Towards an information extraction and knowledge formation framework based on Shannon entropy

Dragoş Iliescu^{1,*}, and *Marian Gheorghe*¹

¹University Politehnica of Bucharest, Faculty of Engineering and Management of Technological Systems, 313 Splaiul Independentei, 060042 Bucharest, Romania

Abstract. Information quantity subject is approached in this paperwork, considering the specific domain of nonconforming product management as information source. This work represents a case study. Raw data were gathered from a heavy industrial works company, information extraction and knowledge formation being considered herein. Involved method for information quantity estimation is based on Shannon entropy formula. Information and entropy spectrum are decomposed and analysed for extraction of specific information and knowledge-that formation. The result of the entropy analysis point out the information needed to be acquired by the involved organisation, this being presented as a specific knowledge type.

1 Introduction

Information acquisition and knowledge formation represents a complex domain for study and research. A specific system is considered to control the information for extraction, verification and storage. This system refers to information management domain.

Related knowledge formation is then expected, based on managed information. Knowledge formation depends with the quality of managed information as only true justified information can be transformed into knowledge [1 - 3].

The 2015 version of the ISO 9001 standard express in clause 7.1.6 a specific requirement regarding the identification and management of organisational knowledge [4]: “The organization shall determine the knowledge necessary for the operation of its processes and to achieve conformity of products and services. The knowledge shall be maintained and made available to the extent necessary.”

The above clause raises several challenges for organisations: how to identify needed or missing knowledge, how to segregate critical important knowledge and how to spread them across organisation. What information source to consider and which particular technology need to be implemented to control the transformation of information to knowledge [5].

* Corresponding author: dragosiliescu2005@gmail.com

Information sources for above objectives are diverse. The international standards ISO 9001:2015 and ISO 9004:2010 recommend that own organisation experience to be considered for above objective, along with certain internal and external sources [4, 6].

Description of a framework to evaluate the quantity of extracted information and how the collected information can be evaluated for knowledge formation represents the main topic of present work.

A case study will be described to exemplify the exposed concept. Information describing the nonconformities occurred in the period of 2014 and 2015 were gathered from a heavy industrial equipment manufacturing company.

Paper is structured as follows: section two defines the categories of identified nonconformities, section three operate the gathered data for information quantity based on Shannon formula and offer an interpretation of obtained results, section four presents the conclusion of the work as well as further possible development.

2 Nonconformities as information source

The records of identified nonconformities during the manufacturing cycle represent the data source of present work. Deviation is herein considered as any individual lack of compliance against a specified requirement. Nonconformity can be determined by one or a group of deviations [7] recorded in a specified moment of time within the manufacturing cycle.

The 2015 edition of ISO 9001 requests in clause 10.2.1 organisations to perform a series of specific activities to control and to correct the identified nonconformities. In clause 10.2.2 it is requested a record log to be maintained by organisation, describing the identified nonconformities by category and nature. Results of applied corrections [4] should also to be recorded. Data gathered for the study herein were extracted from this type of data log within the period of 2014 and 2015.

A total amount of 23 deviation categories, noted $x_1 \dots x_{23}$ (Table 1), have been identified, 278 occurrences for 2014 and 346 in 2015 were recorded.

Table 1. Deviation categories.

Deviation category designation			
x_1	Shape and dimensions in finished condition state	x_{13}	Deviations found by magnetic particle inspection
x_2	Material cleanliness	x_{14}	Wrong or skipped process
x_3	Shape and dimensions in rough condition state	x_{15}	Process parameters
x_4	Product hardness	x_{16}	Defects in weld material
x_5	Deviations found by ultrasonic inspection	x_{17}	Other material characteristics
x_6	Surface condition	x_{18}	Product chemical composition
x_7	Cracks	x_{19}	Deviations found by dye-penetrant inspection
x_8	Deviations in mechanical test results	x_{20}	Deviations found by penetrant radiation inspection
x_9	Deviations found by visual inspection	x_{21}	Shape and dimensions for heat treatment
x_{10}	Shape and dimensions in rough-machined condition state	x_{22}	Microstructure
x_{11}	Chemical composition on liquid steel	x_{23}	Process documentation readability
x_{12}	Weight		

