Bulletin of the *Transilvania* University of Braşov • Vol 6(55), No. 2 - 2013
Series III: Mathematics, Informatics, Physics, 87-94

ON AN APPLICATION OF ENTROPY

Nicoleta ENACHE DAVID[1]

Abstract

In this paper some properties of the entropy of a stationary Markov chain are presented. Using the entropy rate we determine the length of the typical navigation trails in a web application. Our approach can be used for adaptable web sites development and for advertising on Internet.

2000 *Mathematics Subject Classification:* 60J22, 68Q01.
Key words: entropy rate, Markov chain, web application.

1 Introduction

Information theory is a branch of applied mathematics that involves the quantification of information. There are applications in computer networks, statistics, biology, data analysis and other research fields, but information theory is generally used for the study of information transmission systems.

In literature there are connections between information theory and other domains like probabilities theory and statistics, physics, economy. In probabilities theory, the notions of entropy and relative entropy are defined as distribution functions and they characterize the behavior of long random variables sequences. In economy, especially in the investments field, there is a duality between the rate of benefit increase on the market and the market entropy rate.

This paper is structured as follows: in Section 2 we present some properties of the entropy and the Asymptotic Equipartition Property (AEP) that will be used to determine the lengths of navigation trails on a web application. In Section 3 we present a Markov chain model for a web application and in Section 4 we give some introductory notions about LOG files. Finally, in Section 5 we give our experimental results obtained with this model and in Section 6 we give our concluding remarks.

2 Some properties of the entropy

In this section we present some properties of the entropy and the Asymptotic Equipartition Property (AEP) introduced by C.E. Shannon in [9] where he demonstrated the result for i.i.d. processes and for stationary ergodic processes. Lately, P.

[1] Faculty of Mathematics and Informatics, *Transilvania* University of Braşov, Romania, e-mail:nicoleta.enache@unitbv.ro

Bremaud in [3] and B. McMillan in [7] have demonstrated AEP for ergodic alphabet sources.

We consider a random variable X with density function $F_X(x) = P(X \leq x)$. If $F_X(x)$ is continue, then the random variable X is continue. We denote by $f(x) = F_X^{'}(x)$ if the derivative is defined. If $\int\limits_{-\infty}^{\infty} f(x) = 1$ then $f(x)$ is the repartition function of the random variable X. The set S where $f(x) > 0$ is named the support set of the random variable X.

We denote by $h(X)$ the differential entropy of a continuous random variable X having the density function $f(x)$.

Definition 1. *[4] The differential entropy is defined as*

$$h(X) = -\int\limits_S \log f(x)dx, \tag{1}$$

where S is the support set of the random variable X.

The differential entropy depends only on the probability density of the random variable, so in some cases the differential entropy is denoted by $h(f)$.

Theorem 1. *[4] (AEP) Let us consider a sequence of i.i.d. continuous random variables $X_1, X_2, ..., X_n$ with density function $f(x)$. Then we have*

$$-\frac{1}{n}\log f(X_1, X_2, ..., X_n) \overset{P}{\to} M(-\log f(X)) = h(X). \tag{2}$$

In [4] the authors define the typical set for the continuous case and give the properties of the typical set.

Definition 2. *For $\epsilon > 0$ and any n, we define the typical set $A_\epsilon^{(n)}$ in concordance with the density function $f(x)$ so that*

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, ..., x_n) \in S^n : |\frac{1}{n}\log f(x_1, x_2, ..., x_n) - h(X)| \leq \epsilon\right\}, \tag{3}$$

where $f(x_1, x_2, ..., x_n) = \prod\limits_{i=1}^{n} f(x_i)$.

From the previous theorem one can see that

$$-\frac{1}{n}\log f(x_1, x_2, ..., x_n) = -\frac{1}{n}\sum_{i=1}^{n}\log f(x_i) \overset{P}{\to} h(X).$$

It follows that

$$P\{A_\epsilon^{(n)}\} > 1 - \epsilon \tag{4}$$

for n sufficiently big. This property is very important in our current approach.

Let us consider a stochastic process denoted by $(X_1, ..., X_n)$.

Definition 3. *The entropy rate of a stochastic process $(X_1, ..., X_n)$ is defined as*

$$H(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, ..., X_n), \tag{5}$$

if the limit exists.

We consider now a stationary Markov chain $(X_1, X_2, ...)$ having a stationary distribution $\pi = (\pi_1, ..., \pi_k)$ and the transition matrix $P = \{P_{i,j}\}_{i,j=\overline{1,k}}$. We have the following theorem [4]:

Theorem 2. *The entropy rate of a stationary Markov chain $(X_1, X_2, ...)$ having a stationary distribution π and the transition matrix P is*

$$H(\mathbb{X}) = -\sum_{i,j=1}^{k} \pi_i P_{ij} log P_{ij}. \tag{6}$$

3 A Markov chain model

In this section we present a Markov model for a web application. We define the notion of typical navigation trail in a web application and we determine the length of typical trails.

