Characterisation of the diffusion states by data compression

M. Bigerelle, A. Iost

Characterisation of the diffusion states by data compression

M. Bigerelle, A. Iost *

Lab. Matériaux, Equipe des Surfaces et Interfaces, LMPGM UMR CNRS 8517 ENSAM Lille, 8 Boulevard Louis XIV, 59046 Lille cedex, France

Abstract

A new method of characterisation of the diffusion is proposed by analysing the results obtained by data compression. By analysing the results of the diffusion process using a Monte Carlo method, it is shown that the ratio of compression could be a good parameter to quantify the rate of the dynamic process and the time to reach the equilibrium. If an accurate method of data compression is used, then some power laws can be found to characterise the diffusion time with both the size of the system and the increase in entropy.

Keywords: Data compression; Diffusion; Monte carlo; Entropy; Power laws

1. Introduction

A new territory was conquered for science thanks to the development of the theory of information [1]. In the information theory, we must start by giving a precise definition of the word "Information". The problem involves a certain number of possible answers. Information is a function of the ratio of the number of possible answers before and after a logarithmic law is retained to ensure the additivity of information. This theory is used in classical problems relating to information (coding, computer science, and telecommunication). In physics, there is a remarkable likeness between information and entropy which allows physicists to solve the problem of Maxwell's demon. Brillouin [2] proved that information could be considered as the negative of system's entropy called "negentropy". Entropy measures the lack of information; it gives the total amount of missing information on the ultramicroscopic structure of a system. However, the tools used to quantify negentropy remain theoretical mathematical objects rather than applied mathematical tools [3]. In another scientific field, information theory could be a overfull tool to analyse data compression. The more informative, the lower the ratio of data compression. With this evidence, Chaitin has defined the notion of randomness [4]. The definition of randomness could then be summed up as follows: let us define a discrete system of size N that could be described by a computer program: if this one cannot be compressed into a program (considered as a string of N bits) shorter that itself, then the system is random. Most surprisingly no theory was built to unify Chaitin's theory, data compression and the entropy of the system. We have then decided to build this theory

* Corresponding author. Tel.: +33-3-2062-2233; fax: +33-3-2062-2957.

E-mail address: alain.iost@lille.ensam.fr (A. Iost).

as we think that it could correspond to a way of describing a physical system. Quickly summarising our theoretical approach, we postulate the following axiom: a system will be integrally described if the size of the system after compression is minimal, however wide the data compression of all possible algorithms. The knowledge of the physical system is given by the data compression algorithm. Moreover the size of the program becomes a measure (from a mathematical point of view) of the negentropy. However, our mathematical theory is very hard to build because it requires a great number of mathematical fields (Set theory, Topology, Fractals, Statistics, Number theory...).

In this paper, rather than describing our mathematical tools, we decided to present an application of this theory in the field of Materials Science. Thanks to the Monte Carlo simulation of a diffusion process, we will prove that this theory allows us to characterise the state of diffusion and then to quantify the time to reach equilibrium. In a first part, we will describe the two important usual data compression classes of algorithm (run length encoding (RLE) [5] and Huffman coding [6]). In a second part, we will describe the mathematical tools used in this application. Then we will briefly describe the dynamic Monte Carlo simulation and apply the data compression algorithms to those systems, and finally analyse the measure of negentropy.

2. Mathematical definitions

A few mathematical definitions will now be introduced. Let X be a discretized system. We denote F the bijective function that transforms system X into system Y according to an algebra T. This function is then noted as $Y = F(X,T)$. Let $\mathrm{card}(X)$ be the size of the set X then $F(X)$ is said T-contracting if $\mathrm{card}(F(X,T)) < \mathrm{card}(X)$. Let Ω be a finite group of system X, then F said T-contractile in Ω if F is T-contracting and

$$\forall X, Y \in \Omega, \quad \mathrm{card}(X) \leqslant \mathrm{card}(Y) \Rightarrow \mathrm{card}(F(X,T)) \leqslant \mathrm{card}(F(Y,T)) \qquad (1)$$

