KYBERNETIKA ČÍSLO 6, ROČNÍK 6/1970

Quasi-Questionnaires, Codes and Huffman's Length

C. F. PICARD

New concepts are defined, in particular the quasi-question or vertex with an outgoing arc of zero probability. A quasi-questionnaire is a probabilistic homogeneous (rooted) tree with quasi-questions.

It is shown that every instantaneous code is a quasi-questionnaire with precise restrictive conditions; it may also be a questionnaire, without an arc of zero probability.

Also, an approximation is given — without use of the classical construction — of the average length of Huffman's code with a given alphabet and given probabilities of code-words.

INTRODUCTION

A questionnaire is a graph with the set X of vertices having the partition $Q \cup E$ such that:

– Q, the set of questions, is formed by the vertices which are origins of at least two arcs; there is one and only one vertex $x_0 \in Q$ which is terminal extremity of no arc: it is the root.
– E is formed by the set of terminal vertices, called events.
– If there is one and only one path from x_0 to all the other vertices, then the graph is an arborescence (or rooted tree) and there enters exactly one arc in every vertex, but x_0: it is always the case in this paper.
– There exists a mapping $P : i \to p(i)$ from X on the interval $[0, 1]$ such that

$$\sum_{e \in E} p(e) = 1 , \quad p(i) = \sum_{j \in \Gamma_i} p(j) \quad \text{for all} \quad i \in Q ,$$

where Γ_i is the set of successors of i; then $p(x_0) = 1$.
– $|\Gamma_i| > 1$ for all $i \in Q$.

The outward degree of every question is called the basis of the question and it is written a_i (or a if possible). If all the questions have same basis, a, the questionnaire is called homogeneous. There is a compatibility relation between the number of events, $|E| = N$, the number M of questions and the basis: $M = (N - 1)/(a - 1)$. If this relation is true, then the questionnaire is strictly homogeneous; else there is a question with $\beta + 1$ for basis, where β is the rest of the integer division $(N - 1) \div \div (a - 1)$; β is then strictly less than $a - 1$; all the other questions have same basis a and the questionnaire is called homogeneous in the wide sense.

If $a = 2$, the questionnaire is always strictly homogeneous and is called a dichotomic one, if $a > 2$, then it is a polychotomic one.

An heterogeneous (rooted-tree) questionnaire is defined in the case where the bases are not the same for all the questions: $\mathscr{A} = \{a_1, \ldots, a_M\}$ is the set of the bases and the compatibility relation is now

$$\sum_{i=1}^{M} (a_i - 1) = N - 1 .$$

It is possible to associate a code to a questionnaire by a mapping from the set of events on the set of codewords. At every answer $\Gamma^0 i, \Gamma^1 i, \ldots, \Gamma^{a-1} i$ to the question i (without an explanation of the kind of answer), we associate a a-ary digit $j \in \{0, 1, \ldots \ldots, a - 1\}$. The path from x_0 to x must be coded with, say, l a-ary digits if this path contains l arcs: at every arc, corresponds one digit, the left one for the answer to x_0, the right one for x. In doing this, we get a word with l digits and this is the codeword of the code associated to the questionnaire. Because the questionnaire is a rooted-tree, it is possible to code all the events with other words and the code is decipherable and instantaneous. Then Huffman optimal coding procedure is able to give the algorithm for building an optimal homogeneous questionnaire. The case of heterogeneous questionnaire needs to use sometimes an alphabet with a_i letters (0 to a_{i-1}) for the question of basis a_i and we gave already a generalization of Huffman's algorithm to the heterogeneous case. From the coding point of view, it is for example the use of letters and digits to code an event.

In fact the root operates a partition of E in a_{x_0} subsets and every other question i operates a partition of a subset of E in a_i subsets. If all the successors of a question i are events $e_1, \ldots, e_a$, then i operates the final partition of a subset of size a.

The theory gives rules for building algorithms for feazible questionnaires, when some restrictions are done over the partitions of E and of its subsets; for example the unique type of question is to operate an ordinary comparison between two numbers with two (or three) outcomes as $a > b$, $a \leqq b$ (or $a > b$, $a = b$, $a < b$).

Some operations are defined over the questionnaires: they allows to build sophisticated graphs and questionnaires with very ordinary ones; they give too questionnaires with one or other extremum property.

The theory of questionnaires leads to use the information theory in view of the evaluation of the information value of a questionnaire. In this paper, we use essentially the concept of *routing length* i.e. the expectation of the length of a path in a questionnaire and the concept of information transmitted by a questionnaire when the set E is given, with its distribution of probabilities; it has been shown that the information transmitted by a questionnaire is always less than the routing length.

1. QUASI-QUESTIONNAIRES

Definition 1. *A quasi-event* is a terminal vertex of a probabilistic tree with a zero probability.

Definition 2. *A quasi-question* is an inner vertex of a homogeneous tree with at least one terminal vertex of nonzero probability as descendant and one quasi-event as successor.

Definition 3. *The probability of a quasi-question* x is the sum of the probabilities of the terminal vertices of nonzero probabilities which belong to the descendance of x. If a quasi-question x has only one successor y of nonzero probability, all the other successors of x are quasi-events and the information given by x is zero because the conditional probability of y is equal to 1.