3 Estimation of information quantity

Shannon formula for entropy is employed to measure the information quantity [8, 9]: entropy of a variable X (noted as $H(X)$), with discrete values of $\{x_i \mid i=1...n\}$ and corresponding probability distribution of $\{p_i \mid i=1...n, p_i \geq 0\}$ is shown by eq. 1.

$$H(X) = - \sum_i p_i \log_2(p_i)$$
 (1)

Applied for current case, variable X represents the likelihood of nonconformity to occur. The probability p_i (related with x_i deviation category) is approximated by eq. 2 [8, 9], where ω_i represents the occurrence of deviation category x_i and ω represents total occurrence counted for variable X .

$$p_i = \frac{\omega_i}{\omega}; \sum_i p_i = 1; i = 1 \dots n$$
 (2)

For the case of equiprobability ($p_1 = p_2 = p_3 = \dots = p_n$), $H(X)$ gets the maximum value [8, 9] based on formula $H(X)_{\text{Max}} = \log_2(n)$, where n represents the number of X 's discrete values. The meaning of maximum entropy principle is related here with the information quantity to acquire: depending upon the number of discrete values recorded for variable X , a specified number of bits are needed to code that particular information [8 - 10]. It is considered herein that each x_i term can take only two values: yes or no, so that the entropy to be measured in bits. For the current case study, 5 bits (upper of 4.52 in $\log_2(23) = 4.52$) codes the embedded information.

3.1 2014 versus 2015 entropy comparison

Gathered data were analysed, based on the eq. 2, for probability estimation. Results for each x_i term are depicted in Table 2 and Figure 1. The term of $p_i \log_2(p_i)$ reflects the effect on entropy of each individual term x_i .

Table 2. Calculation for occurrence probability as 2014 and 2015.

	2014		2015			2014		2015	
	ω_i	$p_i \log_2(p_i)$	ω_i	$p_i \log_2(p_i)$		$\omega_{i \text{ 2014}}$	$p_i \log_2(p_i)$	ω_i	$p_i \log_2(p_i)$
x_1	67	0.495	86	0.499	x_{13}	6	0.119	2	0.043
x_2	31	0.353	13	0.178	x_{14}	5	0.104	17	0.214
x_3	25	0.313	44	0.378	x_{15}	4	0.088	1	0.024
x_4	20	0.273	61	0.441	x_{16}	2	0.051	3	0.059
x_5	18	0.256	15	0.196	x_{17}	1	0.029	-	-
x_6	18	0.256	6	0.101	x_{18}	1	0.029	-	-
x_7	16	0.237	12	0.168	x_{19}	1	0.029	2	0.043
x_8	16	0.237	26	0.281	x_{20}	1	0.029	1	0.024
x_9	16	0.237	20	0.238	x_{21}	-	-	1	0.024
x_{10}	14	0.217	28	0.294	x_{22}	-	-	1	0.024
x_{11}	9	0.160	6	0.101	x_{23}	-	-	1	0.024
x_{12}	7	0.134	-	-					

The 2014 data are sorted in descendent probability order base (Figure 1). The 2015 results preserved the general trend developed by 2014 records, but some particular variations are noted among the terms and this variation will be further analysed. The change in entropy for the variable X is constraint by the relation of $H(X) \geq 0$, but the change in probability distribution for each of x_i terms, 2015 versus 2014, is unconstrained.

Estimation of entropy for 2014 (eq. 1 based) shows a value of 3.647, versus the entropy of 2015 which shows a value of 3.357. In both cases, a number of 4 bits are needed to code the embedded information. As consequence, even there is a difference between the two values, no significance has been associate to the entropy change.

Variation in probability distribution is shown in Figure 2. The results offer an image over the entropy spectrum, “as before” compared with “as after”.

Fig. 1. Individual effect on entropy for each x_i term.

Fig. 2. Difference 2015 vs. 2014 in effect on entropy.

3.2 Information extraction and knowledge formation

Data gathered to estimate the information quantity in regard with the domain of nonconforming product management show a non-significant difference between the entropies estimated for 2014 ($H(X_{2014}) = 3.647$), versus 2015 ($H(X_{2015}) = 3.357$).

Some of the x_i terms show a significant effect on the entropy variation. Positive values in Figure 2 reflect the terms impacting the information quantity. Terms 4, 14 and 10 (Figure 2) are candidates for the terms which significantly determine a positive variation in entropy. This case is referenced as Case 1 in Table 3. Other terms show influence near zero, or there are lack of records (Case 2), and several other terms show a negative influence (Case 3).

