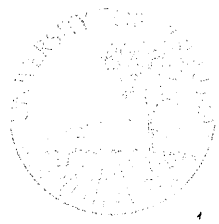
ISE-TR-88-67



A METHOD OF NEIGHBORHOOD FOR ANALYZING FREE ASSOCIATION

by

Sadaaki MIYAMOTO

Shinsuke SUGA

Ko OI

June 3, 1988

A Method of Neighborhood for Analyzing Free Association

S. Miyamoto*, S. Suga**, and K. Oi**

*Institute of Information Sciences and Electronics
University of Tsukuba, Tsukuba, Ibaraki 305, Japan

**Environmental Information Division, The National Institute
for Environmental Studies, Tsukuba, Ibaraki 305, Japan

ABSTRACT

A method for constructing two types of measures of association between a pair of symbols that distribute over a network is developed. The method is called here a neighborhood method in the sense that the construction of the measures is based on neighborhoods of vertices of the network. A family of methods of structural representations is developed using the framework proposed formerly by the authors. Moreover, a new method of hierarchical cluster analysis is developed. These methods are applied to structural representation of data on cognition of living environment of local residents. The data are obtained from a survey by questionnaire that requests free association about living environment. The result of data analysis shows how different structures of the cognition reflect different backgrounds of two districts and serves as a guideline for decision making about improvement of living environment.

1. Introduction

Methods developed for structural modeling [1], [2] have frequently been applied to represent and to understand structure of human cognition of complex systems. When these methods are applied for representing a cognitive structures of a group of people, a problem of aggregation of individual structure into a whole structure should be studied. This means that a method of structural modeling should include a feature of statistical analysis to deal with such a problem. Since important problems in social studies require analysis and representation of a structure of cognition for a large number of people, a method of structural modeling that includes statistical analysis is promising as a powerful tool of analysis in these problems.

Thus, the authors studied a method of digraph representation with cluster analysis and applied it to cognition of living environment of local residents [3]. The present paper deals with the same type of application based on a new mathematical model. Our motivation here is analysis of free (psychological) association which was used in the survey by questionnaire above mentioned (See [3] for the detail.) The present method may be considered as an improved technique and an elaboration of the previous method from applicational viewpoint. On the other hand, the method developed here is new from theoretical viewpoint. The method herein is based on a topological structure of free association which the previous method does not have. The method is called here a neighborhood method in the sense that the theory is developed on a definition of a neighborhood of a network.

One of remarkable characteristics is that the present method uses a framework developed in the previous paper [3]: four methods of analyzing cognitive structure is developed as an analogy of the previous methods in four quadrants (Fig. 1. See also Fig. 2 in [3]). This framework includes a new method of hierarchical clustering which is called here a model reference algorithm. The new algorithm uses the neighborhood model throughout the whole procedure of forming clusters. The idea of a model reference algorithm provides a new principle of developing various algorithms in hierarchical cluster analysis.

When the neighborhood method is applied to the analysis of free association, the network structure in the theory will be replaced by a simple linear structure. This simplification is necessary in the present application, since free association of people does not have, in general, an obvious structure of network. At the same time the theory is developed for a more general structure of a network so that it the theory can be used for wider appplication including analysis of a collection of discourses and/or texts.

2. Measures defined on neighborhood

Let $W=\{a,b,c,\ldots\}$ be a finite set of symbols or words. Assume that $G=\{V,E\}$ is an undirected graph (network). For each $w \in V$, $U_n(w)=\{ v \mid v \in V,$ v is reachable from w with the length of the path less than or equal to n $\}$. If the distance $d(v,w)$ between a pair of vertices v and w is defined as the minimum number of path connecting v and w, $U_n(w)=\{ v \mid v \in V, d(v,w) \leq n \}$.

Assume that there is a map from V into W. that is. each $v \in V$ is named as a or b or c. and so on. In other words. each v is an occurrence of a word in W. Therefore if v is an occurrence of $a \in W$. we write $v=a_i$ with index. Then. V is represented as a set of word occurrences $V=\{a_i, a_j, \ldots b_k, b_\ell \ldots c_m, c_n, \ldots\}$. We define that $A=\{a_i, a_j, \ldots\}$ is a set of all occurrences of $a \in W$ ($A \subset V$) and $B=\{b_k, b_\ell, \ldots\}$, and so on. Moreover M(A) is the number of elements in $A \subset V$. Let L and L' be linear orderings of V. They are regarded as two ordered sequences of all the members of V.

We consider several measures that show degree of association between a pair of elements in W. Below we take a and b as a generic example for the pair in the sense that the results that holds for the pair (a,b) is valid for any pair of elements in W. These measures of association are considered on $U_n(v)$ above defined. Specifically, we consider the following procedure.

1. [Define $r(b,a_i)$] For any $a_i \in A$. we consider a measure $r(b,a_i)$. The value of $r(b,a_i)$ is determined by the occurrence of some b_j in the neighborhood $U_n(a_i)$. In the below we consider ways to determine $r(b,a_i)$.

2. [Define r(b,a)] The measure of association between a and b: r(b,a) is defined as follows.

$$r(b,a) = \sum_{\text{all } a_i \in A} r(b,a_i) \qquad (1)$$

Our first task is to consider what kind of $r(b,a_i)$'s can be defined as reasonable measures of association. For this purpose

we should give explicit definition 1.0-1.3 as the refinement for the step 1.

1.0. As described above. $r(b.a_i)$ should be determined by the occurrence of some b_j in $U_n(a_i)$. Note that there may be several occurrences of b_j in $U_n(a_i)$. When we find. at first. an occurrence b_j in $U_n(a_i)$. we will set $r(b.a_i)=1$. Or. more precisely.

1.1 First set $r(b.a_i) \leftarrow 0$.

1.2 If we find an occurrence $b_j \in U_n(a_i)$. then $r(b.a_i) \leftarrow r(b.a_i) + 1$

Here we have a question. What should we do for other $b_k \in U_n(a_i)$. $b_k \neq b_j$? Of course we must not count doubly the same pair of occurrences $(a_i.b_j)$, therefore we should introduce a marking procedure to inhibit double counting. There are three alternatives of the marking as follows.

1.3-1 After we count b_j in step 1.2. we mark the pair $(a_i.b_j)$ [such as $\overline{(a_i.b_j)}$] so that this pair will not be doubly counted. Then we continue step 1.2 for all the other b_k's in $U_n(a_i)$. Denote the measure r by the present method of the marking as $r_1(b.a_i)$. Namely. $r_1(b.a_i)$={the number of b_j's in $U_n(a_i)$}.