We will introduce a normalised measure denoted $\overline{T}$ with $\overline{T}(X,F) = \mathrm{card}(X)/\mathrm{card}(F(X))$ which can be seen as a compression ratio and we will introduce a reduced normalised measure denoted $\widetilde{T}$ with, $\widetilde{T}(X,F) = \mathrm{card}(X)/[\mathrm{card}(F(X)) - \mathrm{card}(F(\emptyset))]$, with $\emptyset$ an informationless system (for example, a system without any atoms). $\widetilde{T}$ could be seen as the compression ratio reduced by the smallest element. All previous definitions could be applied to $\overline{T}$ and $\widetilde{T}$ transforms.

3. The data compression method

3.1. RLE data compression

RLE is a natural candidate for compressing graphical data [5]. A grid is made up of small dots called cells. Each cell is occupied by 1 bit, indicating the presence or the absence of a particle in the cell. Compressing a grid using RLE is based on the fact that if we select a cell image at random, it is highly probable that its neighbours should get the same state (atom or vacancy). A compressor then scans the cell in rows on the grid looking for runs of cells having the same state. The number of runs has reduced significantly the initial size.

3.2. Huffman data coding

This method [6] starts by building a list of the alphabet symbols in decreasing probability order. Then the smallest number is affected to the largest probability following the bit code of $1, 10, 111, 1101, 1100 \ldots$

4. Monte Carlo simulation

The phenomenon, which is presented above is the diffusion of an A^* atom in A. We have computed this system for two-dimensional systems represented by a N^2 Boolean matrix ($0 =$ vacancy; $1 =$ atom). We put $N^2/2$ A atoms on the right part of the matrix (Fig. 1).

The A jump elementary process occurs as follows: on average, for each iteration, each A^* atom can jump randomly in one of the four directions in order to occupy an empty neighbouring site.

