文部省科学研究費補助金　環境科学特別研究(1) 60030012

ISE-TR-85-52

A METHOD OF NEIGHBORHOOD FOR CLUSTER ANALYSIS OF FREE ASSOCIATIONS IN INVESTIGATIONS OF COGNITIVE STRUCTURES

by

Sadaaki Miyamoto

Ko Oi

Osamu Abe

Atsuo Katsuya

Kazuhiko Nakayama

November 5, 1985

INSTITUTE
OF
INFORMATION SCIENCES AND ELECTRONICS

UNIVERSITY OF TSUKUBA

A METHOD OF NEIGHBORHOOD FOR CLUSTER ANALYSIS OF FREE ASSOCIATIONS IN INVESTIGATIONS OF COGNITIVE STRUCTURES

by

Sadaaki Miyamoto

Ko Oi

Osamu Abe

Atsuo Katsuya

Kazuhiko Nakayama

A Method of Neighborhood for Cluster Analysis of Free Associations in Investigations of Cognitive Structures

S. Miyamoto
Institute of Information Sciences and Electronics
University of Tsukuba, Sakura, Ibaraki 305, Japan

K. Oi
Environmental Information Division
The National Institute for Environmental Studies
Yatabe, Ibaraki 305, Japan

O. Abe
Science Information Processing Center
University of Tsukuba, Sakura, Ibaraki 305, Japan

A. Katsuya
Kyoto Sangyo University, Kyoto 603, Japan

K. Nakayama
Institute of Information Sciences and Electronics
University of Tsukuba, Sakura, Ibaraki 305, Japan

ABSTRACT

The aim of the present paper is to develop a method of generating similarity measures for cluster analysis of the data of free associations in psychological experiments. The free associations are regarded as a sequence of various words. A neighborhood of a word is defined to be a subset of words which occur near to the former word in the associations. Three algorithms of generating similarities between a pair of words based on the presence in the neighborhood are introduced, two of which lead to the same similarity measures when they are symmetrized. Properties of these similarity measures are discussed and normalizations of them are considered. Moreover a method with weighting coefficients, nonsymmetric neighborhoods, and several methods defined on networks and Euclidean spaces are developed. The neighborhood method is applied for analyzing data of free associations obtained by a survey by questionnaire on the living conditions and environment of local residents.

1. Introduction

Various kinds of psychological associations have been studied in different fields of sciences and engineering[1]: In particular, we find application of the associations in the field of social survey as a mean to have an idea of environmental cognition of people [2].

In a foregoing study [2] the authors suggested the use of a free association in a survey to grasp cognitive structure of local residents on his living condition. with the purpose to have macroscopic views for decision makings on improvement of urban environment. In this paper we present a new method which is based on a notion of neighborhood for cluster analysis of the free associations. The method of neighborhood is used to define similarity measures between a pair of words in the free association. Several similarity measures appropriate for this purpose are introduced and their properties are discussed.

2. Neighborhood method for generating similarities

A. Proximity measures and similarity measures

The method here is introduced in a general framework so that it can be used in many applications.

Let $W=\{a,b,c,\ldots\}$ be a finite set of words: we need not specify the number of elements in W. A pair of generic elements in W is denoted by a and b. A text T is a finite sequence of words in W. A word in W occurs in general many times in T. therefore an occurrence of the word a is denoted as a_i with subscript. For example. $T_0 = a_i\, b_j\, c_k\, a_\ell \ldots z_p$. In application data

of the free association is considered to be the text defined above. Sometimes an element in T is denoted by x without subscript when it does not correspond to a specific word in W. Moreover the symbol a may represent twice or more occurrences of a in a text. The set of all occurrences in T is denoted as |T|. For the above T_0. $|T_0|=\{a_i, a_\ell, b_j, c_k, \ldots, z_p\}$. For arbitrary subset $S \subset |T|$, m(S) means the number of elements in S. Moreover the sets of all the occurrences of the words a,b,... in T are denoted as $A=\{a_i, a_j, \ldots\}$, $B=\{b_k, b_\ell \ldots\}$, and so on.

A distance d(x,y) between a pair of occurrences is defined as follows:

$$d(x,y) = \{\text{the number of occurrences of words in T between } x \text{ and } y\} + 1 .$$

Therefore in the above T_0, $d(c_k, b_j)=1$, $d(a_i, a_k)=3$. Then a neighborhood $U_n(x)$ is defined for an occurrence:

$$U_n(x) = \{ y \mid d(x,y) \leq n, \quad y \in |T| \} .$$

For the above T_0, $U_1(b_j)=\{a_i, b_j, c_k\}$.

In a clustering we need a similarity measure s(a,b) defined on W x W which shows degree of relatedness between a and b. Note that a similarity should be symmetric:

$$s(a,b) = s(b,a) .$$

On the other hand, we consider nonsymmetric measures of representing relations of a pair of words directly obtained from the text by using the neighborhood. For the latter measures we use a word proximity p(a,b) instead of similarity. Therefore p(a,b) is nosymmetric in general:

$$p(a,b) \neq p(b,a) .$$

A purpose in this section is to derive similarity measures from proximity measures by a symmetrization and a normalization procedure. First we introduce three algorithms of defining proximity measures based on the neighborhood.

(i) (presence or absence)

Scan the text from the left. For each occurrence a_i of a, examine the neighborhood $U_n(a_i)$. If an occurrence of b is found, then let $p_1(a_i,b)=1$ (two or more occurrences are also reduced to $p_1(a_i,b)=1$). Then a proximity $p_i(a,b)$ is given by

$$p_i(a,b) = \sum_{\text{all } a_i \in A} p_1(a_i,b)$$

where the summation is taken over all the occurrences of a.