1.3-2 After we count b_j in step 1.2 we mark a_i [as $\bar{a}_i$]. We define the measure of association by this method as $r_2(b.a_i)$. Thefefore in this case we have

$$r_2(b.a_i) = 0 \quad \text{iff} \quad B \cap U_n(a_i) = \emptyset$$

$$r_2(b.a_i) = 1 \quad \text{iff} \quad B \cap U_n(a_i) \neq \emptyset .$$

1.3-3 After we count b_j in step 1.2 we mark b_j [as $\overline{b}_j$] to inhibit double counting. Then we continue step 1.2 for all the other b_j's that are not yet counted. We define the measure r obtained by this method as $r_3(b,a_i)$. Moreover we should remark that this measure depends on an ordering of V. Therefore in this method we should define the procedure more precisely.

[Assumption applied only to r_3] In the definition of $r_3(b,a) = \sum r_3(b,a_i)$ for all $a_i \in A$, we will visit each $a_i \in A$ according to the order determined by a particular ordering. That is, if we use the ordering L, we will define $r_3(b,a;L) = \sum r_3(b,a_i;L)$ that shows explicit use of L. Namely,

$r_3(b,a_i;L)$ = {the number of b_j's, $b_j \in U_n(a_i)$, that are not marked yet using the ordering L to visit vertices in V} .

Note that the mark $\overline{(a_i,b_j)}$, $\overline{a_i}$, or $\overline{b_j}$ is valid throughout the procedure to define $r_1(b,a)$, $r_2(b,a)$, or $r_3(b,a;L)$. Therefore an element that is once marked will not be counted again in the procedure of defining the measure for the pair (a,b). For example, a $b_j \in U_n(a_i)$ that is marked may occur later in another $U_n(a_k)$ according to the ordering L. The occurrence b_j will not be counted to add unity to $r_3(b,a_k;L)$. Of course all the marks should disappear in calculating the measure for another pair such as (a,c).

We call the above procedures as simple procedures when we add unity ($r_k(b,a_i) \leftarrow r_k(b,a_i) + 1$, k=1,2.3) when some b_j is counted. On the other hand, we call a procedure as a weighted procedure when we add some numbers that is determined by

$d(a_i, b_j)$. Let us assume a monotone nonincreasing function $f: R^+ - \{0\} \to R^+$. A weighted procedure is obtained by modifying step 1.2:

1.2' If we find an occurrence $b_j \in U_n(a_i)$, then

$$r(b, a_i) \leftarrow r(b, a_i) + f(d(b_j, a_i)) \quad . \qquad (2)$$

Remark. The above consideration shows not only three methods to define three kinds of the measures of association but also it exhibits that it is very difficult to consider another type of a definition of the measure of association so long as we assume the basic procedure of step 1 and step 2. []

Remark. In this paper we are mainly concentrated on simple procedures for two reasons. One reason is that it is difficult to consider weighted measures of r_2 and r_3 types, as we will show later. The other reason is that we do not have a proper interpretation of a weighted measure in terms of probability which we introduce for the simple measures for the purpose of applying these measures to hypothesis testing of association structures of different population in a later section. []

First we show that the r_3 type measure is proved to be independent of the ordering.

Prop. 1 For any two orderings L and L',

$$r_3(b, a; L) = r_3(b, a; L') \quad .$$

(Proof) Note that $r_3(b, a; L)$ is equal to the number of marked b_j's by the procedure 1 and 2. (Or more precisely, 1.1, 1.2,

1.3-3, and 2.) Let $B_a = \{ b_j \mid b_j \in B.\ d(b_j, a_i) \leq n,$ for some $a_i \in A \}$. Then each $b_j \in B_a$ is marked as $\overline{b_j}$, whereas any $b_k \in B - B_a$ is not marked. Therefore for any ordering L, $r_3(b,a;L) = M(B_a)$. The right hand side of the above equation is independent of a particular ordering. []

Hence we write simply as

$$r_3(b,a) = r_3(b,a;L) = M(B_a)$$

by any ordering of V.

Prop. 2

$$r_3(a,b) = r_2(b,a) \quad .$$

(Proof) If we define $A_b = \{ a_i \mid a_i \in A,\ d(a_i, b_j) \leq n,$ for some $b_j \in B \}$, it is easy to see that $r_2(b,a) = M(A_b)$. On the other hand, we already showed in the proof of Prop. 1 that $r_3(a,b) = M(A_b)$. []

Prop. 3

$$r_1(a,b) = r_1(b,a) \quad .$$

(Proof) The number of marked pairs $\overline{(a_i, b_j)}$ for calculating either of $r_1(a,b)$ or $r_1(b.a)$ is equal to the total number of pairs: $\{(a_i, b_j) \mid a_i \in A.\ b_j \in B,\ d(a_i, b_j) \leq n \}$. []

Remark. Proposition 2 shows that we need not distinguish between r_2 and r_3. Thus we obtain only two types of the measures r_1 and r_2 that are simple. It should be noted that in general r_2 (and r_3) does not have the property of symmetry that is valid for r_1.

As is shown above, we can define a weighted procedure. Let us consider 1.3-1 with (2): mark (a_i, b_j). We write the measure defined by this as $r_1(b,a;f)$ to show explicitly the dependence of r_1 on a monotone nonincreasing function f. In other words,

$$r_1(b,a;f) = \sum_{\text{all } a_i \in A} r_1(b,a_i;f)$$

$$r_1(b,a_i;f) = \sum_{\text{all } b_j \in U_n(a_i)} f(d(b_j,a_i))$$

Note that the simple measure $r_1(b,a) = r_1(b,a;I)$, where $I(x)$ is the identity $I(x) = 1$ for all x.

Prop. 4

$$r_1(a,b;f) = r_1(b,a;f)$$

(Proof) Let $\bar{f}: R^+ \to R^+$ be

$$\bar{f}(x) = \begin{cases} 0 & (x = 0) \\ f(x) & (x > 0) \end{cases}$$

and $g(x;n): R^+ \to R^+$ be

$$g(x;n) = \begin{cases} x & (0 \leq x \leq n) \\ 0 & (x > n) \end{cases}$$

Then,

$$r_1(b,a;f) = \sum_{\substack{\text{all } a_i \in A \\ b_j \in B}} \bar{f}(g(d(b_j,a_i);n)) = r_1(a,b;f)$$

[]

On the other hand, if we try to define weighted measures $r_3(b,a;f)$ we must define at first $r_3(b,a;f;L)$ by an ordering L. It is easy to see that the measure $r_3(b,a;f;L)$ now depends on a particular ordering, i.e., $r_3(b,a;f;L) \neq r_3(b,a;f;L')$ in general. Therefore we can not define $r_3(b,a;f)$ that is independent of a particular ordering. We have the same problem in defining $r_2(b,a;f)$. Therefore we do not attempt to define r_2 (or r_3) by a weighting procedure.

