SCIENTIFIC PAPERS
OF THE UNIVERSITY OF PARDUBICE
Series B
The Jan Perner Transport Faculty
12 (2006)

MODELLING OF TRANSPORT PROCESS BY UNARY CODE

Martin KLIMO, Jana URAMOVÁ

Department of information networks, Faculty of management sciences and informatics,
University of Žilina

1. Introduction

Prof. Černý and prof. Kluvánek have created a mathematical theory of transportation, and the basic ideas have been published in 1991 [2]. At least since this date it is well known, that transport and telecommunication network models have the same theoretical roots. The mathematical background consists mainly of graph theory for optimisation of network topology and queuing theory for optimisation of transport system performance. Anyhow, the communication theory has a strong tool at hand - the information theory, and its potential is still unused within the transport theory. In 1996 Anantharam and Verdú has published a paper [3], in which they used the information theory for performance analysis of packet networks for the first time. It seems to be clear, that if performance of transport and packet systems have the same modelling principles, using of information theory should bring some gain also for modelling of transport networks.

Comparing with [3] we have used an alternative approach, finding that there is a direct analogy between the interval between two subsequent transport elements and the word of unary binary code. Properties of these codes have not yet been studied, because of their inefficiency for information transmission. Applying them to studies of intervals in transport flow, their understanding can bring new knowledge for qualitative analysis of transport elements stream.

2. Flow representation by unary code

With respect to throughput modeling, the basic properties of transport elements are described by their placement on a time axis. In addition to this, the individual properties of particular transport elements are included, e.g., transportation type, elements types, etc., but in this article, we will not take these properties into consideration and we will only focus on time positions of transport elements. To simplify our modeling, we will assume, that we can divide a transport system with transport flow present into sections in such a way, that transport element will stay in each section for an equal amount of time. On one hand side, this means simplification, while the time set is countable. One the other hand side, it means, we will restrain the transport flows observation to only those parts of the system, in which the elements cannot stay. This however, is in compliance with the purpose of the article, which describes information properties of transport flows. We will call the process of holding an element in such section a time slot.

From this point of view, the elements flow is fully described by a sequence of intervals between them.

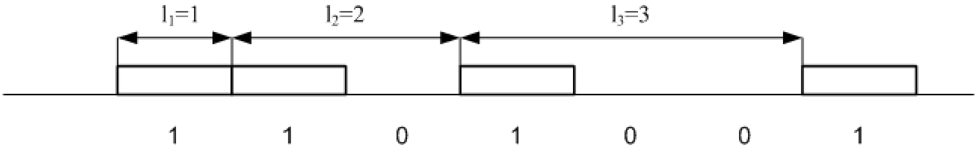

***Fig. 1** Transport elements flow*

We assume that all time slots are of equal length and in each time slot at most one element can appear. Thus, the interval between elements is a natural number. If we designate a time slot in which an element has appeared, by symbol “1”, and time slot in which no element has appeared, by symbol “0”, then the elements flow is described by a sequence of zeros and ones. Thus, each interval corresponds to a sequence of zeros terminated by a one. We will call this sequence a code word.

Interval length	Code word
1	1
2	01
3	001
...	...

***Fig. 2** Unary code words*

Such code words form unary bar code, in which a corresponding number of zeros is terminated by a one bar. To simplify our modeling, we will only call this code unary code, as usually. Thus, intervals between elements are unary code words, which means, the point processes problems are transformed into coding area. This provides a basis for answering questions related to amount of information contained in code words which are generated by a source, or transferred by an information channel. Using these notions, we will be able to talk about the amount of information contained in the position of elements on input or output time axis of a transport subsystem.

In order to be able to use source coding theory, we will not say, that a source generates elements between which the interval is of length l_i with probability p_i, but we will say, that the source has generated the symbol x_i with probability p_i, and this symbol is encoded into a code word (00.01) of length l_i.