(ii) (simple counting)

Scan the text T from the left. For each occurrence a_i of a, Let $p_2(a_i,b)=\{$number of occurrences of b in $U_n(a_i)\}$. Then a proximity $p_2(a,b)$ is defined:

$$p_2(a,b) = \sum_{\text{all } a_i \in A} p_2(a_i,b) \quad .$$

(iii) (count and mark)

Scan the text T from the left. For each occurrence a_i of a, count the number of occurrences of b in $U_n(a_i)$. Let the number be $p_{3\ell}(a_i,b)$. An occurrence b_j once counted is marked so that it will not be doubly counted thereafter. Then a proximity $p_{3\ell}(a,b)$ is defined:

$$p_{3\ell}(a,b) = \sum_{\text{all } a_i \in A} p_{3\ell}(a_i,b) \quad .$$

That is, $p_{3\ell}(a_i,b) = m(\{b_j \mid b_j \in U_n(a_i),\ b \notin U_n(a_k)$ for all a_k's which are left side of a_i in T$\}$).

Moreover scan T from the right. According to the same counting procedure, we have another proximity

$$p_{3r}(a,b) = \sum_{\text{all } a_i \in A} p_{3r}(a_i,b) \quad .$$

Note that in algorithm (i) and (ii), the scanning of T from the left that from the right make no difference. Therefore we need not define $p_{1\ell}$, p_{1r}, $p_{2\ell}$, and p_{2r}.

Remark The method of neighborhood is based on the idea that a pair of words which frequently occur nearby each other is

considered to be related. A simple way to realize this idea without the neighborhood is to use the distance d(a,b) as the similarity. However, the neighborhood method is more appropriate, since a pair of words (c,d) which has a long distance should have a similarity s(c,d)=0. Moreover a similarity based on the distance can be considered within the present framework as a method with weighting coefficients which is a generalization of p_2 in (ii).

First proposition implies that the algorithm (iii) defines an appropriate proximity measure p_3.

Prop. 1 $\quad p_{3r}(a,b) = p_{3\ell}(a,b)$.

(Proof) Let

$$d(b_j,A) = \min_{\text{all } a_i \in A} d(b_j,a_i)$$

and $B_1 = \{ b_j \mid d(b_j,A) \leq n \ , \ b_j \in |T| \}$. Also let us give an edge from a_i to b_j ($a_i \rightarrow b_j$), when b_j is counted as a member in $U_n(a_i)$ by algorithm (iii). Then for each $b_j \in B_1$, b_j has one and only one edge from some a_i by the procedure of forming $p_{3\ell}$. For $b_j \notin B$, b_j has no edge by the same procedure. In the same way, $b_j \in B_1$ (resp. $b_j \notin B_1$) has one edge (resp. no edge) by the procedure of forming p_{3r}. Therefore we have as the total number of the edges:

$$p_{3\ell}(a,b) = p_{3r}(a,b) = m(B_1). \qquad \square$$

Namely we need not distinguish $p_{3\ell}$ and p_{3r}. Therefore we define

$$p_3(a,b) = p_{3\ell}(a,b) = p_{3r}(a,b) \ .$$

Next proposition shows that p_2 is symmetric.

Prop. 2 $\quad p_2(a,b) = p_2(b,a)$.

(Proof) Let us give an edge from a_i to b_j when b_j is counted as a member in $U_n(a_i)$ by the algorithm (ii) in the calculation of $p_2(a,b)$. Also give an edge from b_p to a_q in the same way when $p_2(b,a)$ is calculated. It is clear that $b_j \in U_n(a_i)$ means $a_i \in U_n(b_j)$ and vice versa. Therefore for any pair of occurrences we have only two cases: (a) $a_i \rightarrow b_j$ and $b_j \rightarrow a_i$; (b) a_i and b_j have no edges between them. $p_2(a,b)$ and $p_2(b,a)$ are the number of edges in the respective procedures defined above, which means the required relation. []

It is clear that the proximities p_1 and p_3 are not symmetric. Consider the following example:

Example 1 Let $T_1 = a_1 c_1 b_1 a_2$ and n=2. Then

$$p_1(a,b) = 2,\ p_1(b,a) = 1,\ p_3(a,b) = 1,\ p_3(b,a) = 2.$$

These two measures p_1 and p_3 are proved to be identical when they are symmetrized.

Prop. 3 $\quad p_1(a,b) = p_3(b,a)$.

(Proof) Let $A_1 = \{ a_i \mid d(a_i, B) \leq n,\ a_i \in |T| \}$ as in the proof of Prop. 1, where B is the set of all the occurrences of b in T. As was shown in the proof of Prop. 1, $p_3(b,a) = m(A_1)$. On the other hand, if one connects $a_i \in A$ with some $b_j \in U_n(a_i)$ by an edge by the algorithm (i), each $a_i \in A_1$ has an edge and $a_i \notin A_1$ has no edges. It is clear that the number of the edges in (i) is equal to $p_1(a,b)$,

we have

$$p_1(a,b) = p_3(b,a) = m(A_1). \qquad []$$

Experiences show that for the purpose of the clustering a similarity $s(a,b)$ should be normalized in the sense that the total number of the word occurrences $m(A)$ does not have direct influences on the increase of the similarity. We define normalized similarities from the three proximities. Especially, $s(a,b)$ derived from p_1 (and p_3) has a correlation-like property.

