北海道大学工学部研究報告
第138号（昭63年）

Bulletin of the Faculty of Engineering,
Hokkaido University. No. 138 (1988)

最適関係構造化法による学術文献のクラスタ分析

斉藤たつき

（昭和62年9月30日受理）

Cluster Analysis for Scientific Articles by Optimal Relationship Model

Tatsuki SAITO

(Received September 30, 1987)

Abstract

A methodology for modeling and clustering of large scale complex information space is discussed in order to analize the structure of relation among scientific articles. An object to be modeled in data base system "ANGEL" is expressed by a point, and a binary relation between objects is expressed by points and lines in relational graph. A similarity matrix is derived through operation of an initial relational graph matrix. And various similarity matrices are produced from various relations in the information space to be dealt with. A method which optimizes the consistency is proposed when these real number matrices are combined. And another method is proposed for modeling and clustering by application of characteristics of the relational graph. The performance of the proposal method was superior to the interpretive structural model which applies a 1-0 reachability matrix for cluster analysis of some scientific articles.

1．はじめに

研究者にとって関連する研究論文間の関係を把握することは重要である。研究課題に関する問題を明確にしたり，解決の見通しをたてたり，あるいは方法論の検討をしたりする際には，常に他の研究から影響を受け，あるいは自己の研究が他の研究に何らかの形で刺激を与えている。このようなことから，研究論文間に内在する関係構造を明らかにする方法論の開発が重要と考えられる。他方，最近のおびただしい数の論文生産量は研究分野によっては，研究者がその関連ある文献すべてに目を通すことをますます困難にさせつつある。こうした背景を考慮してできるだけコンピュータで処理可能な方法論をめざすことにした。

2．情報の収集と管理

研究論文間の関係を明らかにするためには，それらの情報を効率的に収集し，それを効果的に利用することが重要である。そこで，文献間の種々の関係を取り扱うことを可能とするデータベースシステムANGELを設計し製作した。文献情報の中で，引用のされかたは様々であるが，引用情報は文献間の強い結び付きを表している情報であり，本研究では処理効率を考えて論文中の引

精密工学科精密機器学第一講座

用箇所やどの節で引用されているかという情報も入力できるようにした。このデータベースは一般的な書誌事項を扱う以外に，他の一般的な文献情報データベースと異なる特長として，文献間の引用関係事項の他にタイトル中に共通に含むタームによる文献の関係，著者関係等の諸関係を取り扱うことができるようになっている。ANGEL には現在までに，引用文献も含めて CAD/CAM 関係の約 2000 件の文献が蓄積されている。その他に，内外の CAD 研究者にアンケートを送り自己またはそのグループの研究に最も関連のある文献を直接アンケートによって指摘してもらったデータ 154 件が入っている。

