AN OPTIMAL COMBINATION METHOD OF RELATIONS AMONG SCIENTIFIC ARTICLES

Saito, Tatsuki
Department of Precision Engineering, Faculty of Engineering, Hokkaido University

https://doi.org/10.5109/13412

出版情報：Bulletin of informatics and cybernetics. 24 (1/2), pp.81-92, 1990-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 24, No. 1 ~ 2, 1990

AN OPTIMAL COMBINATION METHOD OF RELATIONS AMONG SCIENTIFIC ARTICLES

By

Tatsuki SAITO*

Abstract

An optimal combination method is proposed to unify similarity matrices generated from relational graphs having various relations among scientific articles. Three relations are considered, that is, the citation relation, the key word (or title) relation and the author relation. Canonical correlation analysis and alpha factor analysis are applied to determine the optimal combination coefficient. Maximal consistency and maximal generalizability are adopted as the criteria of optimization. The unified similarity matrix is analyzed by combinatorial cluster analysis. The optimal consistency model and its exploratory result are described.

1. Introduction

The information obtained from scientific articles often gives a researcher new motivation. He retrieves articles relevant to his theme of study and can thus obtain valuable ideas. Considering scientific articles as a source of intellectual production, it is necessary to take note of various relations among articles and to investigate their properties.

A fundamental methodology is proposed for modeling a scientific article-information system. Methods of canonical correlation analysis and alpha-factor analysis are applied, in order to optimally model a relational structure among articles. The bibliographical items include the primary and epitomized information. That is, article title, author name, affiliation, journal name, volume, number, pages, published time, publisher and key words are fundamental to identify a scientific article. It is obvious that these items are important for clarifying the relations among articles. The information about references is also important [1, 2]. Bibliographical items are usual in the conventional retrieval system [3–6]. However, it should be recognized that the simple retrieval of bibliographical items alone is not enough to model the structure of the scientific article-space with various relations and analyze these properties. Consequently, a modeling technique is required, and for this graph theory is applicable [7–10]. Usually, the graph used for modeling is nondirected, because of its ease of processing [11–14]. There exists a directionality in the relations between a citing article and a cited article. In this paper, a directed-value graph with the strength of relation

* Department of Precision Engineering, Faculty of Engineering, Hokkaido University, Japan.

and asymmetrical similarity matrices generated from various relations among articles are used for the construction of a precise model. Though a similarity matrix can be analyzed by focusing on a particular relation, more significant results can be obtained by using the total relation among articles. Seven combinatorial cluster analyses are applied to such similarity matrices for numerical classification.

2. Modeling by Using a Relational Graph

2.1. Relational graph

The graph with p points and q lines is called a (p, q) graph, where p points belong to a non-empty finite set P and unordered q pairs belong to specified set L. Pair $l = (u, v)$ of points u and v belongs to set L, where u and v are arbitrary elements of set P, and is called a line of the graph. The graph is called also a graph G = (P; L). There are two types of graphs, directed and nondirected. In a directed graph, the non-empty finite point-set P and the specified set L with the ordered pairs of two different points are dealt with simultaneously. Set S of line (u, v) and set T of line (v, u) are equivalent to a nondirected graph when line (u, v) equals line (v, u). Scientific articles written by the same author are connected by lines. The author relation has no direction, but the citation relation has direction. Thus the citation relation is represented by a directed graph.

While a directed graph can express the presence or absence of a relation between individuals, it cannot express the strength of the relation. In consequence, a valued graph is introduced to make the expression of the strength of the relation possible. A valued graph (P; L; Y) is a graph in which a line of set L of lines is accompanied by the value r, where r is mapped onto the real number set Y. Expressing the value $r(u, v)$ accompanied by line (u, v), it is called the value of a line. In the graph of keyword relations, if a line between articles has a large value, it means that they have many common key words that they discuss similar themes.

2.2. Representation of a relational graph using similarity matrix

The procedure of practical modeling is described hereafter. The important relations among scientific articles are the citation relation, the keyword relation and the author relation. These relations are useful to analyze the structure of scientific article-space. However, they are not pertinent to analysis and thus, it is necessary to model the relation.

The fundamental procedure for modeling is as follows.

1. Consider the binary relation between article i and article j.
2. Represent the binary relation as a relational graph that is expressed by point i and point j.
3. Generate a directed-valued relation graph that corresponds to various relations among articles.
4. Represent this graph as a similarity matrix.