3. Probabilistic argument and framework

In a foregoing paper we discussed a method, or more precisely, a family of methods for representing a system structure [3]. The methods therein were applied to representation of association structures. There were two main points in these methods of representation. (See Fig. 1.)

1. Two kinds of measures of association for a pair of words were used: one was a symmetric measure of similarity and the other was a nonsymmetric measure. The symmetric measure was used for clustering and the nonsymmetric measure was used to obtain a digraph representation of association structure.
2. The two measures had statistical interpretation. More precisely, they were considered as estimates of two probabilistic parameters. Therefore two methods of hypothesis testing were developed for comparing structures of association from different populations.

Here we will try the same approach for analyzing free association using the method of neighborhood. We have already

developed a symmetric measure $r_1(b,a)$ and a nonsymmetric measure $r_2(b,a)$. Therefore the point 1 is satisfied. We show that r_1 and r_2 by the simple procedure have a probabilistic character.

Let us now define two normalized quantities q_1 and q_2:

$$q_1(b,a) = \frac{r_1(b,a)}{M(A)M(B)}$$

$$q_2(b,a) = \frac{r_2(b,a)}{M(A)}$$

Prop. 5 $0 \leq q_1 \leq 1$ and $0 \leq q_2 \leq 1$.

$q_1(b,a) = 0$ iff for any $b_j \in B$ and $a_i \in A$, $d(b_j, a_i) > n$.

$q_1(b,a) = 1$ iff $\max_{a_i \in A,\ b_j \in B} d(b_j, a_i) \leq n$.

$q_2(b,a) = 0$ iff for any $b_j \in B$ and $a_i \in A$, $d(b_j, a_i) > n$.

$q_2(b,a) = 1$ iff for any $a_i \in A$ there exists a $b_j \in B$ such that $d(b_j, a_i) \leq n$.

(Proof) It is easy to show these relations from the definition of q_1 and q_2, therefore the proof is omitted here. []

In relation to two quantities q_1 and q_2, we introduce two probabilistic parameters p_1 and p_2 as follows.

1) Let us choose randomly $a_i \in A$ and $b_j \in B$ and let p_1 be the probability:

$$p_1 = \text{Prob}\{\, d(b_j, a_i) \leq n \,\}.$$

2) Let us choose randomly $a_i \in A$, then

p_2 = Prob{ there exists at least one $b_j \in B$ such that

$$d(b_j, a_i) \leq n \} .$$

It is clear that the above q_1 and q_2 are estimates of P_1 and P_2, respectively.

In the foregoing paper [3] we introduced a "framework" for four methods of structural representation. That is we filled in all the four corners of Fig. 1. Two quantities that satisfy the above points 1 and 2 can be used as elements of the framework. Here we defined two measures q_1 and q_2 that satisfy the point 1 and 2. Therefore we obtained a new family of methods based on the consideration of the neighborhood.

Now we define four methods in Fig. 2 in the framework as follows [3].

i) [Digraph based on q_2] Let α and β be two threshold parameters ($0 < \alpha < 1$, $\beta > 1$). We define two kinds of edges:

$$a \rightarrow b \quad \text{iff} \quad q_2(b,a) \geq \alpha,\ q_2(b,a) \geq \beta q_2(a,b)$$

$$a \longleftrightarrow b \quad \text{iff} \quad q_2(b,a) \geq \alpha,\ q_2(a,b) \geq \alpha,$$
$$\beta^{-1} < q_2(b,a)/q_2(a,b) < \beta .$$

ii) [Clustering based on q_1] We discuss a new algorithm developed for neighborhood method in the next section.

iii) [Hypothesis testing based on P_1 and P_2] Assume that $q_1^k(b,a)$, $q_2^k(b,a)$, and $M(A^k)$ with the superscript k means the quantities for two populations shown by k=1,2. Then the hypothesis $P_2^1 = P_2^2$ is tested by applying the 2 x 2 table in Fig.3. (See [4]). We define

$$O_{11} = q_2^1(b,a), \quad O_{12} = M(A^1) - O_{11}$$

$$O_{21} = q_2^2(b,a), \quad O_{22} = M(A^2) - O_{21} \ .$$

The following T is calculated and comparred against $x(1-\alpha)$. the chi-square distribution with one degree of freedom [4]:

iv) The hypothesis $p_1^1 = p_1^2$ is tested in the same way. We define

$$O_{11} = q_1^1(b,a) \quad O_{12} = M(A^1)M(B^1) - O_{11}$$
$$O_{21} = q_1^2(b,a) \quad O_{22} = M(A^2)M(B^2) - O_{21} \ .$$

4. A model reference algorithm for hierarchical cluster analysis

An algorithm of hierarchical agglomerative clustering that uses the neighborhood method throughout the process of forming clusters is developed. For this purpose we begin by a review of general procedure of hierarchical agglomerative clustering. In the sequel we consider solely hierarchical agglomerative clustering, therefore we call it simply as clustering or hierarchical clustering.

Procedure HC (a general description of hierarchical clustering)
HC0 (assumption) The set of objects to be clustered is $X=\{x_1, x_2, \ldots, x_n\}$. Assume that the similarity between a pair of objects is defined by some mathematical model and is denoted as $s(x_i, x_j)$, $1 \le i, j \le n$, $i \ne j$. (The word similarity means that a large

value of $s(x_i, x_j)$ implies that x_i and x_j are similar; a small value of s means that x_i and x_j are not similar.) Let N be the number of clusters in each step of successive formations of clusters. Let $Y_1, Y_2, \ldots, Y_N$ be clusters, that is, disjoint partition of X: $Y_1 \cup Y_2 \cup \ldots \cup Y_N = X$, $Y_i \cap Y_j = \emptyset$, $i \neq j$. Let $s(Y_i, Y_j)$ be similarity between a pair of clusters. (The way of definition of $s(Y_i, Y_j)$ is given after the description of procedure HC.)

HC1 Let $N := n$; $Y_i := \{x_i\}, i = 1, 2, \ldots, N$;

$s(Y_i, Y_j) := s(x_i, x_j)$, $1 \leq i, j \leq N$, $i \neq j$.