Let

$$s_1^{(\alpha)}(a,b) = \frac{c^{(\alpha)}(a,b)}{m(A)+m(B)-c^{(\alpha)}(a,b)} \quad ,$$

where

$$c^{(\alpha)}(a,b) = \alpha \max(p_1(a,b),p_1(b,a)) + (1-\alpha)\min(p_1(a,b),p_1(b,a)).$$

Prop.3 means that

$$c^{(\alpha)}(a,b) = \alpha \max(p_3(a,b),p_3(b,a)) + (1-\alpha)\min(p_3(a,b),p_3(b,a)).$$

The similarity $s_1^{(0)}$ (resp. $s_1^{(1)}$) is based on the minimum (resp. the maximum) of the two quantities $p_1(a,b)$ and $p_1(b,a)$; $s_1^{(1/2)}$ is based on the arithmetic mean of the two quantities.

Prop. 4 The similarity $s_1^{(\alpha)}(a,b)$ satisfies the following.

(a) For $0 \leq \alpha \leq 1/2$, the relation $0 \leq s_1^{(\alpha)}(a,b) \leq 1$ holds.

(b) For $1/2 < \alpha \leq 1$, the relation $0 \leq s_1^{(\alpha)}(a,b)$ is valid, but $s_1^{(\alpha)}(a,b) \leq 1$ does not hold in general.

(c) For $0 \leq \alpha \leq 1$, $s_1^{(\alpha)}(a,b) = 0$ iff for any $a_i \in A$ and $b_j \in B$,

$d(a_i, b_j) > n$. (That is, a and b are isolated each other.)

(d) For $0 \leq \alpha < 1/2$, $s_1^{(\alpha)}(a,b) = 1$ iff $m(A)=m(B)=m(A_1)=m(B_1)$.

(Proof) Since $c^{(\alpha)}(a,b) \leq \max(m(A),m(B))$ it is clear that $s_1^{(\alpha)}(a,b) \geq 0$ for $0 \leq \alpha \leq 1$. Next, from the identity

$$c^{(\alpha)}(a,b) + c^{(1-\alpha)}(a,b) = p_1(a,b) + p_1(b,a) = m(A_1) + m(B_1),$$

we have

$$s_1^{(\alpha)}(a,b) = \frac{c^{(\alpha)}(a,b)}{m(A-A_1)+m(B-B_1)+m(A_1)+m(B_1)-c^{(\alpha)}(a,b)}$$

$$= \frac{c^{(\alpha)}(a,b)}{m(A-A_1)+m(B-B_1)+c^{(1-\alpha)}(a,b)} \leq \frac{c^{(\alpha)}(a,b)}{c^{(1-\alpha)}(a,b)} .$$

Since $c^{(\alpha)}(a,b)$ is monotone nondecreasing relative to α, the last inequality implies (a). Moreover for $0 \leq \alpha < 1/2$, the above relation implies that $s_1^{(\alpha)}(a,b)=0$ iff $m(A-A_1)=0$, $m(B-B_1)=0$, and $c^{(\alpha)}(a,b) = c^{(1-\alpha)}(a,b)$. The latter three relations are equivalent to $m(A)=m(A_1)$, $m(B)=m(B_1)$, and $m(A_1)=m(B_1)$, which prove (d). For (b), let us consider Example 1. Then it is easily seen that $s_1^{(\alpha)}(a,b) = (2\alpha+(1-\alpha))/(2+1-(2\alpha+1-\alpha)) = (\alpha+1)/(2-\alpha) > 1$ for $1/2 < \alpha \leq 1$. Finally, $c^{(\alpha)}(a,b)=0$, $0 \leq \alpha \leq 1$ means that $\min(p_1(a,b),p_1(b,a))=0$. If $p_1(a,b)=m(A_1)=0$, then for any $a_i \in A$ and any $b_j \in B$. $d(a_i, b_j) > n$. That is, if either of $p_1(a,b)$ or $p_1(b,a)$ is equal to zero, then $p_1(a,b)=p_1(b,a)=0$. Conversely, if $d(a_i, b_j) > n$, for all a_i and b_j, then $m(A_1)=m(B_1)=0$, which means $c^{(\alpha)}(a,b)=0$. Namely, (c) is valid. []

The above proposition means that $s_1^{(\alpha)}$ for $0 \leq \alpha < 1/2$ has

the most desirable properties: among the three measures $s_1^{(0)}$, $s_1^{(1/2)}$, and $s_1^{(1)}$, the measure $s_1^{(0)}$ should be considered first of all.

On the other hand the above type of the normalization is not adequate for p_2. Therefore we define simply

$$s_2(a,b) = \frac{p_2(a,b)}{m(A) + m(B)} \quad .$$

The similarities $s_1^{(d)}$ and s_2 are dependent on the size n of the neighborhood U_n. When their dependence is explicitly shown by a superscript as $s_1^{(d)n}$ and s_2^n, we have

Prop. 5 $s_1^{(d)n}(a,b)$ and $s_2^n(a,b)$ are monotonically nondecreasing with respect to n.

(Proof) It is clear from the algorithms (i) and (ii) that $p_1(a,b)$ and $p_2(a,b)$ are monotonically nondecreasing with respect to n. Immediately it follows that $s_2(a,b)$ has the same property. On the other hand, monotonically nondecreasing property of $c^{(d)n}(a,b)$ and that of the function $f(x) = x/(k-x)$ for positive x and positive constant k mean that the same property is valid for $s_1^{(d)n}(a,b)$. []

A generlization of p_2 with weighting coefficients can be introduced. Let $w = (w(1), w(2), \ldots, w(n))$, be an n-vector of nonnegative components. Define

$$P_{2w}(a_i, b) = \sum_{\text{all } b_j \in U_n(a_i)} w(d(a_i, b_j))$$

according to the counting procedure in algorithm (ii) with the weight w, and put

$$P_{2w}(a,b) = \sum_{\text{all } a_i \in A} P_{2w}(a_i, b) \quad .$$

Note that the same type of weighting is not applicable in algorithms (i) and (iii). Namely, $P_{3w\ell}(a,b) \neq P_{3wr}(a,b)$ in general.