Direct-citation relation matrix

The citation relation is particularly important because the references cited by an author give a good indication study. Accordingly, to begin with, we consider the modeling of the citation relation. The direct-citation relation matrix is defied as $\mathbf{A} = [a_{ij}]$. This is a so-called adjacency matrix, where

$$a_{ij} = 1; \quad \text{when article } i \text{ cites article } j,$$

$$0; \quad \text{otherwise}.$$

Total-citation relation matrix

Considering the total-citation relation that involves indirect citation, the total-citation relation matrix is defined as ${}_1\mathbf{S} = [{}_1s_{ij}]$, ${}_1s_{ij} = \sum_{k=1}^{n} w_k \cdot {}_ka_{ij}$, where k is the length of a directed walk, w_k is a weight such as $1/k$ or $1/k^2$ and ${}_ka_{ij}$ stands for the number of a directed-walk whose the length between article i and article j is k. The upper boundary n does not exceed $\max(k)$ (the maximum length of the directed-walk). The ${}_ka_{ij}$ can be obtained by the kth power of matrix $\mathbf{A}$ by resetting the diagonal elements to 0. It should be noted that while the value of a line is 0 or 1 in a direct-citation relation graph, it becomes various values in a total-citation relation graph. An efficient algorithm that calculates only the nonzero elements of matrix $\mathbf{A}$ was developed.

Key word relation matrix

Considering that the existence of many common key words in two articles indicates similar content, the key word relation matrix is defined as ${}_2\mathbf{S} = [{}_2s_{ij}]$, where ${}_2s_{ij}$ is set to m that means the number of common key words in the key word lists of article i and article j. (If a key word is not described explicitly in the articles, then m is defined as the number of common key words in the titles of article i and article j.)

Author relation matrix

Considering that scientific articles written by the same author discuss similar themes or the same theme, the author-relation matrix is defined as ${}_3\mathbf{S} = [{}_3s_{ij}]$, where

$${}_3s_{ij} = 1; \quad \text{when article } i \text{ and article } j \text{ are written by the same author},$$

$$0; \quad \text{otherwise}.$$

Thus, various relation matrices were defined according to various relations among articles, and the citation, key word and author relations were used for combination and analysis in this paper.

3. Optimal Combination of Relations

3.1. Optimal linear combination of relations

Several kinds of relation matrices are generated as similarity matrices described in Section 2. Each similarity matrix can be analyzed individually. That is, when focusing on a specific relation it can be analyzed by using only one kind of similarity matrix. However, when the overall relational structure is considered, it is desirable that

these similarity matrices are connected optimally according to the purpose of analysis. Hereafter, it is the problem of making an optimal model to connect these relational matrices.

The total relation matrix is defined as

$$\mathbf{R} = [r_{ij}], \qquad r_{ij} = \sum_{l=1}^{m} c_l \cdot {}_l s_{ij}^{p_l} \tag{3.1}$$

where m means the number of relations to be coupled and c_l is a coupling-coefficient and p_l stands for a compressing exponent. If p_l is equal to 1, then (3.1) becomes a linear combination. Various methods can be utilized to determine c_l, heuristic and empirical or based on mathematical criteria. Methods to decide coupling-coefficient c_l are discussed next.

3.2. Considerations on factor analysis method for the optimal combination

The model generated by the combination with the optimal coupling-coefficient c_l and can be called the optimal relation model. Considering how to decide coupling-coefficient c_l, it is noted that method of factor analysis is applicable. The processing of the analysis begins after creating a variance covariance matrix or a correlation matrix from the observed data as described by Ch. Asano [15]. R. B. Cattell reported six techniques utilizing a covariation chart that could be used to create those matrices [16]. Q technique and R technique fix or ignore information about time or occasion at the part of analysis. Paying attention only to the relation between an individual and an item, those techniques process test data. That is, the former analyzes the correlation or the covariance between individuals, and the latter analyzes the items in detail. The Q technique is described as follows because the purpose of this study is to discuss the relation among individuals, i.e., articles.

Let x_{pq} designate an original test value, where p means the test item ($p=1, 2, \ldots, m$), and q stands for the number of individuals ($q = 1, 2, \ldots, n$). Each q is a sampling unit, and n is a sampling number. If the correlation between columns that is generated from x_{pq} is r_{ij} ($i, j = 1, 2, \ldots, n$), the Q technique handles matrix $\mathbf{R} = [r_{ij}]$. That is, the technique analyzes the matrix so that a correlation coefficient between individuals is an element.