A quasi-question is not a question in the usual sense of questionnaires.

To repeat, an event is a terminal vertex of nonzero probability and a *question* is a vertex with *a successors* which are events or questions.

Definition 4. *A polychotomic quasi-questionnaire* is a probabilistic homogeneous tree with the two following properties:

1. the vertices are events, quasi-events, questions or quasi-questions;
2. the sum of the probabilities of the events is 1.

It will be noted that in a quasi-questionnaire* there are two kinds of terminal vertices:

– terminal vertices giving a complete system of events that is the set $E = \{e_i\}$ such as $\sum_{e_i \in E} p(e_i) = 1$, where $P = \{p(e_i) \mid e_i \in E\}$ is given.

– terminal vertices such as $\forall_j \; p(\bar{e}_j) = 0$ giving the set $\bar{E} = \{\bar{e}_j\}$.

This case is very different from the incomplete systems studied by, for example, Renyi [8] and those used in the sub-questionnaires (Dubail ([3]).

In a probabilistic tree, a non-terminal vertex with only quasi-events as successors will be "reduced" in a quasi-event with reduction of the order of the tree.

In a wider sense, a homogeneous questionnaire, such as the number of events N is $N = a^k + \alpha(a - 1) + \beta$, may be considered as a quasi-questionnaire with $(N - \beta - 1)/(a - 1)$ questions and a quasi-question with $(a - \beta - 1)$ quasi-events as outcomes.

* See figure in the annex.

The average path-length of a quasi-questionnaire $\bar{K}$ is defined by the usual sum:

$$\bar{L} = \sum_{e_i \in E} p(e_i)\, r(e_i) \tag{1}$$

where the rank of the event e_i is $r(e_i)$. That is, the event e_i is connected with the roo by a path whose length is $r(\bar{e}_i)$. Of course, the possible extension of the sum to the quasi-event $\bar{e}_j$ does not change $\bar{L}$. $\bar{L}$ is called *routing-length* (or rtng-length)

Theoreme 1. *The routing-length of a quasi-questionnaire is*

$$\bar{L} = \sum_{q \in Q \cup \bar{Q}} p(q) \tag{2}$$

where q is a question $(q \in Q)$ *or a quasi-question* $(q \in \bar{Q})$.

The equivalence between (1) and (2) is proved in the same way as for rooted-tree questionnaires (Picard [7]).

To a polychotomic questionnaire K constructed on (a, P) we can map an infinity of quasi-questions obtained by substituting $\bar{P}$ to P, where $\bar{P}$ is a distribution of the same N nonzero probabilities and of a multiple of $(a - 1)$ zero probabilities.

But reciprocally to a quasi-questionnaire $\bar{K}$ is associated only a couple (a, P) where P is a distribution of N nonzero probabilities.

Let $\bar{N}$ be the number of quasi-events of $\bar{K}$ and let $\bar{M} > 1$ be the number of quasi-questions.

The attempt is made to do some transfers of terminal vertices by interchanging the probabilities of two vertices, an event e_i of rank – the rank of e_i is the path-length from the root to e_i – $r(e_i)$ and a quasi-event of rank r_h.

If $r_h < r(e_i)$ then the rtng-length of the new tree is:

$$\bar{L} - p(e_i)\left[r(e_i) - r_h\right]. \tag{3}$$

This type of transfer must be repeated as often as possible.

If during this process a tree is such that there is a vertex y with only quas-events as descendants, then only the vertex y will be substituted for the subtree issued from y, and y will be taken as quasi-event.

At the end $\bar{N}$ will be reduced by a multiple of $a - 1$ and $\bar{M}$ will also be reduced.

Let r_M be the highest rank of the events; then there is no quasi-event of rank $r_h > r_M$; if not, there would be a vertex of rank r_M of which no descendant would be an event.

If all the quasi-events are of the rank r_M, it is possible to do again a transfer between an event and a quasi-event.

Such transfers will not reduce the rtng-length but may permit to give a quasi-events as successors to the same vertex. Then the number of quasi-events and quasi-questions will be reduced.

If and only if $\bar{N} < a - 1$, no possible transfer can reduce $\bar{N}$.

All transfers performed one after the other make it possible to form a homogeneous questionnaire K' in the strict sense (if $N' = 0$) or in the wide sense (if $0 < N' < a - 1$); the events e_i of this questionnaire K' have as rank $r'(e_i)$ such that $r'(e_i) < r(e_i)\ \forall\ i$.

Consequently the rtng-length L' of the questionnaire K' is

$$L' \leqq \bar{L}. \tag{4}$$

Furthermore, if a quasi-event has a rank $r_h < r_M$ then the inequality yields:

$$L' < \bar{L}. \tag{5}$$

If all the quasi-events are of maximal rank then

1. if $\bar{K}$ has $\bar{N} < a - 1$ quasi-events then, at the most, $\bar{N} - 1$ transfers will permit to form a homogeneous questionnaire, in the wide sense, with a rtng-length

$$L' = \bar{L};$$

2. if $\bar{K}$ has $\bar{N} > a - 1$ quasi-events, then some transfers will permit to form at least a non-terminal vertex with no more than one outgoing event.

If there is no outgoing event this vertex will be replaced by a quasi-event of rank less than r_M in such a way that a new transfer will permit to reduce the rtng-length; if not, the following case will occur.