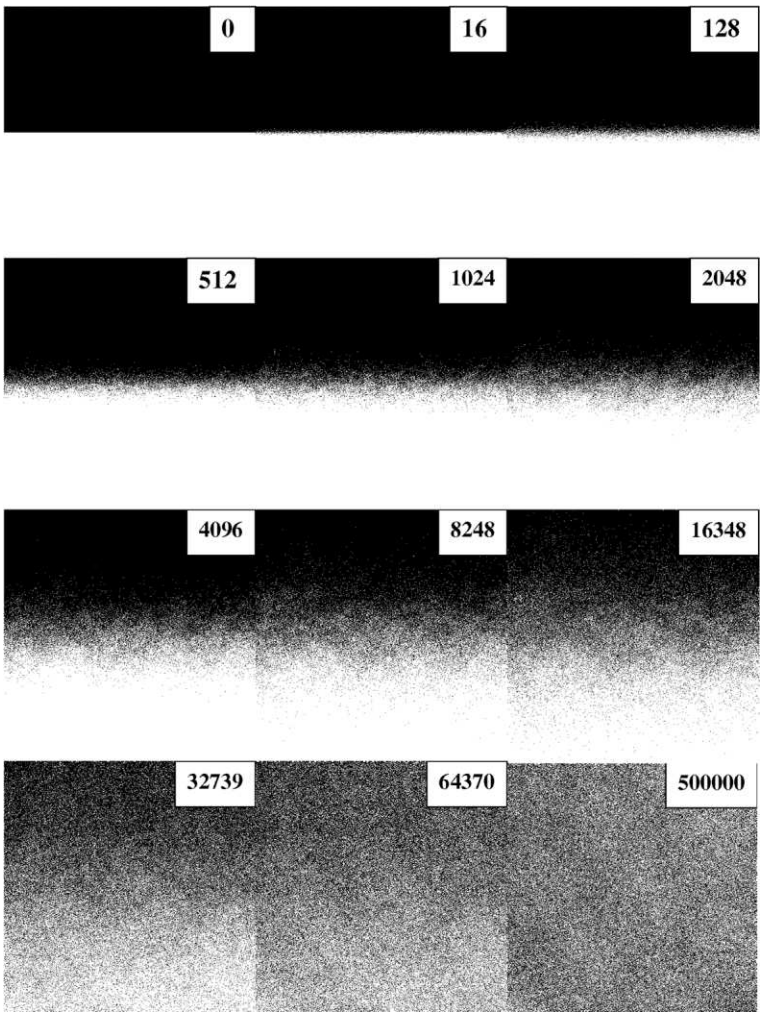

Fig. 1. Monte Carlo simulation of the diffusion of A^* atoms (upper part of the matrix) in A atom (lower part). Iterations represent the number of random jumps $(0, 16, 128, 512, \ldots, 500\,000$ MCS).

1,000,000 Monte Carlo simulations are carried out (MCS) for six different sizes of system (16^2, 32^2, 64^2, 128^2, 256^2, and 512^2 cells).

Then we apply the RLE algorithm on each system at different times. With our notation, $T = \mathrm{RLE}$, Ω is the union of systems of size n noted $X(t,n)$. Fig. 2 shows the $\mathrm{card}(X(t,512),\mathrm{RLE})$ versus simulation time. As can easily be shown, the size of $X(t,n)$ increases with time: diffusion increases the size of the algorithm description of the physical process. As the entropy of the system increases, the program size of the data compression increases. It seems that the RLE algorithm could be seen as RLE differential in time. Moreover, when the process reaches equilibrium, then an asymptote occurs leaving unchanged the size of $F(X(t,512),\mathrm{RLE})$. The RLE data algorithm allows us to quantify the time to reach equilibrium. To analyse more precisely this asymptote, we have simulated 100,000 random systems that represent the definition of the thermodynamical equilibrium. The entropy is then maximal. Then the algorithm RLE is applied to random systems of 512^2 cells noted and Random(512) and next:

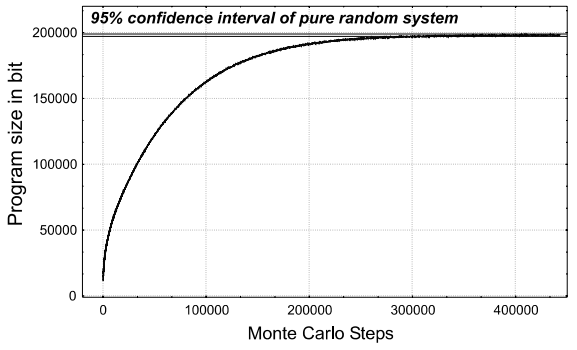

Fig. 2. Size (in bit) of the system of 512^2 cells shown in Fig. 1 versus diffusion time (in MCS).

$$\mathrm{card}(\mathrm{Random}(512), \mathrm{RLE}) = \lim_{t\to\infty} \mathrm{card}(X(t, 512), \mathrm{RLE}) \quad (2)$$

However, this assumption is only true on average and fluctuations are present with an amplitude that depends on $\mathrm{card}(X)$. Then the probability density function of $\mathrm{card}(\mathrm{Random}(512), \mathrm{RLE})$ is plotted (Fig. 3) and follows a Gaussian law that is similar to that of the entropy fluctuations of a system at equilibrium. Entropy variation entails information variation quantified by an algorithm and the size of reduction. Then the question is "Is the equilibrium reached and how long does it take?". We will prove that this question has no physical sense. We postulate that the equilibrium will be the time in which $\mathrm{card}(\mathrm{Random}(512), \mathrm{RLE}) = \mathrm{card}(X(t_e, \mathrm{RLE}))$. However this equality can only be true in a statistical sense with a given confidence interval. As we find a Gaussian PDF with mean $\mu_{\mathrm{card(Random(512),RLE)}} = 197{,}762$ and standard deviation of $\sigma_{\mathrm{card(Random(512),RLE)}} = 287$, at 95% for the confidence interval, we find $t_e = 288{,}000$ MCS. Does this mean that after 288,000 MCS, equilibrium is reached? In fact, in a philosophical sense, it is impossible to give either a positive or a negative answer. This theory only asserts that we cannot answer positively. In fact, we will prove that the quantity of information is not large enough to reject the state of equilibrium. Suppose that instead of observing a system, we observe n systems that follow the same dynamic structure. It is then possible to average the size of programs at each Monte Carlo step. As a consequence, the standard deviation of the mean will be equal to $\sigma_{\mathrm{card(Random(512),RLE)}}/\sqrt{n}$ and t_e will increase with the number of observations: the more frequently we observe a dynamic system, the longer the time of equilibrium and it becomes impossible to assert that equilibrium is reached.