3.3. Optimal consistency combination

The main theme of the present paper is the generation of a totalized-relation matrix introducing a combination method to optimize several relations among articles. Considering the methodology we become aware of the applicability of canonical correlation analysis. In consequence, we consider joint distribution to deal with a multivariate test item in two different groups at the same time. Thus, a new canonical variable is introduced to combine items linearly in groups for each test item. The new variable of each group is decided to maximize the correlation between each pair of groups. Both new variates are fixed. Since the methodology of the canonical correlation analysis is

applicable, this method of analysis is discussed to decide the optimal linear-combination coefficient c_l when p_l is equal to 1 in (3.1).

In this paragraph, the optimal consistency combination is described. It is assumed that test vector $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ is obtained for p kinds of characteristics to k numbers of populations or categories, where the variance covariance matrix $\boldsymbol{\Sigma}_W^{(i)}$ in each individual population is assumed to have the same $\boldsymbol{\Sigma}$. So population variance covariance matrix $\boldsymbol{\Sigma}_W^{(i)}$ in the ith population is given in (3.2) in the form of an unbiased estimator:

$$\hat{\boldsymbol{\Sigma}}_W^{(i)} = \sum_{j=1}^{n_i} (\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})'/(n_i - 1) \tag{3.2}$$

where $\bar{\mathbf{x}}^{(i)} = \Sigma_j \mathbf{x}_j^{(i)}/n_i$.

From the assumption that each $\boldsymbol{\Sigma}_W^{(i)}$ is equal to a common $\boldsymbol{\Sigma}$, the next equation is derived:

$$\hat{\boldsymbol{\Sigma}}_W = \sum_{i=1}^{k} (n_i - 1)\, \hat{\boldsymbol{\Sigma}}_W^{(i)}/(n - k) \tag{3.3}$$

Eq. (3.3) is the merged equations of k numbers $\hat{\boldsymbol{\Sigma}}_W^{(i)}$ as the maximum likelihood estimator of $\boldsymbol{\Sigma}$ and it is called the sample within-variance-covariance matrix.

On the other hand, total variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_0$ of the sample derived with regard to the above k numbers of populations is given as

$$\hat{\boldsymbol{\Sigma}}_0 = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (\mathbf{x}_j^{(i)} - \bar{\mathbf{x}})(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}})'/(n - 1)\,, \qquad \text{where } n = \sum_{i}^{k} n_i\,, \qquad \bar{\mathbf{x}} = \sum_i \sum_j \mathbf{x}_j^{(i)}/n\,.$$

Accordingly, from the dispersion matrix based on $\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}_0$ of the independent variance covariance matrix, sample between-variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_B$ relevant to k numbers populations is represented as

$$\hat{\boldsymbol{\Sigma}}_B = ((n - 1)\hat{\boldsymbol{\Sigma}}_0 - (n - k)\hat{\boldsymbol{\Sigma}}_W)/(k - 1)\,.$$

Hereafter, we consider a new axis $\mathbf{y} = \mathbf{C}\mathbf{x}$ that maximally expresses the sample variance of k numbers of groups at the test point in p dimensional space subject to $p \geqq q$, where $\mathbf{C}$ is a $q \times p$ matrix and the axis has a vector of order q. This q is thought to be the number of canonical axis or common factor. When a new sample is obtained, by the measure of $\mathbf{C}\mathbf{x}$, it is possible to consider that the its character is similar to that of any population or category. However, because this q is unknown a priori, maintaining analysis of $\mathbf{C}$ as the maximum size matrix $p \times p$, from the result, the eigenvalues of the solution are investigated in order of magnitude just as in principal component analysis. That is, it is considered that eigenvalue well expresses the variation of test data. Considering $\mathbf{C}$ as the transform matrix that changes $\mathbf{x}$ into $\mathbf{y}$, to produce between-variance-covariance matrix $\boldsymbol{\Sigma}_B$ that is maximized by this $\mathbf{y}$,

$$\mathbf{C}\hat{\boldsymbol{\Sigma}}_B\mathbf{C}' \tag{3.4}$$

is obtained.