HC2 Find the maximum of $s(Y_i, Y_j)$, $1 \leq i, j \leq N$, $i \neq j$.

Assume that $s(Y_p, Y_q) = \max_{\substack{1 \leq i, j \leq N, \\ i \neq j}} s(Y_i, Y_j)$

HC3 Merge Y_p and Y_q, and let $Y_r := Y_p \cup Y_q$. Save information that is necessary for the output of the dendrogram, such as the similarity level of the merge, members in Y_p and in Y_q, and so on.

HC4 $N := n-1$. If $N=1$, then output the dendrogram that shows process of successive formation of clusters using the saved information.

HC5 Update similarities $s(Y_r, Y_i)$, $i = 1, \ldots, N$, $i \neq r$. Go back to step HC2.

End of HC.

Remark The above procedure does not include detailed description of hierarchical clustering such as what type of information structure is needed for the output of a dendrogram and algorithm for drawing a dendrogram. These details are unnecessary for the discussion of a new algorithm which will be described below. []

The underlined part, method of update of similarity between a pair of clusters, is essential in hierarchical clustering. First we show three well-known methods of the update.

a) (single link) $s(Y_r, Y_i) = \max_{\substack{v \in Y_r \\ w \in Y_i}} s(v,w)$

b) (complete link) $s(Y_r, Y_i) = \min_{\substack{v \in Y_r \\ w \in Y_i}} s(v,w)$

c) (average link) $s(Y_r, Y_i) = \frac{1}{|Y_r| \cdot |Y_i|} \sum_{\substack{v \in Y_r \\ w \in Y_i}} s(v,w)$

($|Y_i|$ is the number of elements in Y_i.)

Let us consider relation between the two kinds of similarities $s(x_i, x_j)$ between a pair of objects and $s(Y_i, Y_j)$ between a pair of clusters. First, $s(x_i, x_j)$ is defined by some mathematical model that is named here as MDL. We write $s(x_i, x_j) = Proc(x_i, x_j; MDL)$ to show explicitly that $s(x_i, x_j)$ is calculated by a procedure based on MDL. In HC1, $s(Y_i, Y_j) = s(x_i, x_j) = Proc(x_i, x_j; MDL) = Proc(Y_i, Y_j; MDL)$. On the other hand, after update of similarity $s(Y_r, Y_i) \neq Proc(Y_r, Y_i; MDL)$ or $Proc(Y_r, Y_i; MDL)$ is not defined. In other words, a hierarchical clustering uses two different mathematical models: one is for defining $s(x_i, x_j)$; the other is for updating $s(Y_i, Y_j)$. The centroid method and the Ward method [5] are exceptional. They are based on a single mathematical model of Euclid geometry. In centroid method, $Proc(x_i, x_j; C)$ = {Euclid distance between x_i and x_j}. Then $Proc(x_i, x_j; C)$ is naturally extended to $Proc(Y_i, Y_j; C)$={Euclid distance between centroids of Y_i and Y_j}.

In case of the Ward method $Proc(Y_i, Y_j; W)$ is directly defined: $Proc(Y_i, Y_j; W)$ = {error sum of squares from centroid of $Y_i \cup Y_j$ } - {error sum of squares from centroid of Y_i} - {error sum of squares from centroid of Y_j}

Remark The symbols C and W represents mathematical models of centroid method and the Ward method, respectively. Note that in these methods the measure for the clustering are not similarity but dissimilarity. []

Now we propose an idea of model reference hierarchical clustering:

Assume that a procedure of calculation of a similarity measure (or dissimilarity measure) between x_i and x_j based on MDL (i.e., $Proc(x_i, x_j; MDL)$) is extended in a natural way to a procedure of calculation of a similarity (or dissimilarity) between a pair of subsets $Proc(Y_i, Y_j; MDL)$ based on the same model MDL. If in HC5, the update uses $Proc(Y_i, Y_j; MDL)$ based on the same model MDL:

$$s(Y_r, Y_i) = Proc(Y_r, Y_i; MDL)$$

then we call the hierarchical clustering with this update procedure as a model reference algorithm for hierarchical clustering.

The centroid method and the Ward method are examples of the model reference algorithms by this definition. The above idea is used not only on Euclid geometry model but also other types of mathematical models. For example, we showed already another type

of a model reference algorithm based on extensions of binary measures of similarity using fuzzy sets [6].

Thus, a model reference algorithm uses a single mathematical model for both the initial definition of similarity and the update of the similarity. An advantage of a model reference algorithm is that levels of similarity of cluster formations are meaningful when we refer to the mathematical model, whereas in a former type of algorithms such as a), b), and c) the merge level is artificial and has no relation to the original model. On the other hand, a drawback of a model reference algorithm is that a tree reversal [5] in a dendrogram may occur. A tree reversal means that a cluster $Y_i \cup Y_j$ formed at the similarity level s_1 may be merged with Y_k to form $(Y_i \cup Y_j) \cup Y_k$ at the level s_2, and it may occur that $s_2 > s_1$. In other words, a cluster formed earlier may have the merge level of similarity less than the merge level of another cluster formed later. In the centroid method this property of tree reversal may occur.

Now we apply this concept of a model reference algorithm to the neighborhood method q_1. If we denote the calculation of q_1 as Proc(a, b ;N), it is shown that the procedure is naturally extended to clusters λ and ζ: Proc(λ, ζ;N). Moreover the derived method does not have the tree reversal.

Let λ and ζ are two disjoint subsets of W. Assume that $\lambda=\{a,c,\ldots\}$ and $\zeta=\{b,d,\ldots\}$. Then occurrences λ_i and ζ_j of λ and ζ are naturally defined as all the occurrences of members in λ and ζ. We define also Λ={set of elements in the occurrences of λ} and Z = {set of elements in the occurrences of ζ}. Therefore, $\Lambda = \{\lambda_1, \lambda_2, \ldots\} = \{a_1, a_2, c_1, c_2, \ldots\}$, $Z = \{\zeta_1, \zeta_2, \ldots\} = \{b_1, b_2, d_1, d_2, \ldots\}$.

Now we can define $r_1(\lambda,\zeta)$ by the same procedure as we define $r_1(a,b)$. Note that

$$r_1(\lambda,\zeta) = \sum_{\substack{a\in\lambda \\ b\in\zeta}} r_1(a,b) \tag{3}$$

is valid.