3．関係グラフモデル

研究論文間の関係を構造化する方法として，文献の間に存在する種々の関係に着目し，それらの関係をグラフとしてモデル化することにした。

3．1　モデル化の基本手順

モデル化の基本手順は以下の要領でおこなう。

(1) 文献 i と文献 j の 2 項関係を考える。

(2) この関係を点 i，点 j とした連関グラフに表現する。

(3) 種々の関係に対応する連関グラフを行列表現する。これらの行列は類似度行列として与えられる。

(4) 各類似度行列を整合性が最適になるように線形結合する。

なお，モデル化と一体になった新しいクラスタアナリシス法については 4．2 で述べる。

3．2　連関構造化モデル

[定義 1]　直接引用関係行列を $A=[a_{ij}]$ と定義する。これはいわゆる隣接行列である。

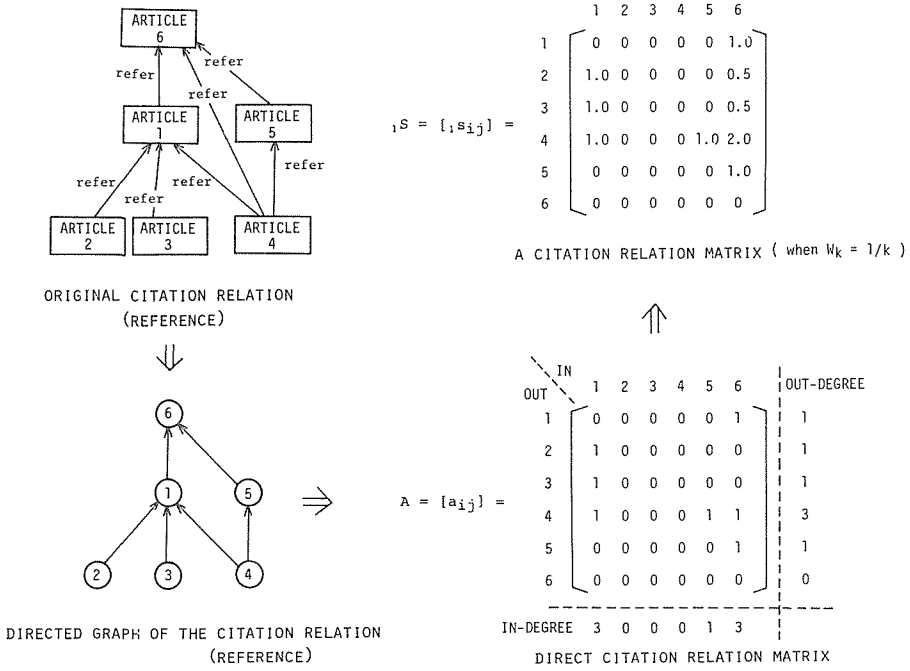

図 1　引用関係グラフ，直接引用関係 A，引用関係行列 ${}_1S$

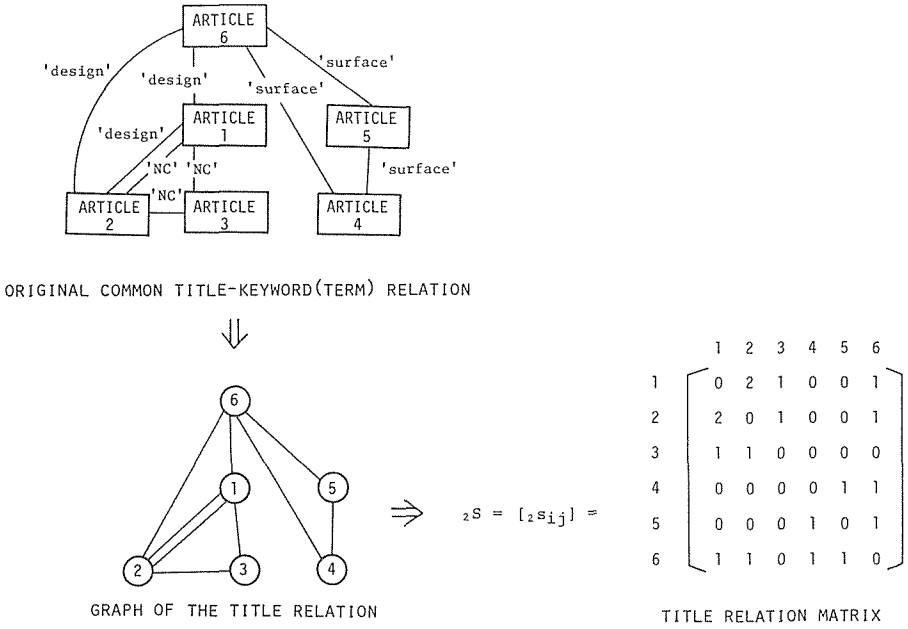

図2　タイトル関係グラフ，タイトル関係行列 ${}_1S$

ここで，

$a_{ij}=1$ ；文献 i が文献 j に引用されている場合

0 ；それ以外の場合　とする。（図1）

[定義2]　引用関係行列を ${}_1S=[{}_1s_{ij}]$ と定義する。ここで，${}_1s_{ij}=\sum_{k=1}^{n} w_k \cdot a_{ij}^k$ とする。ただし，k は有向遊歩道の長さ，a_{ij}^k は長さ k の有向遊歩道の個数，$n \leqq \max(k)$（最大有向遊歩道長），w_k はウエイトで，b は k に等しくない定数としたとき，$1/k$，$1/k^2$ 等の値をとるものとする。なお，a_{ij}^k は A を主対角要素を常に0にしながら k 乗することによって求めることができる。