Every quasi-question of rank r with only one outgoing event e_i may be replaced by this event; consequently, there is a reduction of rtng-length of

$$(r(e_i) - r)\ p(e_i)$$

Thus, the following result has been obtained:

Theorem 2. *Every quasi-questionnaire $\bar{K}$ constructed on (a, P), with a routing-length $\bar{L}$, and with $\bar{N} > a - 1$ quasi-events may be reduced to a questionnaire K' constructed on (a, P) with a routing-length $L' < L$.*

We note that the restriction $\bar{N} > a - 1$ follows from the fact that if $N = a^k + \alpha(a - 1) + \beta$ then the same rooted tree K may be considered both as a questionnaire and as a quasi-questionnaire.

2. INSTANTANEOUS CODES AND QUASI-QUESTIONNAIRES

Some connexions have already been noted between coding and questionnaire (see, for example, Picard [7] § 1–5 and Césari [2]).

The couple (a, P) is the base of the questions and the distribution of the probabilities of the complete system of events E (in the sense of questionnaire) or the number of the letters of the alphabet and the probabilities of the code-words which still form a complete system of events E (in the sense of coding).

The words of an *instantaneous* code are those of the terminal vertices of a probabilistic rooted-tree which have nonzero probability.

The problem of instantaneous coding is to determine the ordered sequence of the arcs of the path connecting the root to the code-words.

This coding is also used in the Theory of Questionnaires to locate the events.

But in this last theory the questions or inner vertices are also to be noted; since the characters used to code a question, i.e. to locate its position on the rooted-tree, are the prefix of the characters of the terminal vertices, we must, to decode them, know the rank of the vertices ([7], § 1.2.1.).

An instantaneous code is a quasi-questionnaire because definition 4 holds.

Let A be a strict homogeneous rooted-tree. Every vertex of A has either no successor or a successors.

Let E be the set of the terminal vertices with N elements, $E = \{e_i \mid i = 1, 2, \ldots, N\}$ and let $r(e_i)$ be their ranks.

Picard [7] and Ash [1], for example, have proved that it is possible to construct an optimal questionnaire on A, without "*discrepancy*" K_0, using an information equal to the rtng-length (called also the absolutely optimal code) assigning the probability $p(e_i) = a^{-r(e_i)}$ to the vertices of rank $r(e_i)$.

The calculation step by step (from the events to the root) of the probabilities of the questions shows that all the vertices of K_0 have a probability connected to the rank by: $p(x) = a^{-r(x)}$. Indeed, a question x of rank r has for successors: the vertices (events or questions) of rank $r + 1$ and of probability $a^{-(r+1)}$ thus $p(x) = \sum a^{-(r+1)} = a^{-r}$. In particular the root x_0 is such that $p(x_0) = a^0 = 1$. From $p(x_0) = \sum_{e_i \in E} a^{-r(e_i)}$ it follows

$$\sum_{e_i \in E} a^{-r(e_i)} = 1 \tag{6}$$

The property (6) holds for every questionnaire constructed on A.

Let then E be the partition $E = E' \cup \bar{E}$ of the terminal vertices of A and let $q(e)$ be the probabilities of an instantaneous code $\bar{K}$, such that:

$$\sum_{e_i \in E'} q(e_i) = 1 \quad \text{and} \quad \begin{cases} (\forall \bar{e} \in \bar{E})\, q(\bar{e}) = 0\,, \\ (\forall e_i \in E')\, q(e_i) \neq 0 \end{cases}$$

the property (6) implies:

$$\sum_{e_i \in E'} a^{-r(e_i)} + \sum_{\bar{e} \in \bar{E}} a^{-r(\bar{e})} = 1 \tag{7}$$

thus,

$$\sum_{e_i \in E'} a^{-r(e_i)} < 1 \tag{8}$$

and nevertheless E' is a complete system of events.

$\bar{E}$ is the set of quasi-events.

The quasi-questionnaire is a questionnaire if an only if $\bar{E} = \emptyset$ so that the following theorem is proved:

Theorem 3. *An instantaneous code constructed on a complete system of evens $E = \{e_i\}$ is a strictly homogeneous questionnaire, if and only if:*

$$\sum_{e_i \in E} a^{-r(e_i)} = 1 .$$

The existence of an instantaneous code for which the rank of every codeword e_i is the integer immediately greater than the logarithm (of base a) of the inverse of the probability $p(e_i)$ has been indicated by Shannon ([9] § 9) and proved by Feinstein [4].

Noiseless Coding Theorem. *There exists an instantaneous code C such that*

$$\log \frac{1}{p(e_i)} \leqq r(e_i) < \log \frac{1}{p(e_i)} + 1 \tag{9}$$

for every code word e_i.

In such a code, the word e_i is coded by a word of $r(e_i)$ characters.