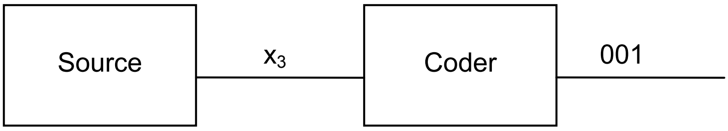

***Fig. 3** Coding of source symbols with unary code*

We will assume, that the source generates symbols of a finite set $X = \{X_1, X_2, ..., X_L\}$ with probabilities $p(x_1), p(x_2), ..., p(x_L)$, which form a probability distribution, i.e., $\sum_{k=1}^{L} p(x_k) = 1$. In the following text, we assume $p(x_L) > 0$. These symbols are generated in such a way, that one symbol is generated in each time slot. We assume, that for each n-tuple of generated symbols beginning in time $i, i \in N$ $x^i = \left(x^{(i)}, x^{(i+1)}, ...x^{(i+n-1)}\right)$ the probability of such sequence is given as follows $p(x^i) = p\left(x^{(i)}, x^{(i+1)}, ...x^{(i+n-1)}\right)$. Furthermore, we will only assume stationary sources, which means, that this probability is not dependent on *i*. Since intervals between transport elements, i.e., symbols of sources are randomly generated, the appearance of each interval (symbol on source output) brings the following information with $I(x) = -p(x)\log p(x), x \in X$. Mean amount of information contained in one symbol on source output is as follows:

$$H(X) = \mathrm{E}_X\{I(x)\} = \sum_{i=1}^{L} p(x_i)\square(x_i) = -\sum_{i=1}^{L} p(x_i)\log p(x_i)$$

and is called entropy of source $\{X, p(x)\}$. If not stated otherwise, we will assume logarithm of basis 2 and the information (i.e., entropy) is expressed in Shanons [Sh]. Thus, in our case, entropy gives the mean amount of information contained in one interval between two elements following one after the other. We also interpret entropy as an amount of interval uncertainty, where deterministic intervals are of zero uncertainty.

Each source symbol represents an interval of a given length, and unary code word is assigned to each interval length (see ***Fig. 4***). Shortly we will say that a code word of unary code is assigned to a source symbol.

Source symbol	Probability of source symbol	Code word	
x_1	$p(x_1)$	c_1	1
x_2	$p(x_2)$	c_2	01
x_3	$p(x_3)$	c_3	001
...	...	...	...
x_L	$p(x_\mathrm{L})$	c_L	00...01

Fig. 4 *Unary code words assigned to source symbols*

By the length of a code word we will designate the number of zeros and ones, with the help of which the code word is written: $m(c(x_i)) = |c(x_i)|, i = 1,....,L$. It is evident, that for a selected unary code, the following statement is true: $m(c(x_i)) = i, i = 1,....,L$.

Theorem 1

Unary code is immediately decodable.

Proof

Although the theorem is trivial, while the code is bar code, i.e., the letter of code alphabet “1” means the end of the code word, we will do a formal proof. Kraft unequality [1] says, that the binary code is immediately decodable IFF the following is satisfied:

$$\sum_{i=1}^{L} 2^{-m(c(x_i))} \leq 1$$

While the code word length in unary code is $m(c(x_i)) = i$, the left side of Kraft unequality is:

$$\sum_{i=1}^{L} 2^{-i} = 1 - \frac{1}{2^L} \leq 1$$

while the equality occurs when L grows to infinity.

Although this theorem does not bring anything new to transport elements flows, from the formal point of view it is necessary for application of information theory results in transport flows.

The code word length m is random variable, which is given by a probability distribution $P(m = i) = p(x_i), i = 1,....,L$. While later we will be interested in, in what time the information is transferred inside the code word, we will use the mean length of a code word $\overline{m} = E_X\{m(c(x))\} = \sum_{i=1}^{L} ip(x_i)$.

Example 1

Let us suppose the following source:

$\{X, p(x)\} = \{x_1, x_2, x_3, x_4, p(x_1) = 0.1, p(x_2) = 0.3, p(x_3) = 0.4, p(x_4) = 0.2,\}$

Entropy of this source is $H(X) = 1.85\,\mathrm{Sh}$. After encoding to unary code, we have the code words $c(x_1) = (1), c(x_2) = (01), c(x_3) = (001), c(x_4) = (0001)$, and the middle length of a code word is $\overline{m} = 2.7$.

This source represents transport elements flow, in which the intervals occur between elements of lengths 1,2,3,4 of slots with probabilities mentioned above.

It is evident, that we can get the same mean information (entropy), which is contained in one interval (unary code word), by changing the symbols order in the source.