Prop. 6 Assume that λ, ζ, and ω are disjoint subsets of W, then

$$M(\Lambda\cup Z) = M(\Lambda) + M(Z)$$

$$r_1(\lambda\cup\zeta,\omega) = r_1(\lambda,\omega) + r_1(\zeta,\omega) .$$

(Proof) The first equation follows from $\Lambda\cap Z = \emptyset$. The second equation follows from (3). []

Cor. 1 Assume that λ, ζ, and ω are disjoint subsets of W, then

$$q_1(\lambda\cup\zeta,\omega) = \frac{r_1(\lambda,\omega) + r_1(\zeta,\omega)}{[\, M(\Lambda) + M(Z) \,]\, M(\Omega)} \tag{4}$$

where Ω is the set of occurrences of ω. []

Now, we state two lemmas.

Lemma 1 Assume that four real numbers $t_1 \geq 0$, $t_2 \geq 0$, $u_1 > 0$, $u_2 > 0$ satisfy

$$\frac{t_1}{u_1} \leq \frac{t_2}{u_2} .$$

Then,

$$\frac{t_1}{u_1} \leq \frac{t_1 + t_2}{u_1 + u_2} \leq \frac{t_2}{u_2} .$$

(Proof) Omitted. []

Lemma 2 A sufficient condition that a dendrogram generated by the procedure HC does not have any tree reversal is that the merged

cluster $Y_r (=Y_p \cup Y_q)$ in HC5 satisfies

$$\max[s(Y_p, Y_i), s(Y_q, Y_i)] \geq s(Y_r, Y_i) \qquad (5)$$

for all $i=1,2,\ldots,N$, $i \neq r$, $i \neq p$, $i \neq q$.

(Proof) Consider a set S_1 of similarities $S_1=\{s(Y_i, Y_j)\}$ before the merge and $S_2=\{s(Y_r, Y_i),\ldots\}$ after the update. The above condition shows that

$$\max_{s \in S_1} s \geq \max_{s \in S_2} s$$

Therefore the merge levels are nonincreasing as the successive formation of clusters proceeds. []

Cor. 2

$$\min[q_1(\lambda,\omega), q_1(\zeta,\omega)] \leq q_1(\lambda \cup \zeta, \omega) \leq \max[q_1(\lambda,\omega), q_1(\zeta,\omega)].$$

(Proof) Obvious from Cor. 1 and Lemma 1. []

Let us consider now hierarchical clustering of W based on the neighborhood method. Consider a version of the algorithm HC in which the update at HC5 is performed by (4). It is clear by the above explanation that this version of HC is a model reference algorithm. Moreover from Cor. 2 any tree reversal does not occur in this method.

5. Application to data obtained from free association

5A. A survey based on free association

A survey by questionnaire was conducted on living environment of local residents in two areas. The questionnaire asked responders the following question (cf. [3]): "What do you associate with the words: easiness of living and happiness in living?" Responders are requested to write down their

associations freely. The two areas of the survey are called here SETAGAYA and DAIGO. SETAGAYA is in a residential area in Setagaya ward in Tokyo. DAIGO is in a village in the northern part of Ibaraki prefecture. The latter region is an agricultural area surrounded by hills and mountains.

The data for the analysis here are free associations of 38 responders in SETAGAYA and those of 52 responders in DAIGO. Associations in each area are connected into one sequence. At the same time in a sequence two associations of two responders are separated by a large number of dummy words as in Fig.4a so that a neighborhood would not connect words in associations of two different responders. Therefore we obtained two sequences G_s for SETAGAYA and G_d for DAIGO. We deal with G_s and G_d as the networks by defining edges that connect only neighboring words (Fig. 4b).

Fifty-four words were selected as members in W in SETAGAYA and another set of 54 words were selected in DAIGO. Then the method of neighborhood was applied: digraphs were defined by i) in Section 3; clusters were generated using the model reference algorithm in Section 4; hypothesis testings were performed using iii) and iv) in Section 3.

Two digraphs with clusters are shown as Fig. 5 for SETAGAYA and Fig. 6 for DAIGO. Edges are defined by i) with the thresholds $(\alpha,\beta) = (0.30,1.5)$ in both figures. Areas surrounded by dashed lines are clusters obtained by the method in Section 4: two dendrograms were generated by the model reference algorithm and clusters formed at the level of similarity $q_1=0.006$ in both dendrograms are shown on the two digraphs. Dendrograms

themselves are omittd here. Result of the hypothesis testings is shown as Tables 1 and 2. Table 1 shows the result of hypothesis testing on p_1 and Table 2 shows that on p_2. The level of significance is alpha=0.05 in both tables. The hypothesis testing is defined on pairs of 26 words that are common to both areas of the survey. The upper part of Table 1 shows pairs of associated words in which the result of the test tells that (p_1 in SETAGAYA) > (p_1 in DAIGO) (Associations in SETAGAYA is stronger than in DAIGO.) The lower part of Table 1 exhibits pairs of associated words in which the test shows that associations in DAIGO is stronger than in SETAGAYA. In the same manner, the upper part (resp. lower part) of Table 2 shows pairs of associated words satisfying (p_2 in SETAGAYA) > (p_2 in DAIGO) (resp. (p_2 in DAIGO) > (p_2 in SETAGAYA).)

5B Discussion on the result of structural representations

Several points of comparison of the two structures are described in the following.

1) [convenience] Each region has a cluster concerning convenience. In Fig. 5, the cluster has the largest number of words and the edges of association concentrate on three words: "traffic", "convenient", and "near". The cluster has also "public", "facilities", "school", and "hospital". In Fig. 6, the cluster contains "traffic", "convenient", and "school". Concentration of edges is weaker in DAIGO than in SETAGAYA and "public", "facilities" are not observed in the cluster of convenience in DAIGO.

2) [clusters with "green"] Each region has a cluster that contains "green". The two clusters in both regions have "green", "air", and "noise" in common. In SETAGAYA, "spacious", "garden", and "trees" are included in the cluster; in DAIGO "river", "stream", "water", and "hills/mountains" are seen. Thus, the cluster in DAIGO covers a larger scenery of natural surroundings, whereas the view in SETAGAYA is limited to a smaller space such as gardens.

3) [household, human factors] In Figs 5 and 6, both regions have words "house", "home", and "family". The word "house" here means a building, whereas "home" and "family" means human relationship of a household. In SETAGAYA we observe a cluster having "family" and "self". The cluster includes also "room", "room arrangement", and "space". In DAIGO, we find a cluster that has "family", "self", "people", "neighbor", and "community". In SETAGAYA "neighborhood", "people", and "relation" form another cluster. In this way, human relationship in SETAGAYA is associated with rooms inside the house. At the same time relationship with neighboring people is considered to be in another category. On the other hand, human relationship in a family is directly related to relationship with community and people in neighborhood in DAIGO.