The multiplication of each member of (9) by $p(e_i)$, and then the summation over the elements e_i of the set E of code-words C, lead to the double inequality:

$$I_N(E) \leqq \bar{L}(C) < I_N(E) + 1 \tag{10}$$

where $I_N(E)$ is the Shannon's information of code C, expressed with logarithms of base a and $\bar{L}(C)$ the rtng-length of the quasi-questionnaire C. One could also write:

$$r(e_i) = \left[\log \frac{1}{p(e_i)}\right] + 1 \tag{11}$$

and

$$\bar{L}(C) = \sum_{e_i \in E} p(e_i) \left(\left[\log \frac{1}{p(e_i)} \right] + 1 \right) \tag{12}$$

where $[x]$ is the greatest integer *strictly* less than x.

Let then C be a Feinstein's code of which the ranks of the words are determined by (9).

If $\log 1/p(e_i)$ is integer for every e_i then $r(e_i) = \log 1/p(e_i)$ $(\forall i)$ and therefore $\sum_{e_i \in E} a^{-r(e_i)} = 1$, thus C is a questionnaire.

If $\log 1/p(e_i)$ is not integer for at least one word e_i, then, according to (9) $r(e_i) >$ $> \log 1/p(e_i)$, that is

$$p(e_i) > a^{-r(e_i)} . \tag{13}$$

Since for every word of C, $p(e_i) \geqq a^{-r(e_i)}$, if (13) holds for at least one word, then

$$\sum_{e_i \in E} a^{-r(e_i)} < \sum_{e_i \in E} p(e_i) \quad \text{and, thus} \quad \sum_{e_i \in E} a^{-r(e_i)} < 1 \; ;$$

and so, according to Theorem 3, C is a quasi questionnaire.

Theorem 4. *A code C satisfying the inequalities* (9) *of the Noiseless Coding Theorem is a strictly homogeneous questionnaire if and only if*

$$r(e_i) = \log \frac{1}{p(e_i)} \quad (\forall i) \, .$$

The homogeneous questionnaires in the wide sense have in fact a property similar to the homogeneous questionnaires in the strict sense.

The sub-tree with one question of base $\beta + 1 < a$ may admit some events of the same probability of the form $a^{-r}/(\beta + 1)$.

Then if each of the $N - (\beta + 1)$ other events e_i forming the subset $E_i \subset E$ has a probability of the form $a^{-r(e_i)}$, then the quasi-questionnaire is a homogeneous questionnaire in the wide sense:

Corollary. *A code C satisfying the inequalities* (9) *of the Noiseless Coding Theorem is a homogeneous questionnaire in the wide sense, if and only if, the following conditions hold*:

1. $N - (\beta + 1)$ *is a multiple of* $a - 1$ *and* $\beta + 1 < a$,
2. $N - (\beta + 1)$ *events have a probability of the form* $a^{-r(e_i)}$,
3. $(\beta + 1)$ *events have a probability of the form* $a^{-s}/(\beta + 1)$,
4. $s \geqq \sup_{e_i \in E_1} r(e_i)$ *where* E_1 *is the set of* $N - (\beta + 1)$ *events of type* 2.

Otherwise C is a quasi-questionnaire.

The last condition places the usual restriction on the optimal homogeneous questionnaires in the wide sense: the question of base $\beta + 1$ has as outcome the events of smallest probabilities.

3. OPTIMAL QUESTIONNAIRES AND QUASI QUESTIONNAIRES

The questionnaires of the minimal rtng-length K_H or optimal questionnaires, may be constructed — within one equivalence — by Huffman's algorithm (1952).

This algorithm allows the evaluation of the rank of the events, thus the determination of the rtng-length L_H, but, to our knowledge, there does not exist a formula which gives L_H directly without using this algorithm.

Now let us study the rank of the vertices of the optimal questionnaires.

Let us suppose that the ranks of the events of a questionnaire K_H are at least equal to k_0. For every rank $r \leqq k_0$, there are exactly a^r vertices and the sum of the probabilities of the vertices of every rank $(0 \leqq r \leqq k_0)$ is 1.

Let X_r be the set of the vertices of rank r

1. $r \leqq k_0$ then it follows:

(14) $$|X_r| = a^r$$

and

(15) $$\sum_{x \in X_r} p(x) = 1$$

Furthermore, if K_H is an optimal questionnaire:

(16) $$x \in X_r \quad \text{and} \quad y \in X_{r+1} \Rightarrow p(x) \geqq p(y)$$

and for the predecessor x_0 of y, $(\forall y \in \Gamma x_0)$:

(17) $$p(x_0) > p(y) .$$

However, these properties are not sufficient to characterize an optimal questionnaire.

If, in K_H, it was possible to find a vertex y of rank $r \leqq k_0 + 1$ such that

$$p(y) \geqq \frac{1}{a^{r-1}},$$

then for all the rank $r - 1$ we would have, according to (14), (16) and (17):

$$\sum_{x \in X_{r-1}} p(x) > a^{r-1} \quad p(y) \geqq 1 ,$$

which is in contradiction with (15).

Thus:

(18) $$r(x) \leqq k_0 + 1 \Rightarrow p(x) < \frac{1}{a^{r(x)-1}}$$

so

$$r(x) < \log \frac{1}{p(x)} + 1 ,$$

which is the second inequality of (9), extended to the questions of K_H.

2. $r > k_0 + 1$ with K_H being always an optimal questionnaire.

If for every vertex of rank $r - 1$ we have $p(x_{r-1}) \leqq a^{-l+1}$, is-it possible to have a vertex of rank $r + 1$ such that $p(x_{r+1}) \geqq a^{-l}$?