Prop. 6 $\quad P_{2w}(a,b) = P_{2w}(b,a)$.

(Proof) According to the process in the proof of Prop. 2, we give an edge from a_i to each occurrence b_j in $U_n(a_i)$, with weight $w(d(a_i, b_j))$. Then it is clear that $P_{2w}(a,b)$ is equal to the sum of weights on the edges defined above. In the same way we obtain $P_{2w}(b,a)$ as the sum of weights of the edges from b_p's to a_q's. As has been noted before, we have only two cases for a pair (a_i, b_j): (a) a and b have no edges between them; (b) $a_i \to b_j$ with the weight $w(d(a_i, b_j))$, $b_j \to a_i$ with the weight $w(d(b_j, a_i))$. Since $d(a_i, b_j) = d(b_j, a_i)$, we have $P_{2w}(a,b) = P_{2w}(b,a)$ []

B. Nonsymmetric neighborhood

Sometimes it appears that nonsymmetric neighborhood is more appropriate for the analysis of free associations, since a word b in the association seems to have more influence on a word

occurring after b than a word before b. Therefore we study some of the measures in nonsymmetric neighborhood, although the results in this subsection are weaker than those in the symmetric case.

A nonsymmetric neighborhood $U_{mn}(x)$ is defined to be a subset

$$U_{mn}(x) = \{ y \mid d(y,x) \leq m, \ y \text{ is left side of } x, \ y \in |T| \} \cup \{ y \mid d(x,y) \leq n, \ y \text{ is right side of } x, \ y \in |T| \}.$$

For example, in $T_0 = a_i b_j c_k a_\ell \ldots z_p$.

$U_{21}(c_k) = \{a_i, b_j, c_k, a_\ell\}$, $U_{01}(b_j) = \{b_j, c_k\}$.

According to U_{mn}, $p_1^{mn}(a,b)$, $p_2^{mn}(a,b)$, $p_{3\ell}^{mn}(a,b)$, and $p_{3r}^{mn}(a,b)$ are defined by the three algorithms (i), (ii), and (iii), respectively, where $U_{mn}(a_i)$ is used instead of $U_n(a_i)$.

Prop. 7 $\quad p_{3\ell}^{mn}(a,b) = p_{3r}^{mn}(a,b)$.

(Proof) Let

$$B_1^{nm} = \{ b_j \mid d(b_j, a_i) \leq n \text{ for some } a_i \in A \text{ which is left side of } b_j \} \cup \{ b_j \mid d(b_j, a_i) \leq m \text{ for some } a_i \in A \text{ which is right side of } b_j \}.$$

By the same way as in the proof of Prop. 1, we have

$$p_{3\ell}^{mn}(a,b) = p_{3r}^{mn}(a,b) = m(B_1^{nm}) \qquad []$$

Therefore we define

$$p_3^{mn}(a,b) = p_{3\ell}^{mn}(a,b) = p_{3r}^{mn}(a,b) .$$

Prop. 8 $\quad p_2^{mn}(a,b) = p_2^{nm}(b,a)$

$$p_3^{mn}(a,b) = p_1^{nm}(b,a) .$$

(Proof) Let us give an edge from a_i to each b_j in $U_{mn}(a_i)$ and from b_p to a_q in $U_{nm}(b_p)$ as in the proof of Prop. 2. It is clear that $b_j \in U_{mn}(a_i)$ means $a_i \in U_{nm}(b_j)$ and vice versa. Therefore we have the first identity. For the second identity, it is sufficient to note that $p_1^{mn}(b,a) = m(B_1^{mn})$. []

Note that $p_2^{mn}(a,b) \neq p_2^{mn}(b,a)$ when $m \neq n$. Consider $T_1 = a_1 c_1 b_1 a_2$, then $p_2^{01}(a,b) = 0$, $p_2^{01}(b,a) = 1$. Therefore symmetrization of p is necessary to define similarities.

Let

$$c_k^{(\alpha)mn}(a,b) = \max(p_k^{mn}(a,b), p_k^{mn}(b,a)) + (1-\alpha)\min(p_k^{mn}(a,b), p_k^{mn}(b,a)), \qquad k=1,2,3,$$

and

$$s_1^{(\alpha)mn}(a,b) = \frac{c_1^{(\alpha)mn}(a,b)}{m(A)+m(B)-c_1^{(\alpha)mn}(a,b)} = s_3^{(\alpha)nm}(a,b) ,$$

$$s_2^{(\alpha)mn}(a,b) = \frac{c_2^{(\alpha)mn}(a,b)}{m(A)+m(B)} .$$

Prop. 9 $$p_2^{0n}(a,b) + p_2^{0n}(b,a) = p_2^{nn}(a,b) \ (= p_2(a,b)).$$
$$2s_2^{(1/2)0n}(a,b) = s_2^{(1/2)nn}(a,b) \ (= s_2(a,b)).$$