Let us create the source as follows: $x_1' = x_3, x_2' = x_2, x_3' = x_4, x_4' = x_1$. Then the source $\{X', p(x')\} = \{x_1', x_2', x_3', x_4', p(x_1') = 0.4, p(x_2') = 0.3, p(x_3') = 0.2, p(x_4') = 0.1\}$ keeps equal entropy $H(X') = 1.85\,\mathrm{Sh}$, but the mean length of a unary code word is $\overline{m}' = 2 < \overline{m}$. Thus a question arises, which re-ordering of symbols leads to the shortest mean length of a unary code word.

Definition 1

If the source $\{X, p(x)\}$ is first re-ordered into the source $\{X', p(x')\}$, and then encoded by unary code, we will call such unary coding of a source $\{X, p(x)\}$ an optimal one, if no other re-ordering exists $\{X, p(x)\}$ which mean length of unary code word is shorter.

Theorem 2

For a source $\{X, p(x)\}$ the following is true $p(x_1) \geq p(x_2) \geq .... \geq p(x_L)$ IFF, the unary code of the source is the optimal one.

Proof

Let the unary code of source $\{X, p(x)\}$ be optimal and let exist *i* > *j*, for which $p(x_i) > p(x_j)$. By changing these symbols a source is created, which unary code has the mean length lower according to this value

$(ip(x_i)+jp(x_j))-(jp(x_i)+ip(x_j))=(i-j)(p(x_i)-p(x_j))>0$, which contradicts the presupposition, that the unary code of source $\{X,p(x)\}$ is optimal.

Thus, when we re-order the source symbols assignments to the intervals lengths between elements, in case of unoptimal division of intervals lengths we can reach the same mean information contained in one interval with the help of a shorter mean length of a unary code. In reality, the division of intervals lengths between transport elements is given, and it is not possible to change it. Then a question arises, why we should study properties of such re-ordering. The reason is, when we know the properties of optimal unary code, i.e., a flow in which the highest mean amount of information is related to one time slot, we will know a flow with the maximum uncertainty, to which the other flows can be related.

For demonstration purposes, we will represent a code by a code tree. A code tree of a unary code of a source $\{X,p(x)\}$ is illustrated in the following figure.

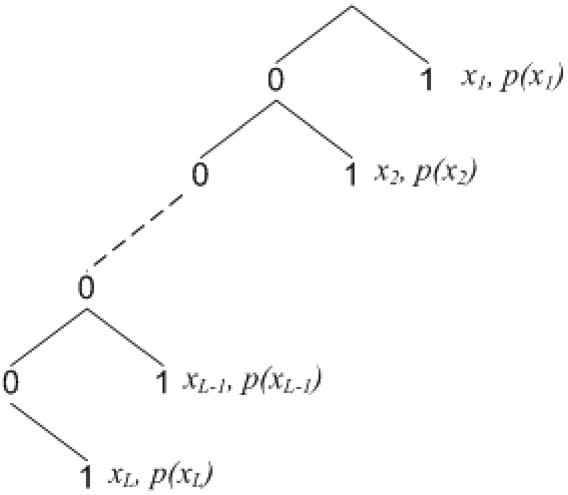

***Fig. 5** Unary code tree for the source $\{X,p(x)\}$*

3. Unary code equivalent to Huffman code

Huffman code is well known (see e.g. [1]), it is immediately decodable and has the shortest mean length of a code word. Next, we will only be using the binary Huffman code. We will use its construction as follows:

- We will order the symbols from left to right in an ascendant way according to probabilities.

- We will assigned codes letters “0” and “1” to the first two symbols on left hand side (with the lowest probability) and we will group them into a new symbol, which probability equals the sum of probabilities of grouped symbols. If the sum is lower then 1, we continue using the previous rule.
- We will stop the code tree beginning at the top level.

An example of such construction is illustrated in the following figure.