In this case, all the vertices of rank r would have the probability $p(x_r) \geqq a^{-l}$, and for at least a question of rank $r : p(q_r) > a^{-l}$; and the questions of rank $r - 1$ would have the probability $p(q_{r-1}) \geqq a^{-l+1}$ and for at least a question of rank $r - 1$ $p(q_{r-1}) > a^{-l+1}$; which leads to a contradiction. Therefore:

Theorem 5. *In an optimal questionnaire in which the events are of rank at least equal to k_0, the probability of every vertex x of rank $r(x) \leqq k_0 + 1$ is such that $p(x) < 1/a^{r(x)-1}$; if for $r > k_0 + 1$ every vertex of rank $r - 1$ is such that $p(x_{r-1}) \leqq \leqq a^{-l+1}$ then all vertices of rank $r + 1$ have a probability $p(x_{r+1}) < a^{-l}$.*

A necassary condition of optimization such that: "No vertex of rank r has a probability greater than the sum of the probabilities of a other vertices of the same rank" [7] leads naturally to theorem 5, but not to stronger inequalities when $r > > k_0 + 1$.

Theorem 5 leads to the comparison of the ranks of the events in Huffman's and Feinstein's codes. Let $r_c(e)$ and $r_H(e)$ be the respective ranks of the same event with the same probability $p(e)$ and let $\inf_{e \in E} r_H(e) = k_0$.

Then:

$$\begin{aligned}
r_H(e) &= k_0 &&\Rightarrow p(e) < a^{-k_0+1} \quad \text{and} \quad r_c(e) \geqq k_0\,, \\
r_H(e) &= k_0 + 1 &&\Rightarrow p(e) < a^{-k_0} \quad \text{and} \quad r_c(e) \geqq k_0 + 1\,, \\
r_H(e) &= k_0 + 2 &&\Rightarrow p(e) < a^{-k_0} \quad \text{and} \quad r_c(e) \geqq k_0 + 1\,, \\
r_H(e) &= k_0 + 3 &&\Rightarrow p(e) < a^{-k_0-1} \quad \text{and} \quad r_c(e) \geqq k_0 + 2\,,
\end{aligned}$$

and also

$$\begin{aligned}
r_H(e) &= k_0 + 2h &&\Rightarrow p(e) < a^{-k_0-h+1} \quad \text{and} \quad r_c(e) \geqq k_0 + h\,, \\
r_H(e) &= k_0 + 2h + 1 &&\Rightarrow p(e) < a^{-k_0-h} \quad \text{and} \quad r_c(e) \geqq k_0 + h + 1\,,
\end{aligned}$$

that is,

$$\begin{aligned}
r_H(e) &= k_0 + 2h &&\Rightarrow r_c(e) \geqq r_H(e) - h\,, \\
r_H(e) &= k_0 + 2h + 1 &&\Rightarrow r_c(e) \geqq r_H(e) - h\,.
\end{aligned}$$

Let then e_1 be the event with the greatest probability

$$p(e_1) = \sup_{e_i \in E} p(e_i)\,;$$

The rank of e_1 is k_0, that is, $r_H(e_1) = k_0$ and furthermore

$$r_c(e_1) = \left[\log \frac{1}{p(e_1)}\right] + 1\,.$$

 The preceding inequalities lead to

$$k_0 \leqq \left[\log \frac{1}{p(e_1)}\right] + 1 .$$

Furthermore, let us suppose:

$$k_0 \leqq \left[\log \frac{1}{p(e_1)}\right] - 1 ,$$

that is:

$$p(e_1) \leqq a^{-k_0 - 1} ;$$

since the questionnaire is optimal, the vertices of rank $k_0 + 1$ have a probability:

$$p(x_{k_0+1}) \leqq p(e_1) \leqq a^{-k_0-1}$$

and the questions of rank k_0:

$$p(q_{k_0}) \leqq a^{-k_0}$$

so that

$$\sum_{x \in X_{k_0}} p(x) \leqq (a^{k_0} - 1)\, a^{-k_0} + a^{-k_0 - 1} < 1$$

which is absurd.

Thus

$$k_0 = \left[\log \frac{1}{p(e_1)}\right] + \alpha$$

where $\alpha = 0$ or 1 and k_0 has no other possible value.

Hence the results:

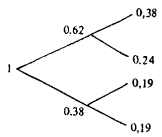
Theorem 6. *In an optimal questionnaire K_H, the event of greatest probability e_1 has a rank $r_H(e_1) = k_0$ such that:*

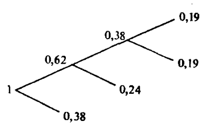
$$k_0 = \left[\log \frac{1}{p(e_1)}\right] \quad \textit{or} \quad k_0 = \left[\log \frac{1}{p(e_1)}\right] + 1 .$$

For $h \geqq 0$, the ranks of the events $r_H(e)$ are connected with the ranks $r_c(e)$ of Feinstein's code by:

$$(19) \qquad r_H(e) = k_0 + 2h \quad \textit{or} \quad r_H(e) = k_0 + 2h + 1 \Rightarrow r_c(e) \geqq r_H(e) - h .$$

Aplications. 1. Let us consider the two following Huffman's questionnaires



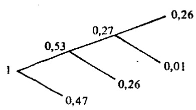


for which $L_H = 2$.