We will now analyse $\tilde{T}(X(t), \mathrm{RLE}, \{X(t_0), X(t_1) \ldots X(t_n)\})$ for different initial sizes. A first study has shown that results will be accurate if $\widetilde{T}$ is used rather than $\overline{T}$. $\mathrm{cards}(\mathrm{RLE}(\emptyset))$ is the size of the system with no atom after RLE algorithm reduction. Fig. 4 represents the evolution of the compression ratio $\widetilde{T}$ versus diffusion time for the six different sizes of the system. For all images, compression will be more and more difficult with the heating time. For short diffusion time, sur-

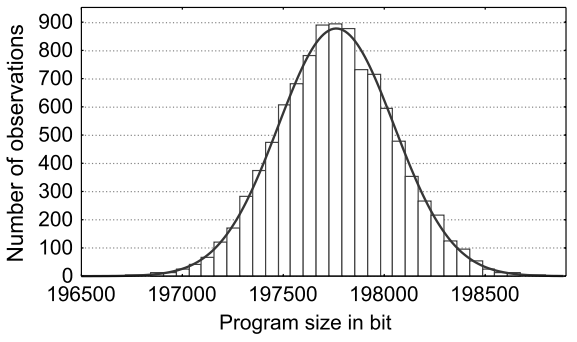

Fig. 3. Probability density function of program size (in bit) of 100,000 random systems of 512^2 cells.

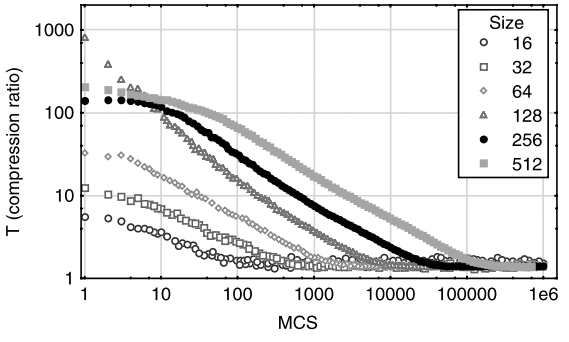

Fig. 4. Evolution of the compression ratio $\widetilde{T}$ (with RLE algorithm) versus diffusion time for six different sizes of the system shown in Fig. 1.

prising results could be observed. For the same diffusion time, the system of size 128^2 is less compressible than systems of 256^2 and 512^2. This problem is characteristic of information coding in the RLE algorithm because the quantity of elementary information is coded on a pre-set number of bits and adding a cell could suddenly increase the size of the system. To avoid this artefact, the Huffman algorithm compresses the system again and the redundant information due to the border effect will be reduced.

From a mathematical point of view, a new application $\tilde{T}(X, \text{Huffman} \circ \text{RLE})$ is defined. Fig. 5 shows the variation of the ratio $\tilde{T}$ versus heating time. Numerous remarks have to be made. Firstly, curves present a linear part (in a log–log scale) from the beginning to a critical value, and after t_e, an asymptote appears. For the first part, all the slopes converge to -0.4, meaning that $\tilde{T} \propto t^{-0.4}$, which could be seen as a characteristic exponent of the diffusion speed that does not depend on the size of the system. This exponent globally represents the kinetics of the diffusion process. The second part represents the state of equilibrium.