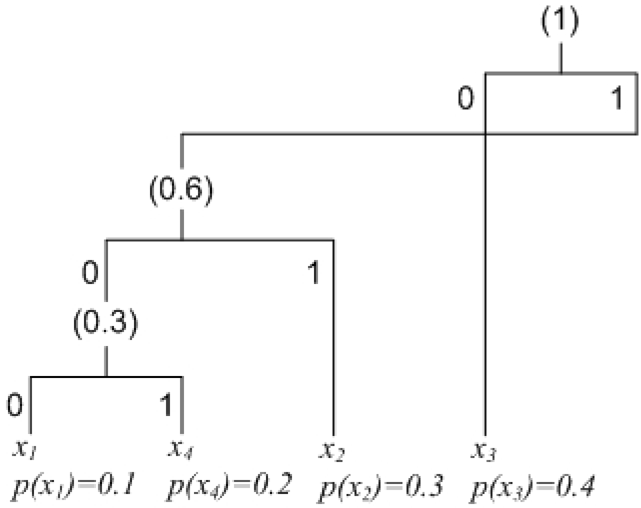

Fig. 6 *Construction of Huffman code for the source*

The respective code tree is illustrated in the following figure.

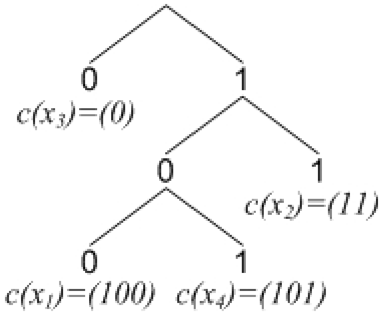

Fig. 7 *Huffman code tree for the source*

Let us look at the source from Example 1. Its entropy is $H(X) = 1.85\,\mathrm{Sh}$. The mean word length of Huffman code is as follows word length of Huffman code is as follows $\overline{m} = 1p(x_3) + 2p(x_2) + 3[p(x_1) + p(x_4)] = 1.9$. Let us remind ourselves, that the mean length of unary code of this source is $\overline{m} = 2.7$ and mean length of the optimal unary code of this source is $\overline{m} = 2$. The example only confirms a well known fact [1], that the Huffman code achieves the minimal mean code word length from all immediately decodable codes, and from above it is approaching the entropy of the encoded source. To complete our notions we add, that it is not the only one with this property and for example Shannon-Fan code achieves this mean code word length too. However, for some sources, the unary code may be almost equal to Huffman code.

Definition 2

We say, that unary code $C_U = \{c_U(x_i), i = 1,....,L\}$ is equal to Huffman code $C_H = \{c_H(x_i), i = 1,....,L\}$, if

$$c_U(x_i) = c_H(x_i), i = 1,....,L-1 \quad \text{a} \quad c_U(x_L) = (c_H(x_L), 1)$$

i.e., all unary code words except the longest one are words of Huffman code and the longest word of unary code is a word of Huffman code after releasing the symbol “1”.

Theorem 3

Unary source code $\{X, p(x)\}$ is equal to Huffman code of this source iff,

$$p(x_k) \geq \sum_{i=k+1}^{L} p(x_i) \ , \quad k = 1,...,L-1 \ , \tag{1}$$

Proof

Equality of unary code and Huffman code is identical to the following request: During the Huffman code construction the rearrangement of symbols of reduced sources according the probability is not desired, i.e., the newly created symbol of reduced source has the lowest probability. Then the symbol “0” will be assigned to it. If the unequality in a theorem is satisfied, then there will be no rearrangement and after adding “1” the resulting code will be also the Huffman code. The other way around, if the unary code is also the Huffman code, then there was no rearrangement of symbols of reduced sources, which means, the unequality in the theorem is satisfied.

It is evident, that the unary code, which is equal to Huffman code is also optimal code, i.e., $p(x_1) \geq p(x_2) \geq .... \geq p(x_L)$.

Example 2

Suppose source:

$$\{X,p(x)\} = \{x_1,x_2,x_3,x_4;p(x_1) = 0.5,p(x_2) = 0.3p(x_3) = 0.1p(x_4) = 0.1\}.$$

The construction of Huffman code of this source is illustrated in the following figure.

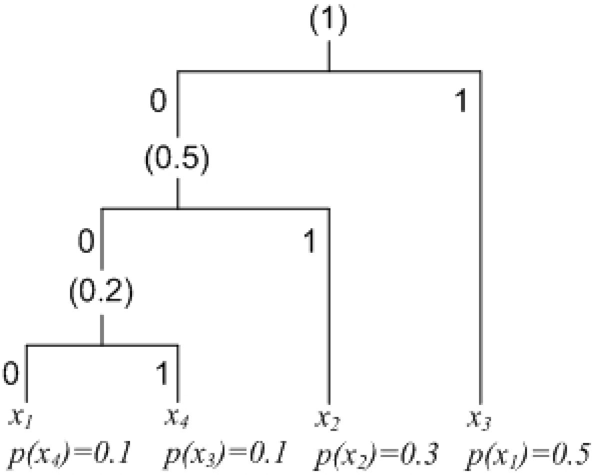

Fig. 8 *Construction of Huffman code of given source*

Huffman code and unary code trees of given source are illustrated in the following figure.