証明

a_{ik}^1 を点 P_i，点 P_k 間の遊歩道の個数，a_{kj}^1 を点 P_k，点 P_j 間の長さ1の遊歩道の個数とすると，$a_{ik}^1 \cdot a_{kj}^1$ は点 P_k を通る点 P_i，点 P_j 間の長さ2の遊歩道の個数である。なぜなら，点 P_i，点 P_k 間の遊歩道と点 P_k，点 P_j 間の遊歩道は独立にそれぞれ a_{ik}^1 個，a_{kj}^1 個存在しているので，その組合せは $a_{ik}^1 \cdot a_{kj}^1$ 個になるからである。したがって，i, j 以外のすべての k に対してこの積を求めると，長さ2の遊歩道の個数がえられる。つまり，$\sum_k a_{ik} \cdot a_{kj} \rightarrow a_{ij}^2$ となる。同様にして，$a_{ij}^2 \cdot a_{jm}$ は点 P_j を通る点 P_i，点 P_m 間の長さ3の遊歩道の個数になる。ゆえに，$A^n=A^{n-1}A$ より，a_{ij}^n は点 i，点 j 間の長さ n の遊歩道の個数に等しい。

（証明終わり）

なお，行列 A のノンゼロ要素のみを計算する効率的なアルゴリズムを開発した。

[定義3]　タイトル関係行列を ${}_2S=[{}_2s_{ij}]$ と定義する。ここで，${}_2s_{ij}=m$ とする。ただし，m は文献 i と文献 j のタイトル中に共通に存在しているタームの個数である。（図2）