Because of instability the elements of $\mathbf{C}$ grow without limit thus, the following constraint is set.

$$\mathbf{C}\hat{\Sigma}_W\mathbf{C}' = I \tag{3.5}$$

In consequence, $\mathbf{C}$ is resolved such that it maximizes (3.4) subject to (3.5). Thus, basing this consideration and treatment on canonical analysis, this method of solving becomes similar to canonical correlation analysis.

Practically, the following three relations were examined where the m meant the number of relations among articles in (3.1). That is, the citation relation, key word relation and author relation. Let a similarity matrix be given initially as ${}_lx_{ij} = c_l \cdot {}_ls_{ij}$, $r_{ij} = \sum_{l=1}^{3} {}_lx_{ij}$, where if $l = 1$ then it means the citation relation, if $l = 2$ then it means the key word relation and if $l = 3$ then it means the author relation. Gathering the matrix made from the reversal relations, let the entire relation matrix $\mathbf{R} = [r_{ij}]$ be

$$\mathbf{R} = \begin{bmatrix} 0 & r_{12} & \cdots & r_{1n} \\ r_{21} & 0 & \cdots & r_{2n} \\ & & \vdots & \\ r_{n1} & r_{n2} & \cdots & 0 \end{bmatrix}.$$

Hereafter, considering the variation of the upper triangular matrix in $\mathbf{R}$ because it makes it possible to obtain the variance of the lower triangular matrix, using the same procedure as the following, between-variation (BV) is $\sum_{i=2}^{n} n(\bar{r}_{i.} - \bar{r}_{..})^2$, or $\sum_{l=1}^{3}\left(\sum_{\substack{ij \\ (i \neq j)}} (c_l \cdot {}_ls_{ij} - c_l \cdot {}_ls_{..})^2\right)$, and within-variation (WV) is $\sum_{i=2}^{n}\sum_{j=2}^{n} (r_{ij} - \bar{r}_{i.})^2$ or $g\sum_{l=1}^{3} (c_l \cdot {}_ls_{..} - c_l \cdot .\bar{s}_{..})^2$, where ${}_ls_{..} = \sum_i \sum_j {}_ls_{ij}$, $.s_{..} = \sum_l {}_ls_{..}$, $.\bar{s}_{..} = .s_{..}/3$, $g = n(n-1)$ and n is the number of articles.

Let TV denote total variation, and there exists the relation of TV = BV + WV. In order to realize maximum consistency, it is necessary to maximize between-variation or to minimize within-variation under the condition of $\sum_{l=1}^{3} c_l^2 = 1$.

Let λ denote the Lagrange's multiplier and let Φ be characterized by $\Phi = \mathrm{WV} + \lambda\left(\sum_{l=1}^{3} c_l^2 - 1\right)$.

From the minimization condition with respect to Φ,

$$\partial\Phi/\partial\mathbf{c}_{l_0} = 0\,, \qquad l_0 = 1, 2, 3 \tag{3.6}$$

is set. Consequently, Eq. (3.7) is obvious,

$$\Phi(c_l, \lambda) = g\sum_{l=1}^{3}\left(c_l \cdot {}_ls_{..} - \left(\sum_{l=1}^{3} c_l \cdot {}_l\bar{s}_{..}\right)\Big/3\right)^2 + \lambda\left(\sum_{l=1}^{3} c_l^2 - 1\right) \tag{3.7}$$

From (3.6) and (3.7), the following equations are deduced as $\partial\Phi/\partial\mathbf{c}_{l_0} = 2g\left[\mathbf{c}_{l_0}\cdot({}_{l_0}\bar{s}..) - \left(\sum_{l=1}^{3} c_{l_0}\cdot {}_{l}\bar{s}..\right)\Big/3\right]\cdot {}_{l_0}\bar{s}.. + 2\lambda\mathbf{c}_{l_0} = 0$, $l_0 = 1, 2, 3$, and,

$$\left[\mathbf{c}_{l_0}\cdot({}_{l_0}\bar{s}..) - \left(\sum_{l=1}^{3} c_{l_0}\cdot {}_{l}\bar{s}..\right)\Big/3\right]\cdot {}_{l_0}\bar{s}.. = 0\,, \qquad l_0 = 1, 2, 3 \tag{3.8}$$