Human relation in neighborhood is an important factor in Japanese living environment. This factor is stronger in DAIGO than in SETAGAYA as we see Figs. 5, 6, and the result of the hypothesis testing. Table 1 shows that relation with neighboring people is related to convenience in DAIGO. This observation agrees with a tendency observed in Fig. 6. We find in Fig. 6

that "neighborhood" and "relation" are included in a cluster of convenience.

To summarize, the association structure in SETAGAYA in centered to convenience, with facilities and house as buildings; whereas the structure in DAIGO shows that human factor is stronger in this region. The human factor includes self, family, neighbors, and community, which is not exhibited in SETAGAYA.

6. Conclusions

As an analogy of Fig. 1, we observe significances of the present method in four quadrants described by two axes of (methodological)/(applicational) aspect and (present results)/(possibilities in the future) of Fig. 7.

(MP: methodology-present) A new method of the neighborhood has been developed that assumes a topological structure in association or discourse. According to the framework of symmetric/nonsymmetric measures and macroscopic/local characteristics, the neighborhood method has been augmented into four methods that fill in the framework.

(AP: application-present) The method has been applied to data of free association obtained from a survey by questionnaire. The result has exhibited categories shown by clusters of cognition on living environment, key concepts in these clusters, and how the same key concepts of living environment reflect different backgrounds of two districts.

(MF: methodology-future) The method of neighborhood should be

applied to a network structure, e.g., semantic net, derived from a discourse, a text, or a set of knowledge. For this purpose relations in a semantic net should be simplified for the aggregation of the structures. At the same time, the basic method of neighborhood should be generalized so that it distinguishes different types of relations. One possibility is to consider a n-ary measure of relation instead of binary measures considered here.

(AF: application-future) A new form of free association or free writing in a survey by quetionnaire has a great advantage over the conventional form in the sense that the free form admits various types of unexpected or exceptional responses. This feature is very important in studying environmental cognition, since decision making for improvement of living environment should not fail to grasp these unexpected structures. The reason why this type of the freely answered questionnaire has not been studied is due to lack of advanced techniques for the data analysis. Now we have shown a new method of analysis using clustering, digraph representation, and statistical hypothesis testing. The result is summarized into a figure that show structures of cognition in a compact manner. Thus the present method will give a major influence for studies of the new type of survey by questionnaire.

References

1. A. P. Sage, Methodology for Large Scale Systems, New York, McGraw-Hill, 1977.

2. J. N. Warfield, Societal Systems, Planning, Policy, and Complexity, New York, Wiley, 1976.

3. S. Miyamoto, K. Oi, O. Abe, A. Katsuya, and K. Nakayama, Directed graph representations of association structures: a systematic approach, IEEE Trans., Syst., Man, and Cybern., Vol.SMC-16, No.1, pp.53-61, 1986.

4. W. J. Conover, Practical Nonparametric Statistics, 2nd ed., New York, Wiley, 1980.

5. M. R. Anderberg, Cluster Analysis for Applications, New York, Academic Press, 1973.

6. S. Miyamoto and K. Nakayama, Similarity measures based on a fuzzy set model and application to hierarchical clustering, IEEE Trans., Syst., Man, and Cybern., Vol.SMC-16, No.3, pp.479-482, 1986.

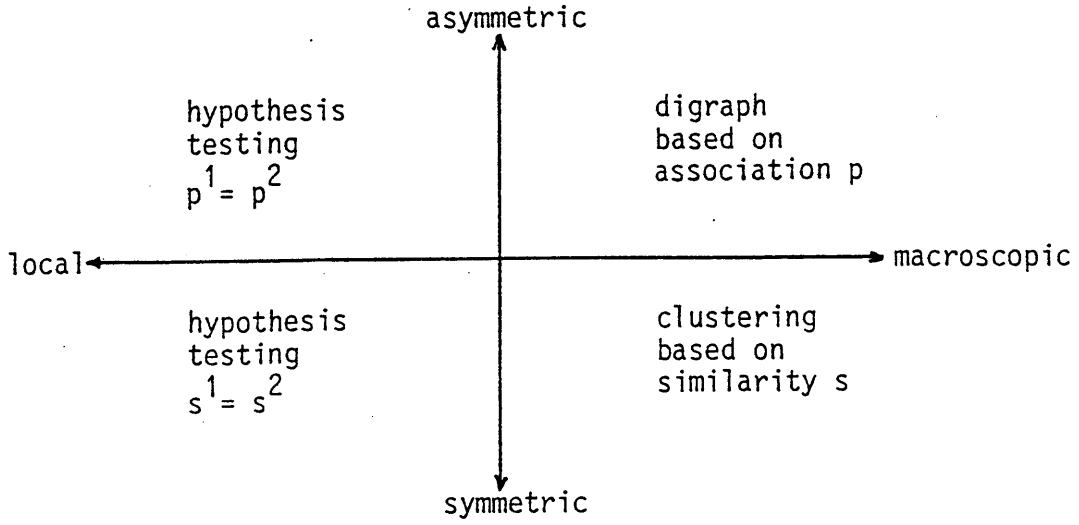


Fig. 1 A framework for structural representations of association: p^k, s^k, k=1,2, means parameters for two populations.

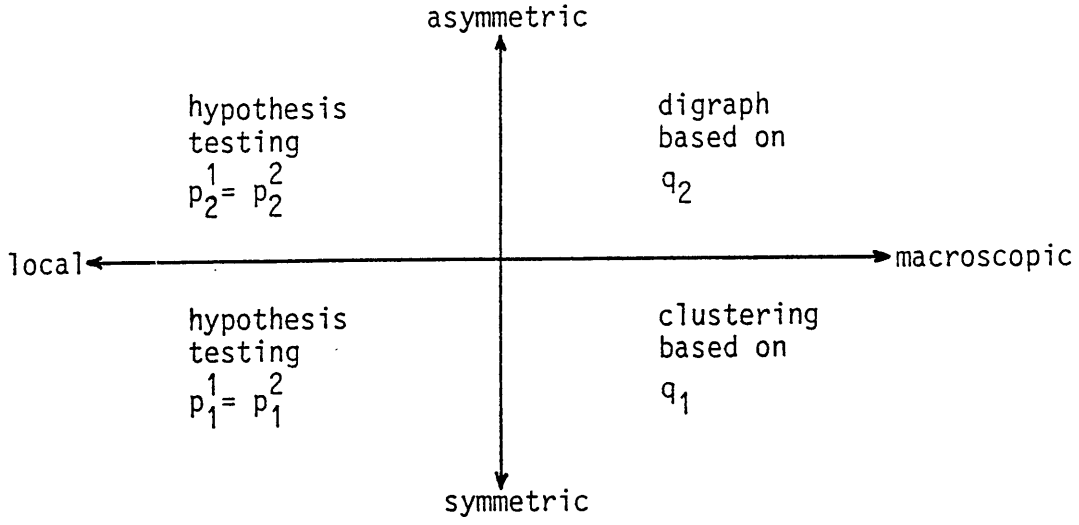


Fig. 2 The framework filled in by the method of neighborhood.

	class 1	class 2
population 1	O_{11}	O_{12}
population 2	O_{21}	O_{22}

Fig. 3 A 2 x 2 table.