The first one is such that $k_0 = \log_2 [(100/38)] + 1 = 2$.

The second one is such that $k_0 = \log_2 [(100/38)] = 1$.

2. Let us consider this example given by Y. Césari to whom I am grateful:



For $p(e_1) = 0{,}47$ we have $r_H(e_1) = k_0 + 2h_1$,

$p(e_2) = 0{,}26$ $\quad r_H(e_2) = k_0 + 2h_2 + 1$,

$p(e_3) = 0{,}26$ $\quad r_H(e_3) = k_0 + 2h_3$,

$p(e_4) = 0{,}01$ $\quad r_H(e_4) = k_0 + 2h_4$,

with

$$h_1 = h_2 = 0 \quad \text{and} \quad h_3 = h_4 = 1$$

we have:

$$r_c(e_1) > r_H(e_1) - h_1$$
$$r_c(e_2) = r_H(e_2) - h_2$$
$$r_c(e_3) = r_H(e_3) - h_3$$
$$r_c(e_4) > r_H(e_4) - h_4\,.$$

The calculation of a bound of the routing-length of the optimal questionnaire is a consequence of:

$$L_H \leqq L' \leqq \bar{L}(C)$$

where L_H is the rtng-length of an optimal questionnaire K_H, $\bar{L}(C)$ the rtng-length of Feinstein's code, L' – the rtng-length of a questionnaire K' deduced from Feinstein's code, these 3 quasi-questionnaires being constructed on the same couple (a, P).

It is possible to bound the preceding inequalities by some informations:

$$I_N(E) \leqq L_H \leqq L' \leqq \bar{L}(C) \leqq I_N(E) + 1\,. \tag{20}$$

 From (12) and (20) we have

$$(21) \qquad L_H < \sum_{e_i \in E} p(e_i)\left(\left[\log \frac{1}{p(e_i)}\right] + 1\right)$$

if at least the probability of an event is not a power of $1/a$ (from theorems 2 and 4).

Let us suppose now that this condition holds; the code C is then a quasi-questionnaire and from theorem 2 there exists a questionnaire K' such that $L' < \bar{L}(C)$.

The evaluation of (11), according to (21) requires the determination for every event e_i, of

$$c_i = \left[\log \frac{1}{p(e_i)}\right] + 1\,.$$

Let us then call Sup $c_i = c_N$ and

$$(22) \qquad \sum_{e_i \in E} a^{-c_i} = 1 - \varrho$$

where

$$\varrho = \sum_{\bar{e}_j \in E} a^{-r(\bar{e}_j)}\,,$$

ϱ is the residue corresponding to the quasi-event of C. ϱ is then sum of $\bar{N} = |\bar{E}|$ powers of $1/a$ (distinct or not) corresponding to the quasi-questionnaire C.

These powers may be obtained explicitly from (22): we shall use a system of numeration of base a to express $\sum_{e_i \in E} a^{-c_i}$ and ϱ may be written:

$$\varrho = \sum_{r_i=1}^{N} t_i a^{-r_i} \quad \text{with} \quad 0 \leqq t_i < a\,.$$

$\bar{N}$ will be then the sum of the nonzero digits:

$$\bar{N} = \sum_{i=1}^{c_N} t_i\,.$$

Example. If $P = \{0{,}90;\, 0{,}10\}$ then $c_1 = 1$ and $c_2 = 4$; $\sum_{e_i \in E} 2^{-c_i} \Rightarrow 0{,}1001$ so that $\rho = 0{,}0111$: thus $\bar{N} = 3$ and the ranks of the quasi-events are 2, 3 and 4.

Also for $a = 3$

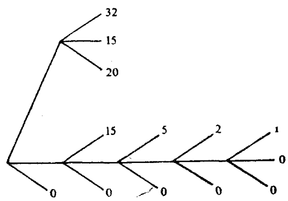
$$P = \left\{\frac{32}{100}, \frac{25}{100}, \frac{20}{100}, \frac{15}{100}, \frac{5}{100}, \frac{2}{100}, \frac{1}{100}\right\}$$

thus $c_i = 2, 2, 2, 2, 3, 4, 5$, and

$$\sum_{i=1}^{N} 3^{-c_i} = 0{,}11111 = 1 - \varrho$$

since $4 \times 3^{-2} = 3^{-1} + 3^{-2}$ and $p = 0{,}11112$ thus

$$\bar{N} = \sum_{i=1}^{c_N} t_i = 6$$



Let us call $\bar{r}_j$ the ranks of the quasi-events, for $j = 1, \ldots, \bar{N}$. A quasi-questionnaire K_1 is deduced from C by a succession of $\bar{N}$ permutations (at the most) among the events and the quasi-events; each one may be followed by a substitution in case there would be a non-terminal vertex of zero probability.

The maximal rank of an event of C is:

$$c_N = \bar{r}_N = 1 + \underset{e_i \in E}{\mathrm{Max}} \left[\log \frac{1}{p(e_i)} \right] \tag{23}$$

and corresponds to the event e_N with the smallest probability p_N.