From (3.8), the next equations are deduced successively as

$$3\cdot\mathbf{c}_{l_0}\cdot {}_{l_0}\bar{\mathbf{s}}^2.. - c_1\cdot {}_1\bar{s}.. + c_2\cdot {}_2\bar{s}.. + \cdots + \mathbf{c}_{l_0}\cdot {}_{l_0}\bar{\mathbf{s}}.. + \cdots + c_g\cdot {}_g\bar{s}..)\cdot {}_{l_0}\mathbf{s}.. = \mathbf{0}\,,$$

$$(3-1)\mathbf{c}_{l_0}\cdot {}_{l_0}\bar{\mathbf{s}}^2.. - (c_1\cdot {}_1\bar{s}.. + c_2\cdot {}_2\bar{s}.. + \cdots + \mathbf{c}_{l_0}\cdot {}_{l_0}\bar{\mathbf{s}}.. + \cdots + c_g\cdot {}_g\bar{s}..)\cdot {}_{l_0}\mathbf{s}.. + \lambda\cdot 3\cdot\mathbf{c}_{l_0} = \mathbf{0} \quad \text{and}$$

$$-\{c_1\cdot {}_{l_0}\bar{\mathbf{s}}..\cdot {}_{l_0}\bar{\mathbf{s}}.. + c_2\cdot {}_2\bar{s}..\cdot {}_{l_0}\bar{\mathbf{s}}.. + \cdots + (3-1)\cdot\mathbf{c}_{l_0}\cdot {}_{l_0}\bar{\mathbf{s}}^2.. + \cdots + c_g\cdot {}_g\bar{s}..\cdot {}_{l_0}\bar{\mathbf{s}}..\} - \lambda\cdot 3\cdot\mathbf{c}_{l_0} = \mathbf{0} \tag{3.9}$$

After dividing both sides of (3.9) by 3, finally letting it express matrix form,

$$(\mathbf{D} - \lambda\mathbf{I})\mathbf{C} = \mathbf{0} \tag{3.10}$$

is obtained, where

$$\mathbf{D} = \begin{bmatrix} (3-1)\cdot {}_1\bar{s}^2_{\cdot 1}/3 & {}_1\bar{s}_{\cdot 1}\cdot {}_2\bar{s}_{\cdot 2}/3 & {}_1\bar{s}_{\cdot 1}\cdot {}_3\bar{s}_{\cdot 3}/3 \\ & (3-1)\cdot {}_2\bar{s}^2_{\cdot 2}/3 & {}_2\bar{s}_{\cdot 2}\cdot {}_3\bar{s}_{\cdot 3}/3 \\ & & (3-1)\cdot {}_3\bar{s}^2_{\cdot 3}/3 \end{bmatrix}, \quad \mathbf{C} = [\mathbf{c}_{l_0}]\,, \quad l_0 = 1, 2, 3\,.$$

Quantity $\mathbf{c}_{l_0}$ is given by the eigenvalue of (3.10).

3.4 Optimal generalizability combination

Another approach is attempted in the this section. That is, optimal generalizability for combination is discussed. This method is used to maximize the coefficient of reliability. The fundamental concept of optimal generalizability is based on the optimization of the infinite numbers of coefficients. Alpha-factor analysis was proposed by L. J. Cronbach and H. F. Kaiser [17, 18, 19] as a method of factor analysis containing psychological considerations. The notation alpha in this section represents the coefficient of reliability.

The following quantities are defined:

$$\mathbf{x}_j = \mathbf{A}\mathbf{f}_j + \mathbf{e}_j\,, \qquad j = 1, 2, \ldots, n \tag{3.11}$$

where $\mathbf{x}_j$ is a p dimensional test vector, $\mathbf{A}$ is the $p \times q$ factor loading matrix, $\mathbf{f}_j$ is the common factor of order q and $\mathbf{e}_j$ is the stochastic error vector or specific factor vector of order p. Assuming the normalization of $\mathbf{x}_i$ and $\mathbf{f}_i$, population variance covariance matrix $\mathbf{\Sigma}$ is equal to population correlation matrix $\mathbf{P}$, and from the assumption of $\mathbf{\Sigma} = \mathbf{P} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}$,

$$\mathbf{A}\mathbf{A}' = \mathbf{P} - \mathbf{\Psi} \tag{3.12}$$

holds, where $\mathbf{\Psi}$ is a diagonal matrix of error variance. Letting $\mathbf{H}^2$ be the communality-

diagonal matrix, the following equation holds as $\mathbf{\Psi} + \mathbf{H}^2 = \mathbf{I}$. Accordingly (3.12) can also be rewritten as $\mathbf{AA}' = \mathbf{P} - \mathbf{I} + \mathbf{H}^2$. Let the part of the common factor in Eq. (3.11) be Eq. (3.13):

$$\mathbf{C} = \mathbf{Af}, \qquad \mathbf{f} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{C} \tag{3.13}$$

where $\mathbf{A}$ is the $p \times q$ factor loading matrix and $\mathbf{f}$ is the common factor score vector of order q.