In literature there are various probabilistic models for web applications. In [8] the authors propose a Markov model that predicts the next web page accessed by a visitor. In [6] the authors present a theoretical approach of Web navigation in the terms of the entropy of a Markov chain that models the web topology.

A very used approach is that a web application can be modeled as a stationary Markov chain

$$M = (Y_1, Y_2, ...),$$

having the set of states $X = \{X_1, X_2, ..., X_k\}$, transition probabilities matrix $P = \{P_{i,j}\}_{i,j=\overline{1,k}}$, the initial probabilities $\mu = \{\mu_1, \mu_2, ..., \mu_k\}$ and following the property:

$$\begin{aligned} P(Y_{n+1} &= X_j | Y_0 = X_{i_0}, Y_1 = X_{i_1}, ..., Y_{n-1} = X_{i_{n-1}}, Y_n = X_i) \\ &= P(Y_{n+1} = X_j | Y_n = X_i) \\ &= P_{i,j} \end{aligned} \tag{7}$$

for all $i, j \in \{1, ...k\}$, all $i_0, ..., i_{n-1} \in \{1, ..., k\}$ and any positive integer n.

In this model the states $X_1, X_2, ..., X_k$ are the web pages of the website and $P_{i,j}(i, j = \overline{1,k})$ represent the transition probabilities from page X_i to page X_j. The elements $P_{i,j}(i, j = \overline{1,k})$ are calculated as the rapport between the number of transitions from X_i to page X_j and the total number of transitions from X_i ([2]).

Let us consider a navigation trail $T = (X_1, X_2, ..., X_t)$ of length $t \geq 2$ in a web application.

Definition 4. *[1] The navigation trail $T = (X_1, X_2, ..., X_t)$ is typical if the estimated probability*

$$P(T) = P(X_1) \cdot \prod_{i=2}^{t} P(X_i|X_{i-1}) \tag{8}$$

exceeds a threshold $\lambda \in (0,1)$.

The parameter $\lambda \in (0,1)$ depends on $k \in N^*$, the number of pages in the web site and on $m \in N^*$, the average number of links on a page. We have

$$\lambda = \frac{1}{k}\frac{1}{m}. \tag{9}$$

Definition 5. *The set of all navigation trails with the length $t \in N^*$ denoted by Γ_t is defined as*

$$\Gamma_t = \{T|T = (X_1, X_2, ..., X_t), t \in N^*\}. \tag{10}$$

AEP can be applied in this case in the following way: if certain web pages appear more frequently in the set Γ_t, there is a subset denoted by $\Psi_t \subseteq \Gamma_t$ that contains these web pages. It follows that $P(\Psi_t)$ is closed to 1, respectively the navigation trails that are not in the subset $P(\Psi_t)$ appear very rarely [1].

It follows that the length t of a typical navigation trail is

$$t \leq \frac{-ln\lambda}{H(M)ln2}, \tag{11}$$

where $\lambda \in (0,1)$ is calculated according with formula (9).

4 LOG files and data processing

LOG files are the files that store information about the activity of a user on a web site, respectively the pages he visited, the computer IP, date and time, browser type, user name if it exists, etc. Recent researches have been conducted to the LOG files analysis, in order to discover navigation patterns on a web site. Navigation patterns can offer support for improving the web site's functionality.

For example, the web site can be improved with a recommendation system, if the navigation patterns suggest the user's sections of interest. Much more, an online bookstore can use such a recommendation system to suggest certain books that can be purchased by the user, taking into account his previous preferences.

The most used format for LOG files is W3C Extended, elaborated by the World Wide Web Consortium ([10]).

5 Experimental results

We consider a web site modeled as a stationary Markov chain $M = (Y_1, Y_2, ...)$ having 10 states. We use a collection of navigation trails extracted from the LOG files of the web site in the period 01.01.2012-01.04.2012. We have eliminated the irrelevant navigation trails, respectively those rarely visited that are not important for our study.

Applying the formula (9) for the Markov chain, we have obtained $\lambda = 0.03$. From the formula (11) and with the parameter $\lambda = 0.03$ it results that the length of the typical navigation trails is $t \leq 2.53$.

In the figures bellow we present the relative frequencies of the lengths of typical navigation trails for our data sets. One can observe that the biggest relative frequencies are obtained for the lengths of trails less or equal with 2.53.