(Proof) Let us give undirected edges in the three counting procedures of forming $p_2^{0n}(a,b)$, $p_2^{0n}(b,a)$, and $p_2^{nn}(a,b)$. Denote the set of edges in each procedure as $E^{0n}(a,b)$, $E^{0n}(b,a)$, and $E^{nn}(a,b)$, respectively. It is clear that $E^{0n}(a,b) \cap E^{0n}(b,a) = \emptyset$, since in $E^{0n}(a,b)$ the connected pair has a_i as the left element and in $E^{0n}(b,a)$ the pair has b_j as the left element. Moreover

$E^{0n}(a,b) \cup E^{0n}(b,a) = E^{nn}(a,b)$, since $b_j \in U_{nn}(a_i)$ iff $b_j \in U_{0n}(a_i)$ or $a_i \in U_{0n}(b_j)$. Counting the numbers of elements in the three sets of the edges, we have the former identity. The latter identity directly follows from the former. []

Prop. 10 $\quad p_k^{mn}(a,b) \leq p_k^{m'n'}(a,b), \quad m \leq m', \; n \leq n',$
$k = 1, 2, 3$.

The proof of Prop. 10 is obvious and omitted.

Remark The similarity $s_1^{(0)0n}(a,b)$ is not recommended. Consider an example T=ab.....ab..... . Then $c_1^{(0)}(a,b) = c_1^{(0)nn}(a,b) = 2$, $n \geq 1$, whereas $c_1^{(0)0n}(a,b) = 0$, which means $s_1^{(0)0n}(a,b) = 0$. The measure $s_1^{(0)0n}(a,b)$ is inappropriate.

C. Neighborhoods in networks and in Euclidean spaces

Although the present method is applied solely to the analysis of the free association here, the neighborhood techniques is applicable to the analysis of many other experiments. In this subsection the method is considered on networks and Euclidean spaces. Some of the applications will be suggested below.

Let (V,E) be an undirected graph whose vertices are occurrences of elements in W. For example $V = \{a_i, b_j, c_k, \ldots, a_\ell, \ldots\}$. The set of all the occurrences in V of a word a is denoted as A. Distance d(x,y) is defined to be the minimum number of edges in the edge sequences which connect x and y. Then $U_n(x)$ is defined:

$$U_n(x) = \{\, y \mid d(x,y) \leq n, \quad y \in V \,\}.$$

The algorithms (i) and (ii) in the subsection 2.A is immediately applied to have $p_1(a,b)$ and $p_2(a,b)$ defined on (V,E), whereas the algorithm (iii) depends on the ordering of occurrences $\{a_i\} \subset V$ of $a \in W$. If we denote an ordering of A as $A'=(a_i, a_j, \ldots)$ and the resulting proximity p_3 according to the ordering as $p_{3A'}$, we have

Prop. 11 For any ordering A' of A.

$$p_{3A'}(a,b) = \sum_{a_i \in A'} p(a_i, b)$$

remains constant. In other words, $p_3(a,b)$ on (V,E) is well-defined.

(Proof) It is sufficient to note that $p_{3A'}(a,b) = m(B_1)$, for any ordering A', where $B_1 = \{\, b_j \mid d(b_j, A) \leq n, \quad b_j \in V \,\}$. []

The proximities $p_1(a,b)$, $p_2(a,b)$ and $p_3(a,b)$ defined in this subsection has the same properties as those in the subsection 2.A. Therefore Props.2~5 are valid for the three proximities in this subsection. Their proofs are also applicable with no essential modification. Therefore we omit the duplicate statements of these facts.

Furthermore, the neighborhood consideration is applicable on Euclidean spaces. Let V be a finite set of points in a Euclidean space R^N. Each point in V is labeled as an occurrence of a word in W. Again, $V = \{a_i, b_j, c_k, \ldots\}$. The distance $d(x,y)$, $x,y \in R^N$ is

defined to be an usual Euclidean distance and the neighborhood $U_\eta(x)$ is defined as a closed spere in R^N with the center x and the radius n. Then it is immediate to have $p_1(a,b)$, $p_2(a,b)$, and $p_3(a,b)$ according to the previous discussions. Props. 2~5 hold also in this case. When the neighborhood is generalized to an arbitrary closed set of a fixed shape containing x, the same results do not hold in general. Let U(0) be an arbitrary closed subset containing the origin. Then define $U(x) = \{ y \mid y-x \in U(0), \ y \in R^N \}$. The neighborhood U(x) is called symmetric if U(0) is symmetric about the origin. Then we have

Prop. 12 If $U(a_i)$ is symmetric, then $p_2(a,b) = p_2(b,a)$ and $p_1(a,b) = p_3(b,a)$. On the other hand, if $U(a_i)$ is nonsymmetric, then the above two identity does not hold in general.

(Proof) The first part is proved in a similar manner as in the proofs of Props. 2,3, and is omitted here. For the second part let us note that for a nonsymmetric neighborhood U(0) there exists a point $x \in U(0)$ but $-x \notin U(0)$. Let W={x,0} and V={x,0}. Then $p_2(0,x)=1$, $p_2(x,0)=0$, $p_1(0,x)=1$, $p_3(x,0)=0$. []

Remark An application of the methods in this subsection is analysis of drawings by subjects. In an experiment subjects are asked to write down their cognitive structures of a certain notion as a network. Then the neighborhood method on the network will be applied to aggregate individual structures into a structure of the whole group. Another experiment requires subjects to make a configuration of objects according to their

cognitive structures. Then the method on Euclidean space is applicable.