Let the sth element of common factor score vector $\mathbf{f}$ of order q be

$$f_s = \sum_{j=1}^{p} w_{sj}c_j, \qquad s = 1, 2, \ldots, q \tag{3.14}$$

where w_{sj} is (s, j) element of $p \times q$ matrix $\mathbf{W} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ and c_j expresses the common part of $\mathbf{x}$ in the set of test properties and w_{sj} expresses its coefficient. However, Eq. (3.14) hold true only for p kinds of test properties. So that, considering the entire property of universe, let the sth element of the common factor score vector be

$$\zeta_s = \sum_{j=1}^{\infty} w_{sj}c_j, \qquad s = 1, 2, \ldots, q \tag{3.15}$$

Taking this under consideration, alpha-factor analysis is used to obtain the factor loading matrix in order to maximize the correlation between common factor score f_s and common factor score ζ_s. Consequently, the concept of canonical correlation analysis is applicable. The psychological considerations that analysts consider to study exact representations such as (3.15) to obtain a highly reliable estimation is similar to the concept of generalizability announced in 1937 and in the same year the formula of Kuder-Richardson was also proposed [20]. These concepts were further developed by L. J. Cronbach et al. and the square of the correlation between f_s and ζ_s is called coefficient of generalizability. As the coefficient of reliability, there are several variations of this coefficient of generalizability. Using similar traditional formulae, one criterion called Kuder-Richardson's coefficient of reliability or the alpha coefficient of Cronbach is introduced as

$$\alpha = \{p/(p-1)\}\{1 - \mathbf{w}'\mathbf{H}^2\mathbf{w}/\mathbf{w}'(\mathbf{P} - \mathbf{\Psi})\mathbf{w}\}.$$

In consequence, letting $\mathbf{P} - \mathbf{\Psi}$ be the total variance $\mathbf{\Sigma}$ of the common part of test v, substituting $\mathbf{H}^2$ by within-variance $\mathbf{\Sigma}_W$ and $\mathbf{P} - \mathbf{I}$ by between-variance $\mathbf{\Sigma}_B$, in order to maximize α, the F. M. Load proposed to maximize μ^2 as defined by

$$\mu^2 = \mathbf{w}'(\mathbf{P} - \mathbf{\Psi})\mathbf{w}/\mathbf{w}'\mathbf{H}^2\mathbf{w} \tag{3.16}$$

To maximize μ^2, partially differentiating Eq. (3.16) by $\mathbf{w}$, the next equation is obtained as

$$\partial\mu^2/2\partial\mathbf{w} = ((\mathbf{P} - \mathbf{\Psi})\mathbf{w} - \mu^2\mathbf{H}^2\mathbf{w})/\mathbf{w}'\mathbf{H}^2\mathbf{w} = \mathbf{0} \tag{3.17}$$

Consequently, the numerator of (3.17) is required to be zero as

$$(\mathbf{P} - \mathbf{\Psi})\mathbf{w} - \mu^2\mathbf{H}^2\mathbf{w} = \mathbf{0}.$$

As a result, the following equation becomes

$$[(\mathbf{P} - \mathbf{\Psi}) - \mu^2\mathbf{H}^2]\mathbf{w} = \mathbf{0} \tag{3.18}.$$

Letting $\mathbf{d} = \mathbf{Hw}$ be, Eq. (3.18) is transformed as

$$[\mathbf{H}^{-1}(\mathbf{P} - \mathbf{\Psi})\mathbf{H}^{-1} - \mu^2\mathbf{I}^2]\,\mathbf{d} = \mathbf{0}$$

Accordingly, let λ be the maximum eigenvalue and let its eigenvector be $\mathbf{l}$, the next equation for the alpha factor is obtained as $\mathbf{a} = \sqrt{\lambda}\mathbf{Hl}$.