We have simulated 10,000 systems randomly chosen, which represent the "ideal" equilibrium and we have calculated the descriptive statistics with their associated confidence intervals (Table 1). For each size of the system, the t_e time when equilibrium is reached, is statistically calculated (more precisely, it is the time which the equilibrium state could not be rejected at the 95% confidence level). Then $\log(t_e)$ is plotted (Fig. 6) versus heating time (in log). The points obey a linear fit with a slope of the 2.22 for the regression line meaning that $t_e \propto s^{2.22}$, where s^2 is the number of elementary cells: the equilibrium time follows a power law with respect to the size of the system.

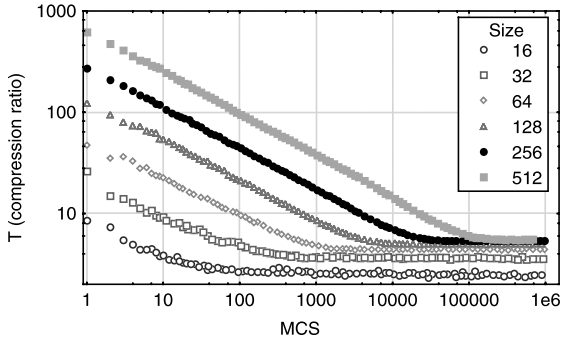

Fig. 5. Evolution of the compression ratio $\tilde{T}$ versus diffusion time for six different sizes of the system shown in Fig. 1 with both RLE and Huffman algorithms reduction.

Table 1
Descriptive statistics of ratio of compression (in log) for 10,000 random systems with different sizes

S	Mean	Std	95% Min	95% Max	t_e
16	0.43887	0.014957	0.40896	0.46878	100
32	0.57852	0.00708	0.56436	0.59268	500
64	0.6531	0.003531	0.64604	0.66017	2000
128	0.70316	0.001825	0.69951	0.70681	10,000
256	0.72814	0.000979	0.72618	0.7301	50,000
512	0.7398	0.000469	0.73886	0.74074	220,000

t_e is the time of equilibrium in MCS.

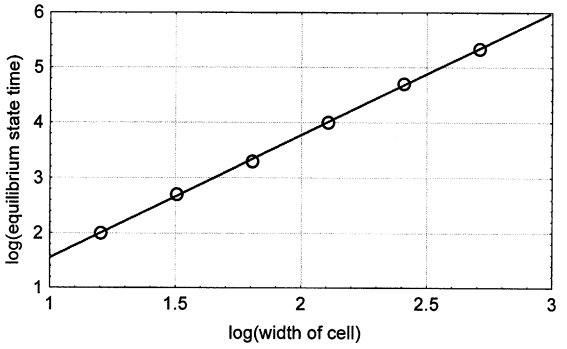

Fig. 6. Equilibrium time (in MCS) versus size of the system (in width).

5. Conclusion

A new method of process characterisation by data compression was presented. This first result shows that some power laws characterise the state of equilibrium and its associated speed to reach this equilibrium. However, the accuracy of the method depends on the reduction algorithm. Some investigations were carried out to introduce this mathematical formalism in the field of Materials Science.

References

[1] C.E. Shannon, Bell Syst. Tech. J. 27 (1948) 379–423, and 623–656.

[2] L. Brillouin, Science and Information Theory, Academic Press, New York, 1956.

[3] W.H. Zurek, Complexity, Entropy and the Physics of Information, Addison-Wesley, 1990.

[4] G.J. Chaitin, Algorithmic Information Theory, Cambridge University Press, 1987.

[5] D. Salomon, Data compression, Springer, New York, 1998.

[6] D. Huffman, Proc. IRE 40 (9) (1952) 1098–1101.