3. Application to the free association

Here the data of a free association is based on a survey by questionnaire on environmental cognition of local residents [2]. Subjects were asked to write freely what they associated with the notions "the easiness of living" and "happiness in living". In this paper 32 selected answers of longer responses which are selected from the whole 600 answers are analyzed. The 32 answers are considered as one text, since a purpose here is to aggregate structures in these answers. Each answer of a responder in the text was separated from others by a large number of blanks so that the neighborhood does not connect word occurrences of two responders.

Figures 1, 2, and 3 show the results of hierarchical agglomerative clusterings by the group average method [3] based on $s_1^{(0)}(a,b)$, $s_2(a,b)$, and $s_1^{(1/2)mn}(a,b)$, respectively. The size of the neighborhood is n=5 for $s_1^{(0)}$ and s_2; it is m=0, n=5 for $s_1^{(1/2)mn}$. The item names in these dendrograms are 56 elements in W which have been associated more than four times. Although several words seem to be synonymous, they should be distinguished at early stages of the data analysis. Otherwise significant implications may be lost.

Remark The original questionnaire and responses are written in Japanese. The names in the dendrograms are translations by the authors.

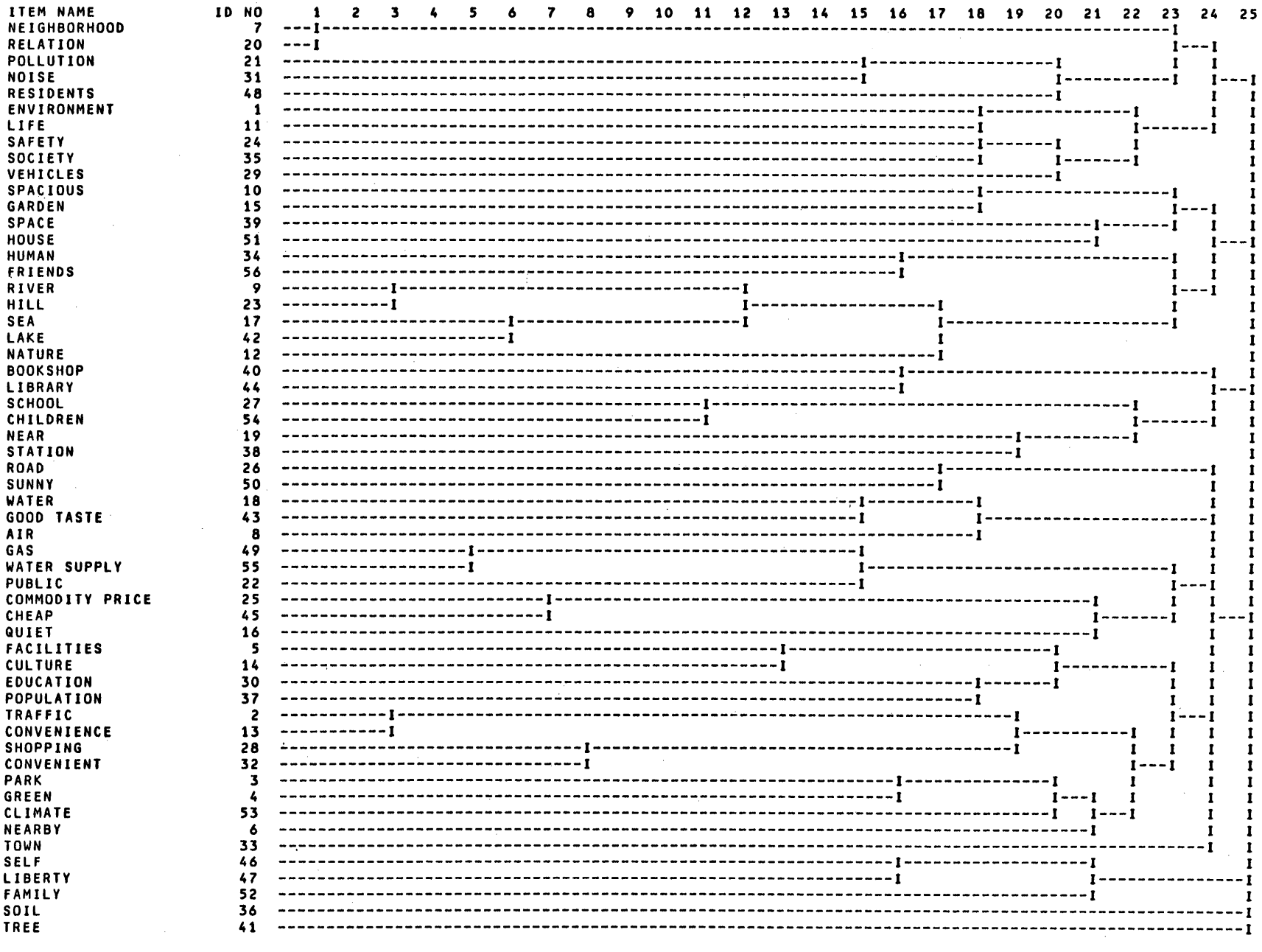

Fig. 1 A dendrogram by the group average method based on $s_1^{(0)}$. The entries are words obtained by free associations on living conditions.