4. Cluster Analysis of The Similarity Matrix

4.1. Characteristics of the citation relation matrix

The matrix to be analyzed is given by $\mathbf{R} = [r_{ij}]$, $r_{ij} = (s_{ij})^p$, where p is an enhancing exponent. This matrix has the following characteristics. First, each element of the matrix has a positive real quantity. Element s_{ij} is similarity as a correlation-like measure and element r_{ij} has a positive real quantity. Second, the matrix is asymmetric. The similarity matrix of the citation relation is asymmetric because there is a time sequence in the citation relation between a citing article and a cited article. In consequence, matrix $\mathbf{R}$ becomes asymmetric. When the methods of the combinatorial cluster analyses are applied, it is necessary to transform the matrix to a symmetric matrix.

4.2. Application of combinatorial clustering methods

All computer programs using the following combinatorial methods have been implemented and applied to the analysis of the relational graph model. Seven hierarchical clustering methods were examined. The nearest neighbour method is known as single linkage because clusters are joined at each stage by the single shortest or strongest link between individuals. The furthest neighbour method is also called the complete-linkage method because all individuals in a cluster are linked to each other by some minimum similarity. The median method adopts the middle value of the nearest neighbour value and the furthest neighbour value. The method of average linkage within the new group is not influenced by extreme values for defining clusters so it cannot make any statements about the minimum or maximum similarity within a cluster. Average linkage between a merged group, also called the group-average method, evaluates the potential merger of clusters i and j in terms of the average similarity between the two clusters. The difference between the latter and the former is whether the sums of within-group pairwise similarities are ignored or not. The centroid method uses both the mean value of similarities and the number of individuals for the merger. The minimum variance method (the Ward method), used to find at each stage those two clusters whose merger gives the minimum increase in the total within-group error-sum of squares, is generally reasonable even if it is not optimum.

When these combinatorial methods were applied, asymmetric similarity matrix **R** was changed to a symmetric matrix by

$$r_{ij} = (\max(s_{ij}, s_{ji}))^p, \qquad (i > j) \tag{3.19}$$

5. Exploratory Result

To examine the present method, 231 scientific articles concerning CAD/CAM were investigated by using four models, that is, the direct-citation relation model, the total-citation relation model, the optimal consistency relation model and the interpretive structural method (ISM). They mainly contained articles about computational geometry and several articles relating to artificial intelligence (AI). In the following comparison with manual classification by experts, the conformance percentage is the separation ratio between the AI-cluster and the proper CAD/CAM cluster. It was expected that this method would suppress the occurrence of inaccuracy due to the analyst's subjectivity since the research filed of AI is obviously different from CAD/CAM. The seven combinatorial cluster analyses programmed by M. R. Anderberg were applied [21]. The average linkage within the new group method was the best. Other combinatorial methods are not useful because these clusters never or rarely agglomerate until the last stage. Accordingly, these methods are not adequate for analyzing the relation structure of scientific articles. In order to compare the present method with the interpretive structural model (ISM) [22], enhancing factor p in Eq. (3.19) was set to 0 to examine these models by combinatorial analyses. The present model is construed to correspond to ISM since total-citation relation matrix $\mathbf{S}_1$ is equal to the reachability matrix when $c_2 = c_3 = p_1 = 0$ and weight w_k of matrix S_1 is not equal to 0. While ISM used the reachability matrix derived from the total-citation relation, the present method was better than ISM in contrast; i.e., the model using the direct-citation relation was 71.1% and the model using the total-citation relation was 72.2% and ISM was 71.1% as shown in Table 1. Using the optimal consistency combination with the relations of total citation, key word and author, the conformance percentage attained by the present modeling was 86.8%.

Table 1. Conformance with human classification for 231 articles in the field of CAD

Relation	Conformance with human classification
Direct citation	71.1%
Total citation	72.2%
Optimal consistency combination	86.8%
(ISM)	71.1%

6. Conclusion

The methods examined are modeling with the direct-citation relation, modeling with the total-citation relation and modeling by the optimal consistency com-

bination of the citation relation, the key word (title) relation and the author relation, and the ISM method. Since the model using the optimal consistency combination was the best, it was concluded that some relations among scientific articles should be considered totally and this method can express the original space of articles more accurately than ISM or models using one kind of relation. Though the final result does not necessarily conform with manual classification completely, it can be expected that the present method gives us significant information that can only be extracted with difficulty by human classification. This method of modeling will be applicable not only to classification of scientific articles but also to analysis of relational structures in the engineering field.