If e_N is the only event of rank c_N, before a permutation we can replace the quasi-question with outcome e_N by an event of probability p_N; the rank of e_N is $c_N - 1$ and the reduction of the rtng-length is p_N (recursive operation).

If the difference of rank between e_N and the event of probability immediately higher, e_{N-1} is $c_N - c'_N$, we shall have reduced the rtng-length by $p_N(c_N - c'_N)$ and the number of quasi-events by $(a - 1)(c_N - c'_N)$.

Let us suppose then that $c_N = c'_N$ and let c_N be the rank of e_N in C.

If there exists a quasi-event of rank $\bar{c}_j < c_N$, then the permutation of e_{N-1} with this quasi event gives a reduction of $(c_N - \bar{c}_j)\, p(e_{N-1})$.

But we can increase this reduction by doing a permutation between the quasi-event of minimal rank $\bar{c}_0$ and the event of maximum probability among the events of rank higher than $\bar{c}_0$; then, by other steps (j), by substituting a new value to $\bar{c}_0$ (greater or equal to the previous value) we can repeat the process.

This algorithm may be expressed as follows:

1. Let $\bar{c}_0(j)$ be the smallest rank of a quasi-event at the step j (j starting at 1); the permutation of a quasi-event of rank $\bar{c}_0(j)$ with the event of greatest probability, having a rank c_i greater than $\bar{c}_0(j)$, leads to the reduction of

$$\underset{c_i > \bar{c}_0(j)}{\mathrm{Max}} \{p(e_i)\} \times (c_i - \bar{c}_0(j)) \,.$$

2. For the next step $(j + 1)$ we shall substitute $\bar{c}_0(j + 1)$ to $\bar{c}_0(j)$ where

$$\bar{c}_0(j + 1) \geqq \bar{c}_0(j)$$

the algorithm will stop at step t such that for every event $c \leqq \bar{c}_0(t)$.

At the end, one will obtain a quasi-questionnaire K_1 with $\bar{N}_1$ quasi-events and a routing-length:

$$L_1 = \bar{L}(C) - \sum_{j=1}^{t} (c_i - \bar{c}_0(j)) \operatorname*{Max}_{c_i > \bar{c}_0(j)} \{p(e_i)\} \tag{24}$$

C et K_1 have the *property P*:

The events ordered according to the decreasing probabilities have non-decreasing ranks.

This is a necessary property for optimal questionnaires. This property is true for K'. K' will be deduced from K_1 according to theorem 2.

Furthermore, the rtng-length of K' is less than that of K: every time $(a - 1)$ quasi-events have been suppressed, then the event of probability $p_i > p_N$ may receive a rank lower.

Thus

$$L' < L_1 - (\bar{N}_1 \,\%\, (a - 1))\, p_N\,. \tag{25}$$

($a \,\%\, b$ means the integer part of the quotient of a by b.)

This formula may be expressed more accurately

(1) by writing $\bar{N}_1$, in a polynomial form

$$\bar{N}_1 = \sum \alpha_v (a - 1)^v$$

and using the successive remainders;

(2) by substituting to p_N the $(\bar{N}_1 \,\%\, (a - 1))$ smallest probabilities among the $p(e_i)$.

From (24) and (25):

Theorem 7. *It is possible to find an upper limit to the routing-length of a polychotomic questionnaire by*

$$L_H \leqq \bar{L}(C) - \sum_{j=1}^{t} (c_i - \bar{c}_0(j)) \operatorname*{Max}_{c_i > \bar{c}_0(j)} \{p(e_i)\} - p_N(N_1 \,\%\, (a - 1)) \tag{26}$$

where $\bar{L}(C)$ *is defined by* (12), $\bar{N}_1$ *is the number of quasi-events of the quasi-questionnaire* K_1 *of which all the quasi-events are of maximal rank,* p_N *is the smallest probability of P.*

For the ranks c_i, $\bar{c}_0(j)$ and t see formula (24). This formula is not trivial.

Nevertheless, it gives the determination of the upper limit of L_H without using Huffman's construction and has the theoretical interest to use the events by their probabilities $p(e_i)$ without using the questions.

By substituting a questionnaire K'' to the code C without using the questionnaire K_1, we found a more elementary formula:

Since C has $\bar{N}$ quasi-events, we can deduce from the polynom $\bar{N} = \sum \beta_w (a-1)^w$ how many times the rank of an event may reduced by one so that:

$$L_H \leqq L'' < \bar{L}(C) - (\bar{N} \,\%\, (a-1))\, p_N \,. \tag{27}$$

L'' may be stated more accurately as in (25), but we are not sure that the questionnaire K'' with the routing-length L'' has the property P. (26) is more exact than (27).

According to a communication by E. Gilbert (Bell C°), it seems that E. Moore has shown a simpler formula but less precise than (27) (see also Jelinek's Annex [6])

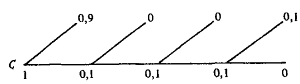
$$L_H \leqq I_N(E) + 1 - 2\, p_N \quad \text{for} \quad a = 2\,.$$

ANNEX. EXAMPLES OF QUASI-QUESTIONNAIRES AND CODES

1. Let $a = 2$, $P = \{90/100, 10/100\}$.
For $c_1 = 1$, $c_2 = 4$ we have found $\bar{c}_0(1) = 2$. (24) leads to

$$L_1 = \bar{L}(C) - 2 \times \frac{10}{100}$$

then $L' = L_H = 1$ and $\bar{L}(C) = 1{,}3$.