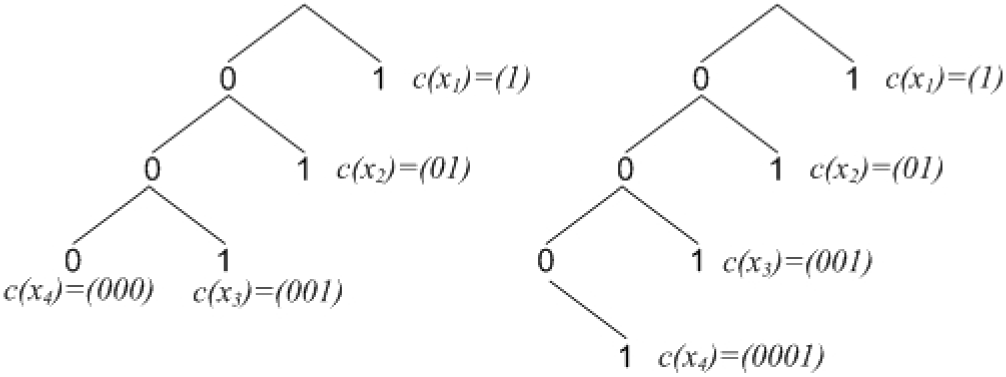

Fig. 9 *Huffman code and unary code trees of given source*

Entropy of this source is $H(X) = 1.69\,\text{Sh}$. After encoding into Huffman code, its mean length is $\overline{m}_H = 1.7$ and after encoding into unary code, the mean length of a code word is $\overline{m}_U = 1.8$.

If the unary code C_U is equal to Huffman code C_H, their mean length of a code word only differs in a contribution of added “1” to the last word, i.e., $\overline{m}_U = \overline{m}_H + p(x_L)$.

4. Unary code with the shortest mean length

The previous theorem answered the question related to the condition which the source must satisfy in order for its unary code to be equal to Huffman code. However, which source from those, which satisfied the given condition, had the shortest mean length of word? The following theorem is related to this question.

Theorem 4

From these sources $\{X, p(x)\}$, $X = \{x_1,...,x_L\}$, the source with the following distribution has the shortest mean word length of unary code.

$$p(x_i) = 2^{-(i+1)},\ \ i = 1,...,L-2,\ 0 < p < 1$$
$$p(x_{L-1}) = 2^{-(L-1)} - \varepsilon$$
$$p(x_L) = \varepsilon,\ \varepsilon \to 0^+$$

Proof

We will get the Huffman code equal to unary code by removing the symbol “1” from the longest word of the unary code. From the basic theorem of sources encoding it is well known, that Huffman code will achieve the shortest length if

$$p(x_i) = 2^{-(i+1)},\ \ i = 1,...,L-1$$

$$p(x_L) = p(x_{L-1}) = 2^{-(L-1)}$$

However, the mean length of Huffman code will stay the same, if we move arbitrary part of probability $p(x_L)$ to probability $p(x_{L-1})$, since the code words are of the same length. This is not valid for unary code, which mean length of code word is $\overline{m}_U = \overline{m}_H + p(x_L)$ and thus the lower the probability $p(x_L)$, the lower it will be. But it must be non-zero, in order to be able to assign a code word to a symbol of a source x_L and this way to keep the condition of maximal length of a word satisfied.

5. Entropy rate of a transport flow

It is a well known fact, that the highest entropy has the interval between transport elements at uniform distribution $p_i = L^{-1}, i = 1,...,L$ and its value is $H = log\, L$. This however, does not mean that a flow with such interval distribution will have the highest uncertainty for a given time.

We have to realize, that while interval distribution is not interval distribution with the highest entropy, it may lead to the highest entropy contained in one time slot and thus leading to a flow with the maximum uncertainty. This is caused by the fact that lower entropy compared to uniform distribution can by compensated by shorter mean length of interval. Thus in order to be able to compare flows of transport elements from the uncertainty point of view, we will re-calculate the entropy of an interval for one time slot.