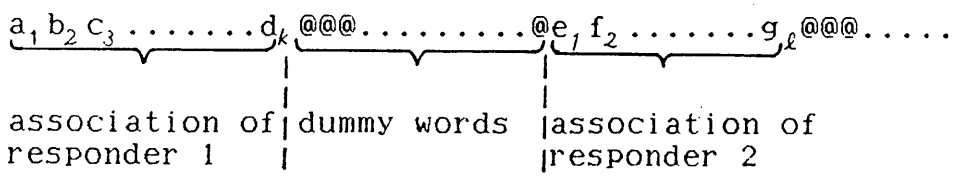


Fig. 4a Associations of responders and a sequence.

$a_1 - b_2 - c_3 - \ldots - d_k - @ - @ - @ - \ldots - @ - e_1 - f_2 - \ldots$

Fig. 4b A sequence as a network.

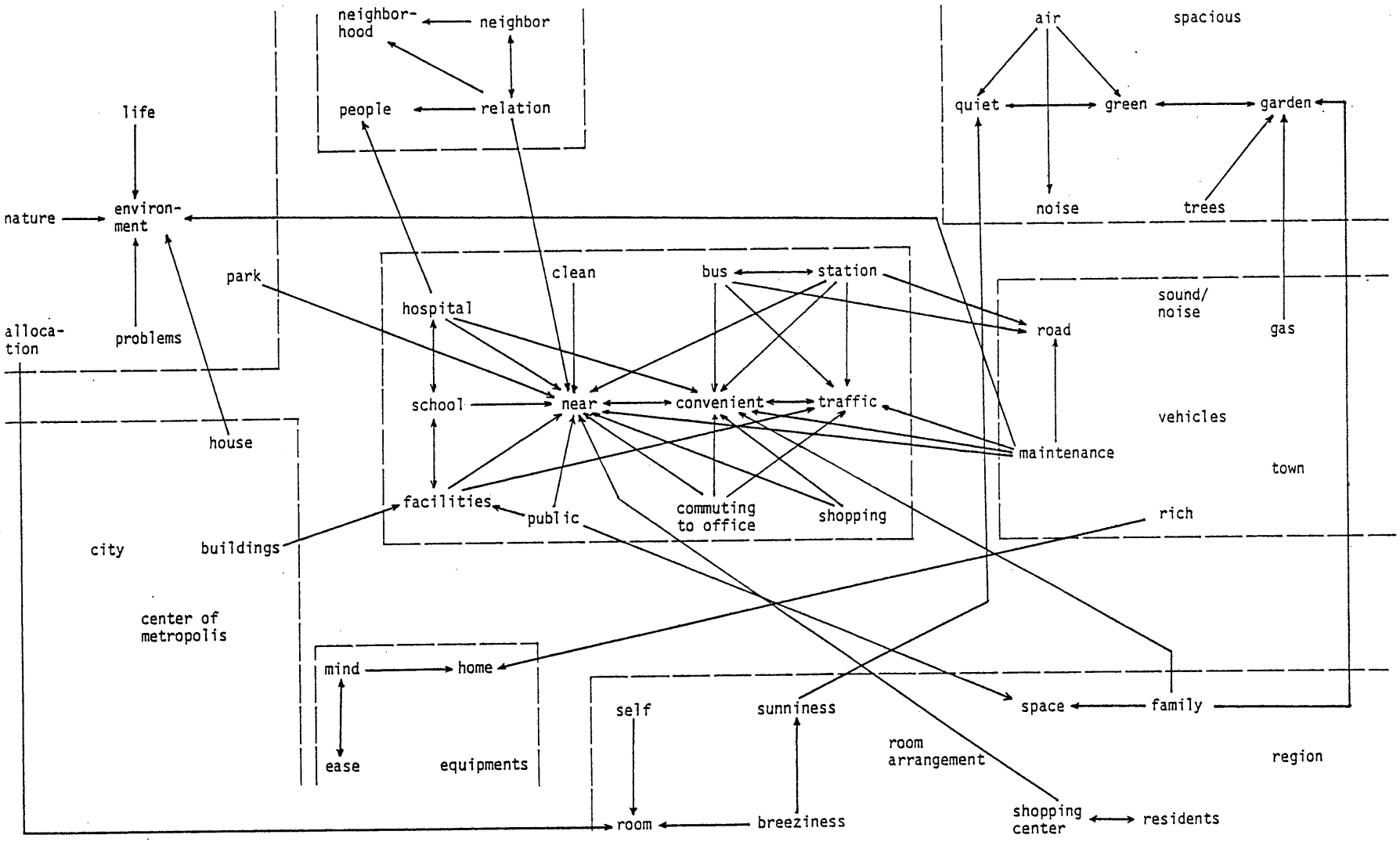


Fig. 5 Digraph with clusters obtained from SETAGAYA.
Threshold parameters for the edges are $(d,\beta) = (0.30,1.5)$.

