INFORMATION AND CONTROL 32, 368–378 (1976)

Useful Information and Questionnaires

BERNADETTE BOUCHON

Centre National de la Recherche Scientifique, Paris, France

A questionnaire is an inquiry process, using a probabilistic latticoïd. We suppose that a positive valuation, called an utility, characterizes every terminal vertex. The useful information and the useful routing length of a questionnaire have been introduced in a particular case. We propose, in this article, to define and to study these quantities in the general case, and to exhibit some properties relative to a product of questionnaires, corresponding to dependent or independent processes.

1. INTRODUCTION

During the last ten years, the theory of questionnaires has been used to study different problems in various domains, for example to form a diagnostic as a result of different medical tests (Terrenoire, 1970), to organize flowcharts for computer programs (Picard, 1972, Chap. 13.5.3), to describe various sorting methods and to exhibit the best one, according to given criteria (Picard, 1972, Chap. 14.3; Bouchon, 1972), to look for some optimal codes (Picard, 1970).

A questionnaire is an inquiry process, which allows us to disjoint the n elements of a complete set E of events, according to some restrictive conditions (Picard, 1972; Bouchon, 1974). More precisely, a questionnaire Q is a directed valued graph without cycles, which has an unique starting vertex, called the root α of Q. Its set of vertices is denoted by X, and the binary relation on X determining this graph by Γ; every terminal vertex is called an answer, corresponding to one element of E; then, $F = X - E$ is the set of nonterminal vertices, called questions. The valuation p defined on every arc of Q is a probability function. We can characterize Q by the triplet (X, Γ, p). The probability of a vertex x, also denoted by p, is defined as the sum of the valuations of the arcs (y, x) ending in x, which is assumed to be equal to the sum of the valuations of the arcs (x, z) issued from x. The questionnaire is said to be arborescent, if there exists only one path between the root and every answer, latticoïd in the other case. If Q has only one question, it is said to be elementary.

Copyright © 1976 by Academic Press, Inc.
All rights of reproduction in any form reserved.

The product $\hat{Q}$ of two questionnaires Q_1 and Q_2 corresponds to the process obtained by doing the process Q_2 after having finished Q_1; if Q_2 occurs only after an answer x of Q_1, we say that $\hat{Q} = Q_1 \diamondsuit Q_2$ is the restricted product of Q_1 by Q_2 in x; if Q_2 is realized after every answer of Q_1, we say that $\hat{Q} = Q_1 Q_2$ is the direct product of Q_1 by Q_2.

We suppose in this article that every element of E is characterized by a positive valuation, called its utility. In the examples considered in the first paragraph, this valuation allows us to give particular importance to every diagnostic obtained after a sequence of tests, to every operation described in a flowchart, to every result of a sort, or to every word of a given code. In order to study the questionnaires which may be constructed, we consider a weighted quantity of information, the axioms of which were given by Belis and Guiaşu (1968), called the useful information.

Let $(P, U) = \{(p_i, u_i) \mid 1 \leqslant i \leqslant n\}$ be a set associated with $E = \{e_i \mid 1 \leqslant i \leqslant n\}$, p_i being the probability of e_i, and u_i its utility. Then the *useful information transmitted by* Q is defined as:

$$G(P, U) = \sum_{i=1}^{n} \left(u_i p_i \Big/ \sum_{i=1}^{n} u_i p_i\right) \log(1/p_i).$$

It is convenient to introduce the *preference* $v_i = u_i p_i$ of each answer $e_i \in E$, and its *normalized preference* $w_i = (v_i / \sum_{i=1}^{n} v_i)$. Then, the useful information transmitted by Q may be written as:

$$G_u(Q) = G(P, U) = \sum_{i=1}^{n} w_i \log(1/p_i).$$

In the arborescent case, some properties of this useful information have been exhibited (Guiaşu and Picard, 1971), and a definition of the useful information processed by a questionnaire has been proposed.

In this paper, we propose to study the latticoïd case, and to obtain some results about products of questionnaires, corresponding to dependent and independent processes.