Acknowledgment

The author would like to express his hearty gratitude to Professor Chooichiro Asano of Kyushu University for his guidance on theoretical ground, valuable suggestions and continuous encouragement. Thanks are also owed to Professor Hajime Tanaka (Sapporo Gakuin University) and Professor Yoshinori Akaishi of Hokkaido University for their significant advice.

References

[1] BORENIUS, G. and SCHWARZ, S.: *Remarks on the use of citation data in predictive models of scientific activity*, Info. Stor. Retr., **8**, (1972), 171–175.

[2] CAWKELL, A. E.: *Evaluating scientific journals with Journal Citation Reports,—a case study in acoustics*, J. Amer. Soc. Info. Sci., **29**, (1978), 41–46.

[3] JARDINE, N. and RIJSBERGEN, C. J. V.: *The use of hierarchic clustering in information retrieval*, Info. Stor. Retr., **7**, (1971), 217–240.

[4] RIJSBERGEN, C. J. V.: *Further experiments with hierarchic clustering in document retrieval*, Info. Stor. Retr., **10**, (1974), 1–14.

[5] VOORHEES, E. M.: *Implementing agglomerative hierarchic clustering algorithms for use in document retrieval*, Info. Proc. Manag., **22**, 6, (1986), 465–476.

[6] LUCARELLA, D.: *A document retrieval system based on nearest neighbour searching*, J. Info. Science, **14**, 1, (1988), 25–33.

[7] SAITO, T., TEJIMA, S., KAWAI, N. and OKINO, N.: *Study on development of methodology for grasping research trend and efficiency of the scientific information retrieval by it*, Formation Process of Information Systems and Organization of Scientific Information-Report (a Grant-in-Aid of the Ministry of Education of Japan), (1977), (in Japanese).

[8] SAITO, T., OKINO, N., ASANO, Ch. and TANAKA, H.: *A modeling and clustering method by relational graph model and an exploratory application for cluster analysis of scientific article space*, SCIENCE AND TECHNOLOGY OF SOCIETY: Proc. IFAC 81, **XII**, (1982), 62–67.

[9] CUMMINGS, L. J. and FOX, D. A.: *Some mathematical properties of cycling strategies using citation indexes*, Info. Stor. Retr., **9**, (1973), 713–719.

[10] TODOROV, R. and GLANZEL, W.: *Journal citation measurers—a concise review*, J. Info. Science, **14**, 1, (1988), 47–56.

[11] KOCHTANEK, T. R.: *Bibliographic compilation using reference and citation links*, Info. Proc. Manag., **18**, 1, (1982), 33–39.

[12] NOMA, E.: *Untangling citation networks*, Info. Proc. Manag., **18**, 2, (1982), 43–53.

[13] TOMER, C.: *A statistical assessment of two measures of citations: The impact factor and the immediacy index*, Info. Proc. Manag., **22**, 2, (1986), 251–258.

[14] LOIZOU, G.: *Graph model of complex information sources*, Info. Proc. Manag., **15**, (1979), 127–131.
[15] ASANO, Ch.: *Factor analysis, theory and applications*, Kyoritsu (1971), (in Japanese).
[16] CATTELL, R. B.: *Factor analysis*, Harper & Brothers, (1952).
[17] CRONBACH, L. J.: *Test 'reliability': its meaning and determination*, Psychometrika, **12**, (1947), 1–16.
[18] CRONBACH, L. J.: *Coefficient alpha and the internal structure of tests*, Psychometrika, **16**, (1951), 297–334.
[19] CRONBACH, L. J., RAJARATNAM, N. and GLESER, G. C.: *Theory of generalizability: a liberalization of reliability theory*, Brit. J. Statist. Psychol., **16**, (1963), 137–163.
[20] KUDER, G. F. and RICHARDSON, M. W.: *The theory of the estimation of test reliability*, Psychometrika, **2**, (1937), 151–160.
[21] ANDERBERG, M. R.: *Cluster analysis for applications*, Academic Press, New York, (1973).
[22] WARFIELD, J. N.: *Societal systems-planning, policy and complexity*, John Wiley, (1976).

Received September 5, 1989
Revised September 27, 1989
Communicated by Ch. Asano