Definition 3

Let $C_U = \{c_U(x_i), i = 1,...,L\}$ be a unary code of a source $\{X, p(x)\}$, $X = \{x_1,...,x_L\}, L \geq 1$ and $\overline{m}_U$ the mean length of this code. The following quantity we call the entropy rate of the unary code C_U

$$h_U = \frac{H(X)}{\overline{m}_U}.$$

Another article will be dedicated to the searching for a source with the maximum entropy rate. As the reader probably anticipates, for $L \to \infty$ it will be a source, which probability distribution is approaching the geometric one with a quotient of $p = \frac{1}{2}$. However, for sources with finite number of symbols the situation is more complicated.

6. Conclusion

Interpreting intervals between elements in a transport flow as words of unary code allows us to use information theory results for its studies. In this article, we have mostly concerned with flows, in which the construction of unary code leads to Huffman code, a code with the shortest mean length. This means, we have been mostly concerning with flows, which have the shortest possible mean interval length, while they have non-changing uncertainty. The fact, that Huffman code contains in its longest code word one "1" less compared to equal word of unary code (interval) means, that we cannot automatically take the results about uncertainty related to one symbol of Huffman code also for an uncertainty, which is related to one time slot of a transport flow. This means, that we have left the problems of transport flows with the highest uncertainty to the next article. In order to make the reader understand, which direction in our research we head, we have at least introduced the entropy rate definition of a transport flow.

This work has been supported by institutional grant FRI 8/2007

Lektoroval: prof. Ing. Karel Šotek, CSc.

Předloženo: 8.3.2007

Literature

1. REZA, F. M. *An Introduction to Information Theory*. General Publishing Company, Canada, ISBN 0-486-68210-2, (1994).
2. ČERNÝ, J., KLUVÁNEK, P. *Základy matematickej teórie dopravy*. Veda, Bratislava, ISBN 80-224-0099-8, (1991).
3. ANATHARAM, V., VERDÚ, S. *Bits Through Queues*. IEEE Trans. on Information Theory, Vol. 42, No.1, January (1996).

Resumé

MODELOVÁNÍ DOPRAVNÍHO PROCESU UNÁRNÍM KÓDEM

Martin KLIMO, Jana URAMOVÁ

Článek analyzuje analogii náhodných intervalů mezi dvěmi jednotkami v dopravním proudu a slovy unárního kódu. Tím poukazuje na otázky, které jsou společné pro teorii dopravy a teorii informací, což umožňuje požít výsledky teorii informací v dopravě. Nejsou to jenom otázky kvantifikace neuspořádanosti (entropie) dopravního toku, ale rovněž precizování pojmu kapacity dopravního systému. Podrobněji jsou prezentovány vlastnosti náhodných intervalů s maximální entropií při minimální střední délce.

Summary

MODELLING OF TRANSPORT PROCESS BY UNARY CODE

Martin KLIMO, Jana URAMOVÁ

Analogy of random intervals between two transport units and words of unary binary code is analyzed. This way, article is pointing to questions, which are common for transport theory and theory of information, which allows usage of theory of information results in tranport. These are not only questions of quantification of assortness (entrophy) of transport stream, but also better specification of term capacity of transport system. Properties of random intervals with maximum entropy under minimum length are presented.

Zusammenfassung

DIE SIMULATION DES VERKEHRLICHEN PROZESSES MIT DER HILFE DES UNÄREN KODES

Martin KLIMO, Jana URAMOVÁ

Die Analogie wurde der ZufallsIntervalls zwischen zwei Transporteinheiten und Wortes des monadischen Binärcodes analysiert. Damit weist sie auf die Fragen, die fur die Verkehrstheorie und Informationstheorie gemeinsam sind, was die Resultate aus der Informationstheorie in dem Verkehr benutzen ermoglichen wurden. Es sind nicht nur die Quantifizierungfragen der Unordnung der Transporteinheit, aber auch die Fragen der Prezision des Begriffs uber die Kapazitat des Verkehrssystems. Die Eigenschaften der Zufallsintervalls mit der Maximalentrophie bei der minimalen Mittellange wurden umfassend prasentieret.