2. Definitions

Let us consider a latticoïd questionnaire $Q = (X, \Gamma, p)$; every answer, belonging to E, has an utility and a probability. We want to study the acquisition of information during the processing; we need to affect an utility to every question and every one of Q, in order to introduce the useful information processed by Q. We can affect utilities to the b_i arcs ending in $e_i \in E$, by dividing u_i in b_i parts with the help of arbitrary parameters, in a way similar to Duncan (1974b). Nevertheless, in order to obtain easily a property of preference conservation in each vertex, we are going to restrict our study to the case when the utility $u(i, j)$, $(u(j))$, and the preference $v(i, j)$, $(v(j))$ of an arc (i, j), (a vertex j), are defined in the following way:

$$\forall j \in E \begin{cases} u(j) = u_j, \\ v(j) = v_j, \end{cases}$$

$$\forall j \in X - \{\alpha\} \ \forall i \in \Gamma^{-1} j \begin{cases} u(i, j) = u(j), \\ v(i, j) = u(i, j)\, p(i, j), \end{cases}$$

$$\forall j \in F \begin{cases} v(j) = \sum_{i \in \Gamma j} v(j, i), \\ u(j) = v(j)/p(j). \end{cases}$$

It follows easily that $v(i) = \sum_{j \in \Gamma^{-1} i} v(j, i)$.
Thus, the preference of each vertex is conserved, and

$$v(\alpha) = \sum_{i=1}^{n} v(e_i).$$

We shall call $w(i) = v(i)/v(\alpha)$ the normalized preference of a vertex i. The preference of an arc (i, j) is defined as:

$$w(i, j) = v(i, j)/v(\alpha) \qquad \forall i \in F \qquad \forall j \in \Gamma i.$$

The useful information $G(i)$ treated by a question $i \in F$ is defined as:

$$G(i) = \sum_{j \in \Gamma i} (w(i, j)/w(i)) \log(p(i)/p(i, j)).$$

The useful information processed by Q can therefore be defined as:

$$G(Q) = \sum_{i \in F} w(i)\, G(i) = \sum_{i \in F} \sum_{j \in \Gamma i} w(i, j) \log(p(i)/p(i, j)).$$

We know that, in the arborescent case, this later information can be written as:

$$\begin{aligned} G(Q) &= \sum_{i \in F} \sum_{j \in \Gamma i} w(j) \log(p(i)/p(j)) \\ &= \sum_{j \in X - \{\alpha\}} w(j) \log(1/p(j)) - \sum_{i \in F} w(i) \log(1/p(i)) \\ &= \sum_{j \in E} w(j) \log(1/p(j)). \end{aligned}$$

Then $G(Q) = G_u(Q)$.

Let us study the connection between $G(Q)$ and $G_u(Q)$ in the latticoïd case. If $G^1(i) = \sum_{j \in \Gamma^{-1} i} (w(j, i)/w(i)) \log(p(i)/p(j, i))$ is the information consumed by a vertex i, $i \neq \alpha$, we obtain:

$$G(Q) = G_u(Q) + \sum_{i \in X - \{\alpha\}} w(i)\, G^1(i).$$

This gives us the property:

PROPOSITION 1. *$G_u(Q) \leqslant G(Q)$. The equality holds if and only if Q is arborescent.*

3. Useful Information Transmitted by a Product of Questionnaires

3.1. *Branching Property of G_u*

Let Q_1 be a questionnaire, with $(n - 1)$ answers. We denote by Q its restricted product by an elementary dichotomous questionnaire Q_2 in its $(n - 1)$th answer x. We suppose that the utility of an answer y of Q is obtained as the product of the utility of x by the utility of the corresponding vertex in Q_2, if y is reachable from x. We also suppose that the roots of Q_1 and Q have the same preference, and that Q is associated to the couple (P, U). It follows that x has an utility equal to $u_x = (p_{n-1}u_{n-1} + p_n u_n)/(p_{n-1} + p_n)$, according to the assumption of Guiașu (1971), and that its probability is $p_x = p_{n-1} + p_n$.