Table 3. Extracted information and acquired knowledge.

Case	Extracted information	Knowledge definition description
1	These terms can be treated for informativeness significance:	Terms were segregated and sorted as informativeness significance order base. The three most informativeness were selected.
a	Product hardness (x_4).	
b	Wrong or skipped process (x_{14}).	
c	Shape and dimensions as rough-machined condition state (x_{10}).	
2	These terms ($x_1, x_9, x_{12}, x_{16}, x_{17}, x_{18}, x_{20}...x_{23}$) should be potentially considered for control improvement by:	Terms with no induced entropy change were selected to represent those domains were, potentially, an improved process control should be expected as necessary.
a	Process parameters control.	
b	Output measuring technique, capability, etc.	
c	Clear (noise reduction) communication channel	
3	These terms ($x_2, x_5...x_7, x_{11}, x_{13}$, and x_{15}) should not be considered for informativeness purposes.	Terms selected for being part in indistinguishable group should be confirmed for past effective corrective actions.

For those cases where the induced variation on entropy exists, either positive or negative, the meaning can be related with a successful transmission from the observed process to the observer. Only the terms inducing the positive entropy change will be counted as informativeness. These cases will transfer the information to the observer.

Case of the terms having an insignificant effect, for example terms 1, 9 and 16 in Figure 2, can be interpreted as being not for interest in the described context. However, they still should be counted as domains potentially requesting process control capability improvement, either in terms of process parameters control or in terms of communication by decreasing the noise over the communication channel.

4 Conclusions

Domain of nonconforming product management has been considered as information source, the main topic in present research was oriented on estimation the information quantity, how information extraction can be possible and how information to be transformed into knowledge.

The entropy concept was used to evaluate the occurred changes into the current manufacturing states. Comparison of collected data was main engine in measuring the change in entropy spectrum, and finally, getting the main factors to be considered for contributions in information quantity.

Regarding the cases where improvements in process control capability are needed, the term of shape and dimensions in finished condition state was identified. This identification represents the second information type revealed by the analysis herein.

The third type is represented by the identification of that group of terms that should not be counted for informativeness. The last identified information showed a particular aspect. If the process under observation has recorded an improvement action in the past, then the results can be interpreted as a confirmation of the improvement effectiveness, this information reclaiming further investigations. All above acquired knowledge are part of the knowledge-that domain.

Returned benefits, as coming out from the three cases depicted above, refer to the selection of appropriate knowledge to be acquired, based on minimum collection of data. Overall, gathered data are discrete, and not continuous, cost of collecting being minimal.

Several limitations can be described into the present work. The entropy formula considers a noiseless communication channel. Other way said, there are no expected modifications in the data content, or meaning, during transmission within the information system. This limitation can eventually affect all the cases depicted in the paperwork, mostly Cases 2 or 3, as the change in entropy seems to be minimal as a result of a blocked communication channel, thus not reflecting actual state of the process. Another implication refers to the state of correlation that potentially exists among the considered terms, but the used entropy formula presumed the existence of independent discrete terms.

References

1. M. Carrara, D. Chiffi, Recent Trends in Philosophical Logic **41**, 31 (2014)
2. D. Davidson, *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson* (Blackwell, Cambridge, 1986)
3. E. Gettier, *Analysis* **23**, 121-3 (1986) <http://fitelson.org/proseminar/gettier.pdf> (accessed on 27.01.2016)
4. ***, ISO 9001:2015 Quality management systems - Requirements (2015)
5. ***, <http://www.cavendishscott.com/articles-news/iso-90012015/what-is-organizational-knowledge-in-iso-90012015> (accessed on 21.03.2016)
6. ***, ISO 9004:2010 Managing for sustained success of an organization. A quality management approach (2010)
7. ***, ISO 9000:2015 Quality management systems. Fundamentals and vocabulary (2015)
8. E. C. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948)
9. A. Caticha, *Proc. 27th Int. Conf. Workshop on Bayesian Inference and Maximum Entropy in Sci. Eng.* (Saratoga Springs, New York, 2007)
10. K. Conrad, <http://www.math.uconn.edu/~kconrad/blurbs/analysis/entropypost.pdf> (accessed on 17.02.2016)