2. Let $a = 3$, $P = \{32/100, 25/100, 20/100, 15/100, 5/100, 2/100, 1/100\}$

$$j = 1 \quad \bar{c}_0(1) = 1 \quad c_1 = 2 \quad p(e_1) = \frac{32}{100}\,,$$

$$j = 2 \quad \bar{c}_0(2) = 2 \quad c_2 = 3 \quad p(e_2) = \frac{5}{100}\,,$$

$$j = 3 \quad \bar{c}_0(3) = 2 \quad c_3 = 4 \quad p(e_3) = \frac{2}{100}\,.$$

$$j\,\mathrm{Max} = 4 \quad \bar{c}_0(4) = 3 \quad c_4 = 5 \quad p(e_4) = \frac{1}{100}\,,$$

$$\bar{N}_1 = 2 = a - 1 \Rightarrow L' = L(C)$$

$$L' \leqq \overline{L}(C) - \frac{32 + 5 + 2 \times 2 + 1 \times 2}{100} - \frac{1}{100} = 1{,}68$$

Thus $L_H = 1{,}51$ and $\overline{L}(C) = 2{,}12$.

3. Let $a = 3$, $P = \{22/45, 1/5, 1/15 \text{ (three times)}, 1/54 \text{ (six times)}\}$; the ranks are determined by

$$3^0 > \frac{22}{45} \geqq 3^{-1} > \frac{1}{5} \geqq 3^{-2} > \frac{1}{15} \geqq 3^{-3} > \frac{1}{54} \geqq 3^{-4}$$

Thus $c_i := 1, 2, 3, 4$, we have

$$3^{-1} + 3^{-2} + 3 \,.\, 3^{-3} + 6 \,.\, 3^{-4} = 3^{-1} + 2 \,.\, 3^{-2} + 2 \,.\, 3^{-3}$$

Thus $\varrho = 3^{-1} + 3^{-3}$ and $\overline{N} = 2$

$$L_H \leqq \overline{L}(C) - \left[\frac{1}{5} \times 1 + \frac{1}{15} \times 1 + \frac{1}{54} \times 2\right] - \frac{1}{54},$$

$$L_H \leqq \overline{L}(C) - \frac{29}{90}.$$

Thus:

$$\overline{L}(C) = 1 + \frac{84}{90} \quad \text{and} \quad L_H = 1 + \frac{54}{90}$$

so that the bound is close to L_H.

4. Let C be the code defined by: $a = 2$ and $P = \{0{,}425/0{,}250/0{,}08125 \text{ (four times)}\}$; the ranks of C are:

$$r(e_1) = r(e_2) = 2$$

$$r(e_3) = \ldots = r(e_6) = 4$$

Thus

$$\sum_{e_i} 2^{-r(e_i)} = 0{,}11$$

and

$$\rho = 0{,}01\,.$$

There is only a quasi-event of rank 2. The difference $\overline{L}(C) - L'$ is obtained in the following way:

e_3 will change from rank 4 to rank 2 by permutation, then e_4 will change from rank 4 to rank 3 since $\overline{N}_1 = 1$.

Thus $\overline{L}(C) - L' = 3 \times 0{,}08125 = 0{,}24375$.

(Received April 6, 1970.)

[1] R. Ash: Information Theory. Interscience Pub., New York 1965.
[2] Y. Cesari: Questionnaire, Codage et Tris. Publication de l'Institut Blaise Pascal, Paris 1968.
[3] F. Dubail: Algorithmes de questionnaires optimaux au sens de divers critères. Thèse 3ème Cycle (Lyon) 1967.
[4] A. Feinstein: Foundations of information theory. Mc Graw Hill, New York 1958.
[5] D. A. Huffman: A method for the construction of minimum-redundancy codes. Proc. IRE (1952), 9, 1098.
[6] F. Jelinek: Probabilistic Information Theory discrete and memoryless models. Mc Graw Hill, New York 1968.
[7] C. F. Picard: Théorie des Questionnaires. Gauthier-Villars, Paris 1965.
[8] A. Renyi: Wahrscheinlichkeitsrechnung. VEB Deutscher Verlag der Wissenschaften, Berlin 1962.
[9] C. E. Shannon: The mathematical theory of communications. Bell System — Technical Journal (1948).

VÝTAH

Kvazi-dotazníky, kódy a Huffmanova délka

C. F. PICARD

Definují se nové koncepce, a to zejména kvazi-otázka neboli uzel s vycházející větví o nulové pravděpodobnosti. Kvazi-dotazník je pravděpodobnostní homogenní strom s kvazi-otázkami.

Ukazuje se, že každý okamžitý kód je kvazi-dotazník s přesně vymezenými podmínkami; může být též dotazníkem bez větve o nulové pravděpodobnosti.

Též je dána aproximace — aniž by se použilo klasické konstrukce — průměrné délky Huffmanova kódu s danou abecedou a s danými pravděpodobnostmi kódových slov.

C. F. Picard, Équipe de recherche structures de l'information, Tour 42, Faculté des Sciences, 9 Quai St. Bernard, Paris V^e. France.