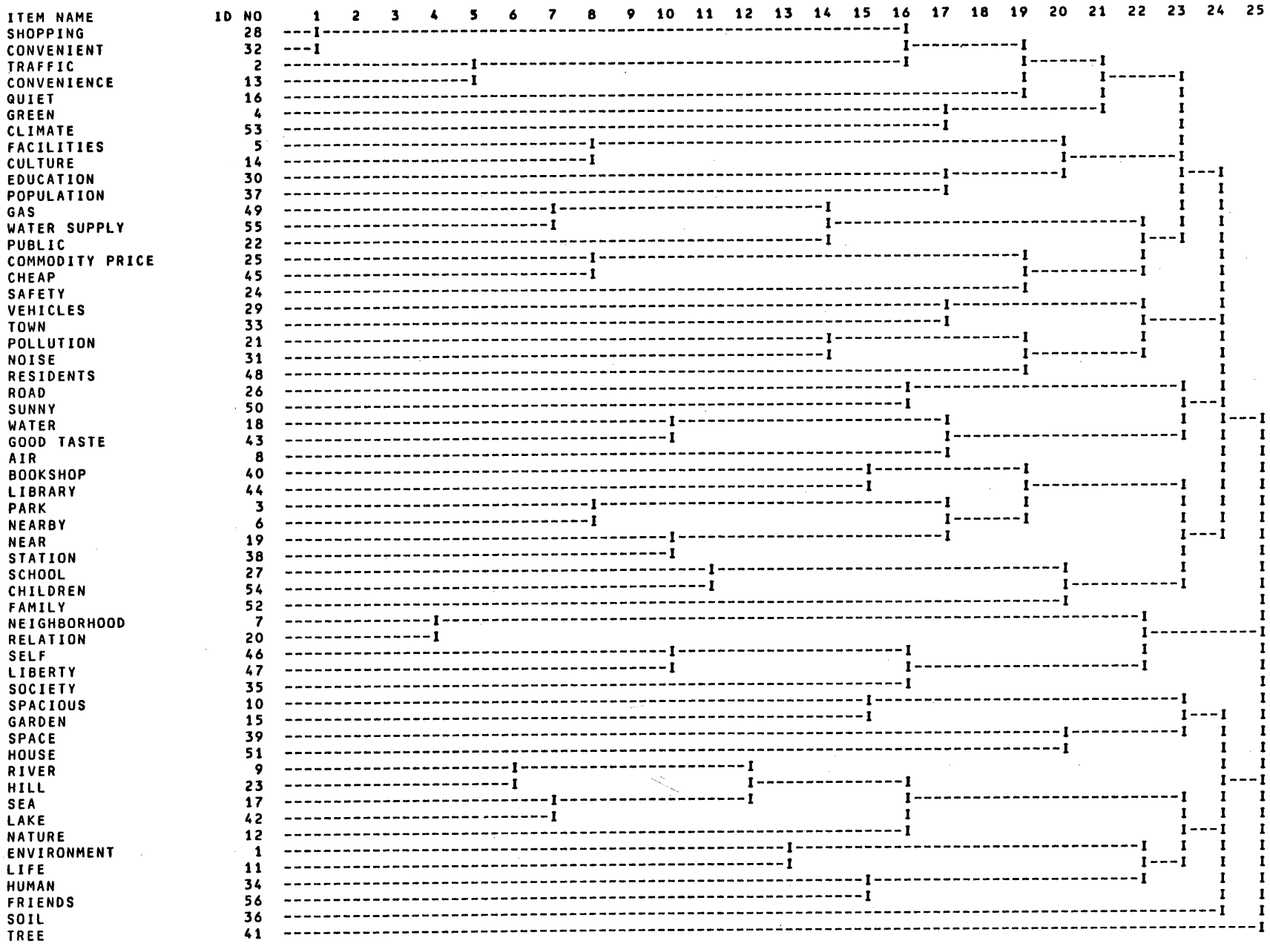

Fig. 2 A dendrogram by the group average method based on s_2. The entries are words obtained by free associations on living conditions.