We conclude that the trails having the lengths $t \leq 2.53$ appear with biggest relative frequencies, respectively between 15% and 33% in Figure 1, between 16% and 19% in Figure 2 and between 8% and 16% in Figure 3.

In conclusion, the typical navigation trails have the lengths $t \leq 2.53$.

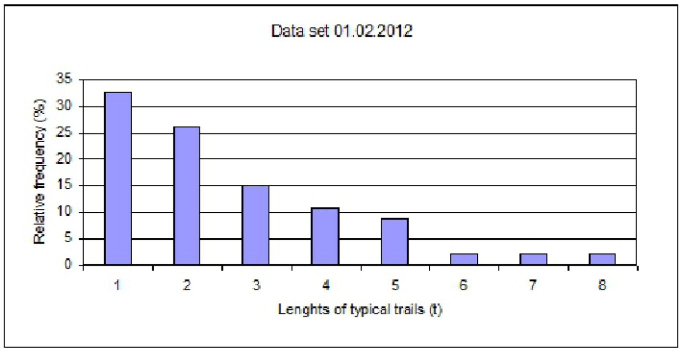

Figure 1: The lengths of typical trails for the data set 01.02.2012.

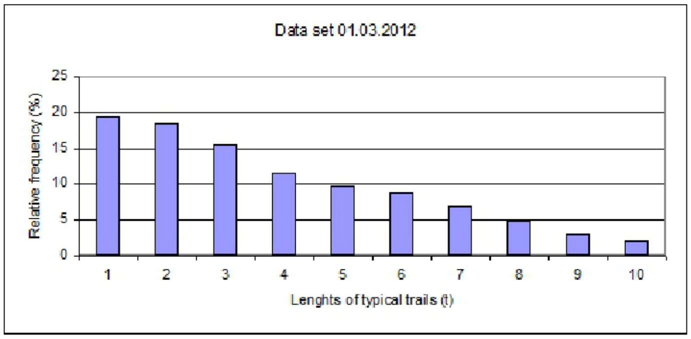

Figure 2: The lengths of typical trails for the data set 01.03.2012.

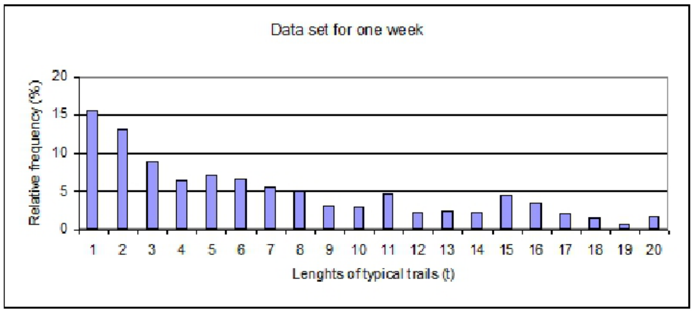

Figure 3: The lengths of typical trails for one week.

6 Conclusion

In literature there are many models used for the study of the behavior of users during the navigation on a web site. Among all these, Markov models play an important role. In this paper we have used the entropy rate of a Markov chain in order to determine the lengths of the typical trails. In practice, the computation of the lengths of the typical navigation trails on a web site can be used in:

- Web site personalization, by recommendations for certain pages that the user is interested in;
- Internet advertising, by placing the banners on the pages that are on the typical navigation trails.

References

[1] Borges, J., Levene, M., *An average linear time algorithm for web usage mining*, International Journal of Information Technology and Decision Making **3(2)** (2004), 307-319.

[2] Borges, J., Levene, M., *Testing the Predictive Power of Variable History Web Usage*, Soft Computing - A Fusion of Foundations, Methodologies and Applications, **11** (2007).

[3] Bremaud, P., *Markov Chains: Gibbs fields, Monte Carlo Simulation and Queues*, Springer, New York, 1998.

[4] Cover, T.M., Thomas, J. A., *Elements of Information Theory*, John Wiley and Sons Inc., 1991.

[5] David, N., *Informational Systems and Informatic Applications in Businesses*, LAP Lambert Academic Publishing, Saarbruken, 2011.

[6] Levene, M., *Loizou, G., Computing the Entropy of User Navigation in the Web*, International Journal of Information Technology and Decision Making **2 (3)** (2003), 459-476.

[7] McMillan, B., *The Basic Theorems of Information Theory*, The Annals of Mathematical Statistics, **24(2)** (1953), 196-219.

[8] Sarukkai, R.R., *Link prediction and path analysis using Markov chains*, Proceedings of the 9th International World Wide Web Conference, Amsterdam, Holland, 2000.

[9] Shannon, C.E., *A Mathematical Theory of Communication*, Bell System Technical Journal, **27** (1948), 379-423.

[10] World Wide Web Consortium, http://www.w3.org.