We obtain easily the following result:

$$G_u(Q) = G_u(Q_1) + (w_{n-1} + w_n)\, G_u(Q_2).$$

Thus:

PROPOSITION 2. *The useful information G_u verifies the branching axiom, relatively to probabilities*:

$$\begin{aligned} &G_u(p_1, u_1; \ldots; p_n, u_n) \\ &\quad = G_u\left(p_1, u_1; \ldots; p_{n-2}, u_{n-2}; p_{n-1} + p_n, \frac{p_{n-1}u_{n-1} + p_n u_n}{p_{n-1} + p_n}\right) \\ &\qquad + \frac{p_{n-1}u_{n-1} + p_n u_n}{\sum_{i=1}^{n} p_i u_i} \\ &\qquad \times G_u\left(\frac{p_{n-1}}{p_{n-1} + p_n}, \frac{u_{n-1}(p_{n-1} + p_n)}{p_{n-1}u_{n-1} + p_n u_n}; \frac{p_n}{p_{n-1} + p_n}, \frac{u_n(p_{n-1} + p_n)}{p_{n-1}u_{n-1} + p_n u_n}\right). \end{aligned}$$

3.2. *Restricted Products of Q_0 by $Q_1, \ldots, Q_n$*

Let $Q_i = (X_i, \Gamma_i, p_i)$ be a questionnaire, $0 \leqslant i \leqslant n$; let its set of answers be denoted by E_i, and its root by α_i; let u_i, v_i, w_i be the utilities, preferences, and normalized preferences in Q_i, $0 \leqslant i \leqslant n$. If $E_0 = \{e_i \mid 1 \leqslant i \leqslant n\}$, we perform the successive restricted products of Q_0 in e_i by Q_i, $1 \leqslant i \leqslant n$, and obtain the questionnaire $Q = Q_0 \diamondsuit Q_1 \diamondsuit \cdots \diamondsuit Q_n$, the elements of which are denoted as in Sections 1 and 2. We suppose that the process represented by Q_i, $(1 \leqslant i \leqslant n)$, depends on the event associated with the answer e_i of Q_0. Let $e_k{}^i \in E_i$, and $x_k{}^i \in X_i$, denote the answers and vertices of Q_i, $1 \leqslant i \leqslant n$; we write $x_k{}^i \cdot e_i$ for the homologous vertex of $x_k{}^i$ in Q. The probabilities of $x_k{}^i$ in Q_i, conditioned by e_i are called $p_i(x_k{}^i \mid e_i)$. We define a conditional utility of $x_k{}^i$ in X_i by:

$$u_i(x_k{}^i \mid e_i) = u(x_k{}^i \cdot e_i)/u_0(e_i)$$

and a *conditional preference* by

$$\begin{aligned} v_i(x_k{}^i \mid e_i) &= u_i(x_k{}^i \mid e_i)\, p_i(x_k{}^i \mid e_i) \\ &= v(x_k{}^i \cdot e_i)/v_0(e_i). \end{aligned}$$

We may note that the utility $u(x_k{}^i \cdot e_i)$ characterizes the event associated with the vertex $x_k{}^i$ of Q_i when coming after the occurrence of the event associated with the answer e_i of Q_0, whereas the utility $u_0(e_i)$ characterizes the single event associated with e_i.

This leads to a *conditional normalized preference*

$$\begin{aligned} w_i(x_k{}^i \mid e_i) &= v_i(x_k{}^i \mid e_i)/v_i(\alpha_i \mid e_i) \\ &= w(x_k{}^i \cdot e_i)/w(\alpha_i \cdot e_i). \end{aligned}$$

The useful information of Q is equal to the following expression:

$$G_u(Q) = \sum_{e_i \in E_0} \sum_{e_k{}^i \in E_i} w(e_k{}^i \cdot e_i) \log(1/p(e_k{}^i \cdot e_i)).$$

If we introduce a conditional information of Q_i, $1 \leqslant i \leqslant n$:

$$G_u(Q_i \mid Q_0) = \sum_{e_k{}^i \in E_i} w_i(e_k{}^i \mid e_i) \log(1/p_i(e_k{}^i \mid e_i))$$

and if we consider the questionnaire $\overline{Q_i}$ obtained from Q_i by replacing the preference of $x_k{}^i$ by the preference of $x_k{}^i \cdot e_i$, $\forall x_k{}^i \in X_i$, $0 \leqslant i \leqslant n$, we obtain the following property:

PROPOSITION 3. *If the processes Q_i, $1 \leqslant i \leqslant n$, depend on the process Q_0, the useful information transmitted by the questionnaire Q verifies:*

$$G_u(Q) = G_u(\overline{Q_0}) + \sum_{i=1}^{n} w(\alpha_i \cdot e_i)\, G_u(Q_i \mid Q_0).$$