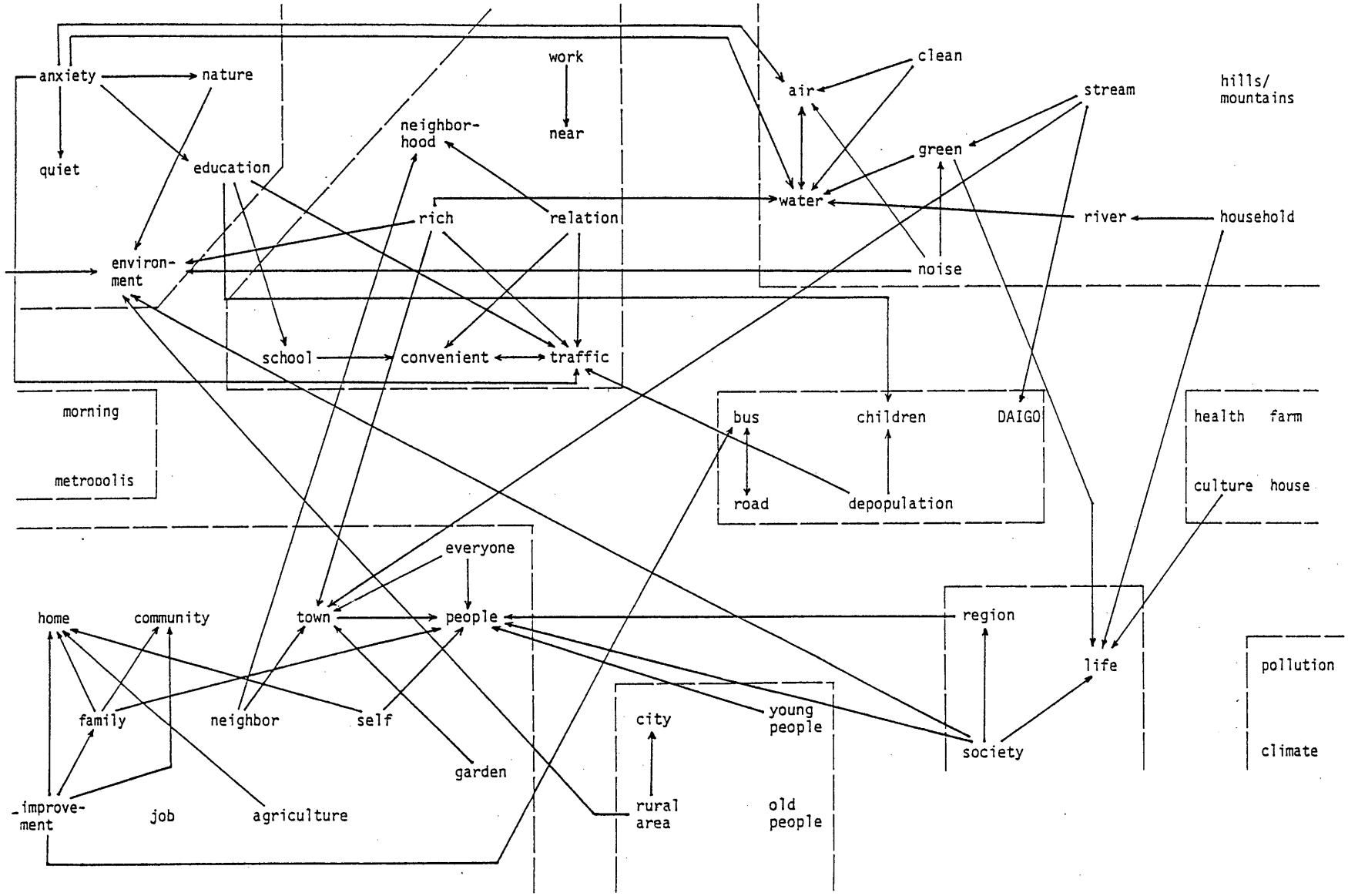


Fig. 6 Digraph with clusters obtained from DAIGO.
Threshold parameters for the edges are $(\alpha,\beta) = (0.30, 1.5)$.

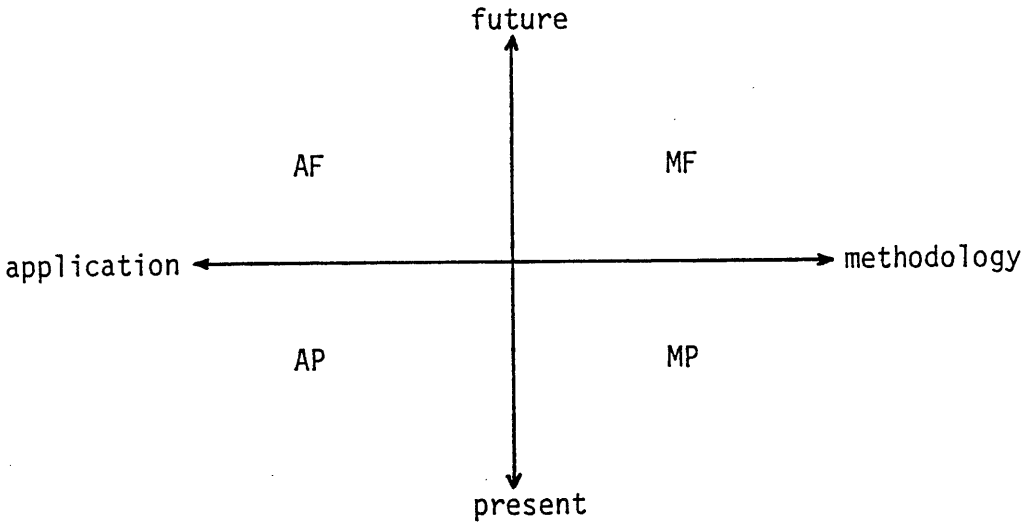


Fig. 7 A framework on significances of the neighborhood method.

INSTITUTE OF INFORMATION SCIENCES AND ELECTRONICS
UNIVERSITY OF TSUKUBA
SAKURA-MURA, NIIHARI-GUN, IBARAKI 305 JAPAN

REPORT DOCUMENTATION PAGE	REPORT NUMBER ISE-TR-88-67

TITLE

A Method of Neighborhood for Analyzing Free Association

AUTHOR(S)

Sadaaki Miyamoto
Shinsuke Suga
Ko Oi

REPORT DATE	NUMBER OF PAGES
June 3, 1988	25+figures
MAIN CATEGORY	CR CATEGORIES
INFORMATION SYSTEMS	H.1.2, G.3

KEY WORDS

cluster analysis, digraphs, algorithms, free association, survey by questionnaire

ABSTRACT

A method for construction of two types of measures of association between a pair of symbols that distribute over a network is developed. The method is called here a neighborhood method in the sense that the construction of the measures is based on neighborhoods of vertices of the network. A family of methods of structural representations is developed using the framework proposed formerly by the author. Moreover, a new method of hierarchical cluster analysis is developed. These methods are applied to structural representation of data on cognition of living environment of local residents. The data are obtained from a survey by questionnaire that requests free association about living environment. The result of data analysis shows how different structures of the cognition reflect different backgrounds of two districts and serves as a guideline for decision making about improvement of living environment.

SUPPLEMENTARY NOTES