping capacity. In our case the documents that are agnostic of system structure do not model KPI naturally – however, the apparent envelope in Fig. 4 proves that similar values of *statistics* on process data mean similarity of cost *trajectories*, which is not intuitive.

## VII. Conclusion

We made an attempt to model industrial system behavior by preparing vectors of statistical values, calculated on a given horizon – without getting into the detail of system operation. Those vectors serve to find similar patterns in system history, by using algorithms rooted in natural language processing (cosine similarity and TF-IDF). All three specific methods of document creation reveal similarity of operation of a gas network, considered here as example.

The introduction of typical values used for comparison (methods B and C) can reduce the amount of historical data needed to be stored, but at the cost of accuracy of the calculated similarities between scenarios.

Sensitivity of results to parameters of the used methods needs to be verified – especially the influence of horizon length $H$ and the way the "synonyms" (typical values for statistics) are chosen. Independently, testing of the methods against operational data from a huge N. American water supply network is in progress.

Although the proposed modeling approach falls into category of big data analysis, its properties in terms of control theory can and should be devised for certain classes of systems (e.g. linear). Particularly, it is worthy to assess which sorts of statistical properties are relevant to typical system characteristics (inertia, lag, integration etc.). Here we proposed the mean, standard deviation, skewness and kurtosis. In some cases Fourier or wavelet transforms could be more accurate. This still remains an open direction of theoretical research.

## References

[1] "The rise of industrial big data," white paper by GE Intelligent Platforms, online: http://www.geautomation.com/download/the-rise-of-industrial-big-data/13476 (accessed 2015.02.24).

[2] M. Desenuk, "Big data, physics, and the industrial internet", Rock Stars of Big Data Analytics, San Jose, CA, 2014, online: http://media.computer.org/pdfs/Denesuk.pdf (accessed 2015.02.24).

[3] G. Kreß, "Big data in industrial applications", European Data Forum 2013, Dublin, online: http://www.slideshare.net/EUDataForum/edf2013-keynote-gerhard-kre-big-data-in-industrial-applications (accessed 2015.02.24).

[4] H. Chen, Hsinchun, RHL Chiang, and VC Storey. "Business intelligence and analytics: from big data to big impact", MIS quarterly 36.4 (2012): 1165-1188.

[5] A. Wang, "Big data in industrial sector is different: industrial internet", Big Data Innovation Summit, Hong-Kong, 2014.

[6] JW Williams, KS Aggour, J. Interrante, J. McHugh, E. Pool, "Bridging high velocity and high volume industrial big data through distributed in-memory storage & analytics," Big Data, 2014 IEEE International Conference on , pp. 932-941, 2014.

[7] V. Jirkovský, et al. "Big Data analysis for sensor time-series in automation." Emerging Technology and Factory Automation (ETFA), 2014 IEEE. IEEE, 2014.

[8] L. Zhang, "A framework to model big data driven complex cyber physical control systems." Automation and Computing (ICAC), 2014 20th International Conference on. IEEE, 2014.

[9] N. Lu, J. Bin, L. Jianhua, "Data mining-based flatness pattern prediction for cold rolling process with varying operating condition." Knowledge and Information Systems 41.2 (2014): 355-378.

[10] V. Jirkovský, M. Obitko, "Semantic Heterogeneity Reduction for Big Data in Industrial Automation", Znalosti 2014 Conference, online: www.znalosti.eu/images/accepted_papers/znalosti2014_paper14.pdf (accessed 2015.02.24).