If we consider the case when the events represented by the answers of Q_i, $1 \leqslant i \leqslant n$, are independent of those represented by the answers of Q_0, we can write:

$$\forall i \in \{1,..., n\} \qquad w(\alpha_i \cdot e_i)\, w_i(e_k{}^i) = w(e_k{}^i \cdot e_i) \qquad \forall e_k{}^i \in E_i$$

and

$$w(\alpha_i \cdot e_i)\, G_u(Q_i \mid Q_0) = G_u(\overline{Q_i})$$

Thus:

PROPOSITION 4. *If the processes Q_i, $1 \leqslant i \leqslant n$, are independent of the process Q_0, the useful information transmitted by the questionnaire Q verifies:*

$$G_u(Q) = \sum_{i=0}^{n} G_u(\overline{Q_i}).$$

In the particular case, when all the questionnaires Q_i, $i \leqslant i \leqslant n$, are the same, say Q^*, we have:

$$G_u(Q) = G_u(\overline{Q_0}) + G_u(Q^*) \sum_{i=1}^{n} w(\alpha_i \cdot e_i).$$

Since $w(\alpha_i \cdot e_i) = w_0(e_i)$, we obtain

$$G_u(\overline{Q_0}) = G_u(Q_0)$$

and

$$\sum_{i=1}^{n} w(\alpha_i \cdot e_i) = 1.$$

PROPOSITION 5. *If the processes Q_0 and Q^* are independent, the useful information transmitted by their direct product verifies:*

$$G_u(Q_0 Q^*) = G_u(Q_0)\, G_u(Q^*).$$

4. Useful Information Processed by a Product of Questionnaires

The assumptions being those of Section 3, we have to extend our definitions of conditional utility and preference, to the arcs of every questionnaire Q_i, $1 \leqslant i \leqslant n$; $\forall x_k^i \in X_i - \{\alpha_i\}$ $\forall x_j^i \in \Gamma_i x_k^i$:

$$\begin{aligned}
u_i((x_k^i, x_j^i) \mid e_i) &= u_i(x_j^i \mid e_i) \\
v_i((x_k^i, x_j^i) \mid e_i) &= u_j(x_j^i \mid e_i)\, p_i((x_k^i, x_j^i) \mid e_i) \\
w_i((x_k^i, x_j^i) \mid e_i) &= v_i((x_k^i, x_j^i) \mid e_i)/v_i(\alpha_i \mid e_i) \\
&= w_i((x_k^i \cdot e_i, x_j \cdot e_i))/w(\alpha_i \cdot e_i).
\end{aligned}$$

The useful information processed by Q can be written as:

$$\begin{aligned}
G(Q) &= \sum_{t \in F} \sum_{r \in \Gamma t} w(t, r) \log(p(t)/p(t, r)) \\
&= \sum_{x_k^0 \in F_0} \sum_{x_j^0 \in \Gamma_0 x_k^0} w((x_k^0, x_j^0)) \log(p(x_j^0)/p((x_k^0, x_j^0))) \\
&\quad + \sum_{i=1}^{n} \sum_{x_k^i \cdot e_i \in \Gamma e_i \cap F} \sum_{x_j^i \in \Gamma_i x_k^i} w((x_k^i \cdot e_i, x_j^i \cdot e_i)) \\
&\quad \times \log(p(x_k^i \cdot e_i)/p((x_k^i \cdot e_i, x_j^i \cdot e_i))).
\end{aligned}$$

If we note that

$$p(x_k^i \cdot e_i)/p((x_k^i \cdot e_i, x_j^i \cdot e_i)) = p_i(x_k^i \mid e_i)/p_i((x_k^i, x_j^i) \mid e_i)$$

$\forall i \in \{1, \dots, n\}$, $\forall x_k^i \in X_i$, $\forall x_j^i \in \Gamma_i x_k^i$, and if we introduce a conditional information processed by Q_i, $1 \leqslant i \leqslant n$:

$$G(Q_i \mid Q_0) = \sum_{x_k^i \in F_i} \sum_{x_j^i \in \Gamma_i x_k^i} w_i((x_k^i, x_j^i) \mid e_i) \log(p_i(x_j^i \mid e_i)/p_i((x_k^i, x_j^i) \mid e_i))$$

we obtain the results analogous to those found in the case of transmitted information:

Proposition 6. *The useful information $G(Q)$ processed by the questionnaire possesses the property*:

(i) *if Q_i is dependent on Q_0, $1 \leqslant i \leqslant n$*:

$$G(Q) = G(\overline{Q_0}) + \sum_{i=1}^{n} w(\alpha_i \cdot e_i)\, G(Q_i \mid Q_0)$$

(ii) *if Q_i is independent of Q_0*, $1 \leqslant i \leqslant n$:

$$G(Q) = \sum_{i=0}^{n} G(\overline{Q_i}),$$

(iii) *if* $Q_i = Q^*$, $\forall i \in \{1, \ldots, n\}$:

$$G(Q) = G(Q_0) + G(Q^*).$$

5. Useful Routing Length

5.1. *Definition*

The classical routing length $L(Q)$ of a questionnaire Q is the mathematical expectation of the length of the paths going from the root to the answers. The elaboration of a useful routing length $L_u(Q)$ allows us to attach importance to some particular answers, by means of normalized preferences.

It is possible to define $L_u(Q)$, by an axiomatization analogous to the one which has allowed Picard (1972) to introduce $L(Q)$. We should use preferences of arcs and vertices conditioned by a path. We are going to restrict ourselves to introducing $L_u(Q)$ as a generalization of the results obtained in the arborescent case (Picard, 1972, Chap. 13.2)

$$L_u(Q) = \sum_{i \in F} w(i).$$

In the special case when, for every $e \in E$, all the paths going from the root to e have the same length $e(r)$, the questionnaire is said to be dedekindian. The following properties state:

Proposition 7. *If the questionnaire Q is dedekindian*:

$$L_u(Q) = \sum_{e \in E} w(e)\, r(e).$$

Proposition 8. *The useful routing length $L_u(Q)$ verifies the inequality*:

$$L_u(Q) \geqslant G_u(Q) + \log u(\alpha) - \sum_{e \in E} w(e) \log u(e).$$

Let us consider two questionnaires Q_0 and Q_1, and let us do their product Q, using evident notations:

$$L_u(Q) = \sum_{i \in F} w(i) = \sum_{i \in F_0} w(i) + \sum_{e_i \in F_0} \sum_{j \in \hat{\Gamma} e_i \cap F} w(j).$$

We know that

$$\sum_{i \in F_0} w(i) = \sum_{i \in F_0} (v_0(i)\, v_1(\alpha_1)/v(\alpha)) = \sum_{i \in F_0} (v_0(i)/v_0(\alpha_0)) = L_u(Q_0)$$

and that

$$\sum_{e_i \in E_0} \sum_{j \in \hat{\Gamma} e_i \cap F} w(j) = \sum_{e_i \in E_0} (v_0(e_i)/v_0(\alpha_0)) \sum_{j \in \hat{\Gamma} e_i \cap F} w_1(j) = L_u(Q_1).$$

PROPOSITION 9. *If the processes represented by* Q_0 *and* Q_1 *are independent, the useful routing length is additive*:

$$L_u(Q_0 Q_1) = L_u(Q_0) + L_u(Q_1).$$

5.2. *Huffman Algorithm*

C. F. Picard has stated that the analogy between L and L_u allows us to use the Huffman algorithm for constructing a L_u-minimal questionnaire Q_H, when the families A of the bases of questions and W of the normalized preferences of answers are given: we arrange A and W in ascending order, and we build a question by affecting to its issues the a_0 first elements of W, a_0 being the first element of A. Then we eliminate the used elements of W and insert their sum in W; we remove a_0 of A and do the same operation with the new sets A and W. The algorithm stops when we use the last element of A (1972, Chap. 13.2).

This algorithm, used in the case of the classical routing length, minimizes, at each step, the contribution of information of the constructed question. We may wonder if this property is conserved, under some condition, when we study L_u.

In the huffmanian arborescent questionnaire Q_H, the contribution of useful information $G''(i)$ of a question i is:

$$G''(i) = \sum_{j \in \Gamma_i} w(j) \log(p(i)/p(j)) \qquad \forall i \in F.$$

Denote by $x(w_1, ..., w_\alpha, p_1, ..., p_\alpha)$ a question obtained by connecting up α answers, of normalized preferences w_i and probabilities p_i, $1 \leqslant i \leqslant \alpha$. Put $p(x) = \sum_{i=1}^{\alpha} p_i$, and $w(x) = \sum_{i=1}^{\alpha} w_i$. Let us compare the contribution of information of two questions x and y, that we can build when we construct Q: we examine the sign of $\Delta = G''(y) - G''(x)$.