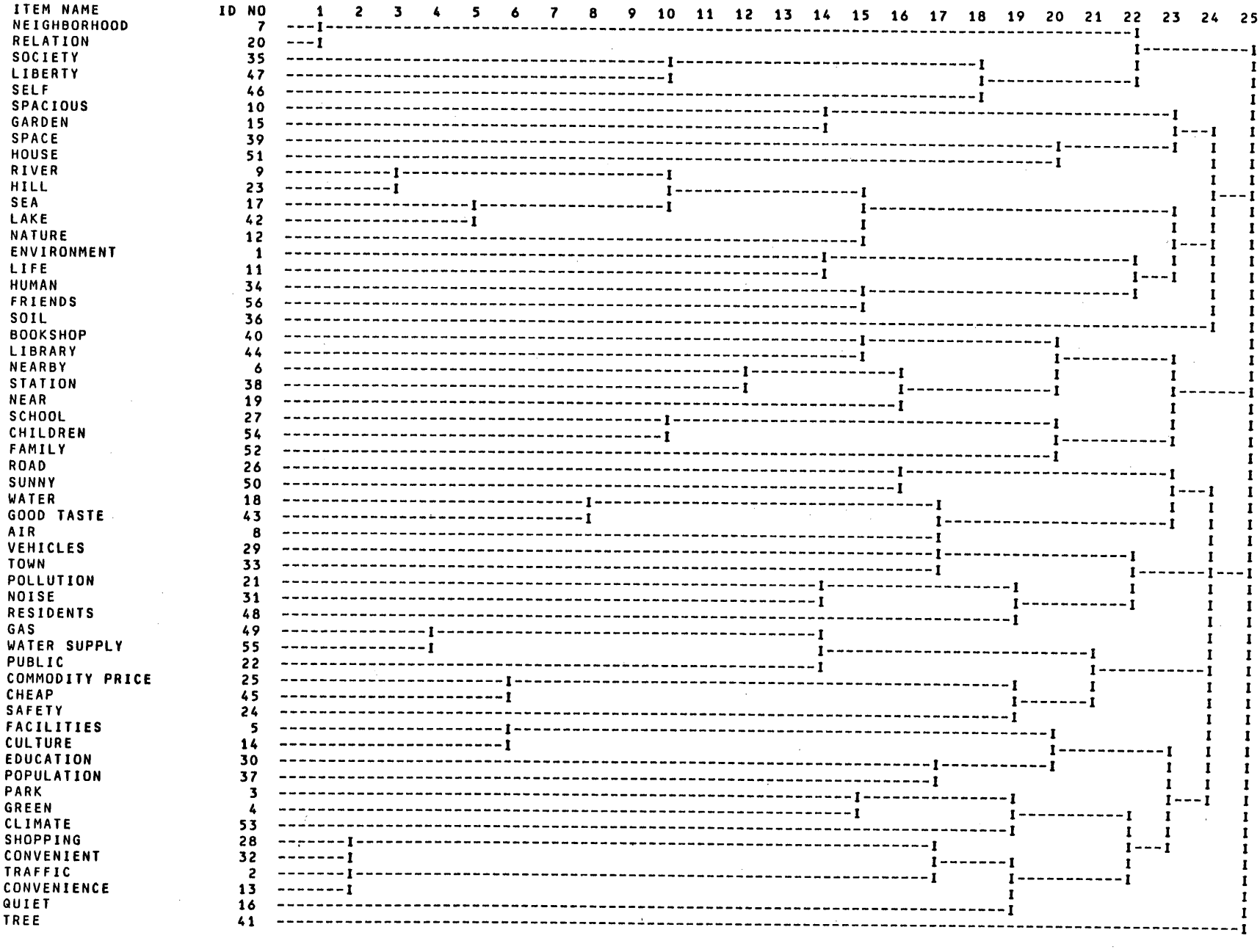

Fig. 3 A dendrogram by the group average method based on $s_1^{(1/2)05}$. The entries are words obtained by free associations on living conditions.

The three dendrograms show similar categories as a macroscopic coginitive structure of the subjects. We find several categories:

(a) "nature" which contains 'river', 'hill', 'lake', 'sea', etc;

(b) "house" which contains 'space', 'garden', 'house', etc;

(c) "human relation" which contains 'neighborhood', 'relation', 'liberty', etc;

(d) "convenience" which contains 'convenient', 'shopping', 'traffic', etc;

(e) "culture" which contains 'bookshop', 'library', 'school', 'near', 'children';

and so on.

They are considered to be important categories in cognitive structures on living environment of these subjects.

Remark Changes on the size n of the neighborhood have been investigated but their results are omitted here. No great changes on the categories have been observed. It appears that smaller values of the size n produces a more definite results. Actually, the value n=3 is rather too small in this example, since some of the definite structures are lost. For example, related words such as 'library', 'bookshop', 'school', 'near' which forms a group in the above three figures do not form a cluster in the latter case when n=3.

4. Conclusions

The neighborhood method considered here is more appropriate than a method based on distance for generating similarities for cluster analysis. A multidimensional scaling should be studied based on the present method; the application is immediate.

Another application of the method here is the analysis of documents. When the documents have loose structures to which syntactic approaches are inappropriate, or if the amount of data is very large, the present method is useful. Structure on Chinese classical literature which is very symbolic has been already studied by the present method [4].

Moreover the neighborhood method defined on networks and on Euclidean spaces suggests a new kind of experiments in the study of cognitive structures by providing a new technique of the data analysis.

ACKNOWLEDMENT

This research was partially supported by the Grant in Aid for Scientific Research of the Educational Ministry in fiscal 60030012.

References

1. J. R. Anderson. Cognitive Psychology and Its Implications. San Francisco. W.H.Freeman and Co., 1980.

2. 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.. to appear.

3. M. R. Anderberg. Cluster Analysis for Applications. New York. Academic. 1973.

4. K. Matsumoto. S. Miyamoto. K. Nakayama. and S. Hoshino. "Statistical analysis of Chinese classical text: structure of keywords in commentaries of Huang-ti Yin-fu-ching." Library and Information Science. No.22. 1-10. 1984 (in Japanese).

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

REPORT DOCUMENTATION PAGE	REPORT NUMBER ISE-TR-85-52

TITLE

A method of neighborhood for cluster analysis of free associations in investigations of cognitive structures

AUTHOR(S)

Sadaaki Miyamoto (Institute of Information Sciences and Electronics)
Ko Oi (The National Institute for Environmental Studies)

Osamu Abe (Science Information Processing Center, University of Tsukuba)
Atsuo Katsuya (Kyoto Sangyo University)

Kazuhiko Nakayama (Institute of Information Sciences and Electronics)

REPORT DATE	NUMBER OF PAGES
November 5, 1985	24
MAIN CATEGORY	CR CATEGORIES
Psychology	3.36, 3.71, 3.63

KEY WORDS

cluster analysis, free associations, neighborhood, cognitive structures, living environment

ABSTRACT

The aim of the present paper is to develop a method of generating similarity measures for cluster analysis of the data of free associations in psychological experiments. The free associations are regarded as a sequence of various words. A neighborhood of a word is defined to be a subset of words which occur near to the former word in the associations. Three algorithms of generating similarities between a pair of words based on the presence in the neighborhood are introduced. Properties of these similarity measures are discussed and normalizations of them are considered. Moreover a method with weighting coefficients, nonsymmetric neighborhoods, and several methods defined on networks and Euclidean spaces are developed. The neighborhood method is applied for analyzing data of free associations obtained by a survey by questionnaire on the living conditions and environment of local residents.

SUPPLEMENTARY NOTES