1st *Case*.

$$\begin{cases} x(w_1, ..., w_{\alpha-1}; p_1, ..., p_{\alpha-1}) \\ y(w_1, ..., w_\alpha \ ; p_1, ..., p_\alpha) \end{cases}$$

$$\Delta = w(x) \log(p(y)/p(x)) + w_\alpha \log(p(y)/p_\alpha).$$

Therefore

$$\Delta \geqslant 0.$$

2nd *Case.*

$$\begin{cases} x(w_1, \ldots, w_\alpha; \quad p_1, \ldots, p_\alpha) \\ y(w_1, \ldots, w_{\alpha-1}, w_\alpha'; p_1, \ldots, p_\alpha) \end{cases}$$

$$\Delta = (w_\alpha' - w_\alpha) \log(1/p_\alpha).$$

Thus $\Delta \geqslant 0$ if and only if $w_\alpha' > w_\alpha$.

PROPOSITION 10. *The minimum of useful information brought by a question is obtained by choosing the least base α_0, and connecting the α_0 answers having the least preferences among those having the same probabilities.*

3rd *Case.*

$$\begin{cases} x(w_1, \ldots, w_\alpha; p_1, \ldots, p_\alpha) \\ y(w_1, \ldots, w_\alpha; p_1, \ldots, p_{\alpha-1}, p_\alpha') \end{cases}$$

$$\Delta = w(x) \log(p(y)/p(x)) + w_\alpha \log(p_\alpha/p_\alpha').$$

Let us introduce

$$z = p(y) - p(x) = p_\alpha' - p_\alpha$$

and

$$\varphi(z) = w(x) \log(p(x) + z) - w_\alpha \log(p_\alpha + z),$$

then $\Delta = \varphi(z) - \varphi(0)$.

By studying the sign of $\varphi'(z)$, we obtain the result:

(i) if $u(x) > u_\alpha$ and $z > 0$ then φ is increasing;

(ii) if $u(x) < u_\alpha$ and $z < 0$ then φ is decreasing.

In both of these cases, $\Delta \geqslant 0$.

We conclude in saying that the Huffman algorithm, used in the case of useful routing length, does not always minimize the useful information brought at each step.

ACKNOWLEDGMENT

We express our thanks to the referee for his valuable suggestions.

RECEIVED: July 3, 1975; REVISED: March 19, 1976

REFERENCES

BELIS, M., AND GUIAŞU, S. (1968), A quantitative-qualitative measure of information in cybernetic systems, *IEEE Trans. Information Theory* **14**, 593–594.

BOUCHON, B. (1972), Réalisation de questionnaires et propositions logiques, Thèse 3° Cycle, Paris.

BOUCHON, B. (1974), Quelques résultats de la théorie des questionnaires réalisables, 7th Prague Conference on Information Theory, Statistical decision functions and random processes.

DUNCAN, G. T. (1974a), Heterogeneous questionnaire theory, *SIAM J. Appl. Math.* **27**, 59–71.

DUNCAN, G. T. (1974b), A characterization of Shannon entropy through lattice questionnaires, *Inform. Contr.* **26**, 219–224.

DUNCAN, G. T. (1975), Optimal diagnostic questionnaires, *Operation Research* **23**, 22–32.

GUIAŞU, S. (1971), Weighted entropy, *Reports on Mathematical Physics* **2**, 165–179.

GUIAŞU, S., AND PICARD, C. F. (1971), Borne inférieure de la longueur utile de certains codes, *C. R. Acad. Sci. Paris* **273A**, 248–251.

PICARD, C. F. (1970), Quasi-questionnaires, codes and Huffman's Length, *Kybernetika 6* **6**, 418–435.

PICARD, C. F. (1972), "Graphes et Questionnaires," Gauthier–Villars, Paris.

TERRENOIRE, M. (1970), Un modèle mathématique de processus d'interrogation: les pseudoquestionnaires, Thèse Sci. Math. Grenoble.