[定義4]　著者関係行列を ${}_3S=[{}_3s_{ij}]$ と定義する。

${}_3s_{ij}=$　1 ；文献 i と文献 j が同一著者による場合

0 ；それ以外の場合　とする。（図3）

つぎに，これらの関係行列を結合した総合関係行列を考える。

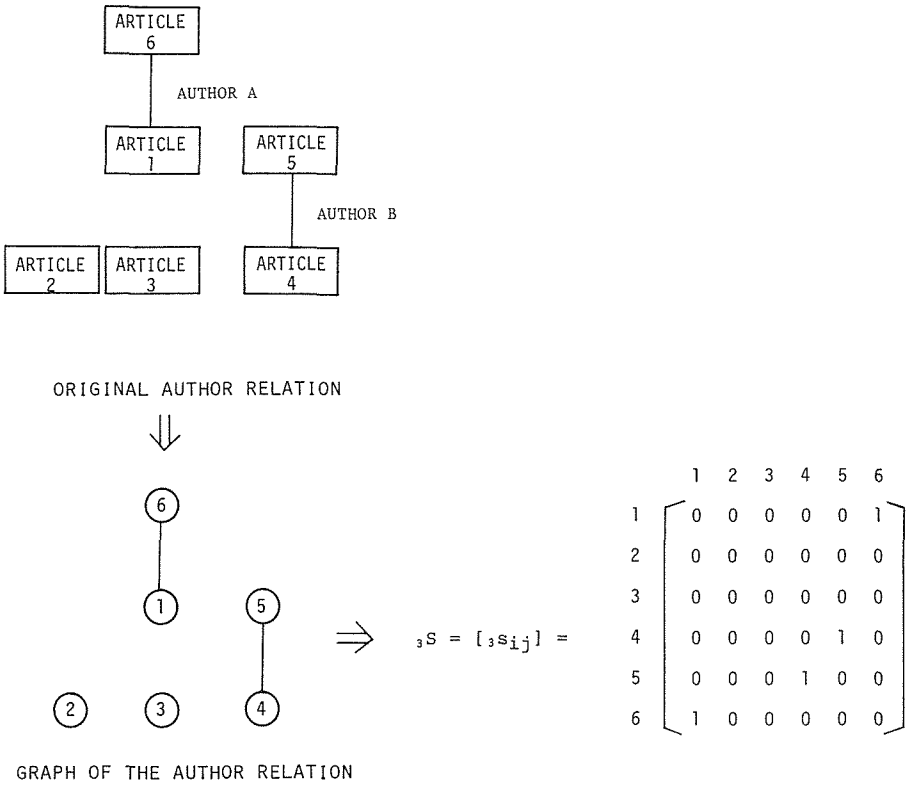

図3　著者関係グラフ，著者関係行列 $_3S$

[定義5]　総合関係行列を $R=[r_{ij}]$ と定義する。

$r_{ij}=\sum_{l=1}^{h} c_l \cdot {}_lS_{ij}^{p_l}$ である。ただし，c_l は結合係数，p_l は圧縮係数で実験では $h=3$ とした。

[注1]　このモデルは引用関係行列 $_1S$ のウエイト $w_k \neq 0$ とし，$c_2=c_3=0$，$p_1=0$ としたとき，Interpretive Structural Model に対応する。

[注2]　電気抵抗回路網にアナロジーをとる場合は i, j 間の合成抵抗値に対応するものとして引用関係行列を求め，その逆数を要素 $_1s_{ij}$ として与える。

[注3]　従来の組合せ論的クラスタアナリシスによる場合は R を対称化する必要がある。こうして生成された各関係行列を4に述べるクラスタ分析にかける。

3．3　最大整合性を基準とするクラスタ化

$R=[r_{ij}]$, $r_{ij}=\sum_{l=1}^{n} c_l \cdot {}_ls_{ij}$ とし，そこですべての文献間の2項関係を行方向に，引用関係（$l=1$），タイトル関係（$l=2$），著者関係（$l=3$）を列方向にとった行列を考える。このとき，Total Variance＝Between Variance＋Within Variance の関係があり，ここに $g=(n-1)n/2$ としたとき，Between Variance：$\sum_{l=1}^{3}(\sum_{ij}(c_l \cdot {}_ls_{ij}-c_l \cdot {}_ls_{..})^2)$ かつ Within Variance：$g\sum_{l=1}^{3}(c_l \cdot {}_ls_{..}-\bar{s}_{..})^2$ である。このとき，λ をラグランジェの未定乗数，$\sum_{l=1}^{3} c_l^2=1$ として，$\Phi=$Within Variance$+\lambda(\sum_l c_l^2-1)$ を最小化する $\{c_l\}$，λ を求める。この解は

$$D=\begin{bmatrix} \frac{(3-1)\cdot {}_1\bar{s}^2_{\cdot\cdot}}{3} & \frac{{}_1\bar{s}_{\cdot\cdot}\cdot {}_2\bar{s}_{\cdot\cdot}}{3} & \frac{{}_1\bar{s}_{\cdot\cdot}\cdot {}_3\bar{s}_{\cdot\cdot}}{3} \\ & \frac{(3-1)\cdot {}_2\bar{s}^2_{\cdot\cdot}}{3} & \frac{{}_2\bar{s}_{\cdot\cdot}\cdot {}_3\bar{s}_{\cdot\cdot}}{3} \\ & & \frac{(3-1)\cdot {}_2\bar{s}^2_{\cdot\cdot}}{3} \end{bmatrix}$$

として，$(D-\lambda\ 1)C=0$ を解けばよいことになる。

4．関係度依存法によるクラスタ分析法

4．1　関係度の算出

方向性をもった構造を表現した非対称な類似度行列も扱える新しいモデル化法およびクラスタ法を以下に述べる。

[定義 6]　結合行列を $X=[x_{ij}]$ と定義する。

$x_{ij}=$　$1/od(i)$ ；点 i における出次数 $od(i)$ が 0 でない場合

　　0 ；それ以外の場合

ただし，t は点の総数であり，また $od(i)=\sum_{j=1}^{t} a_{ij}$ である。

[定義 7]　直接関係度行列を $Y=[y_{ij}]$ と定義する。ここで，$y_{ij}=a_{ij}x_{ij}$ とする。

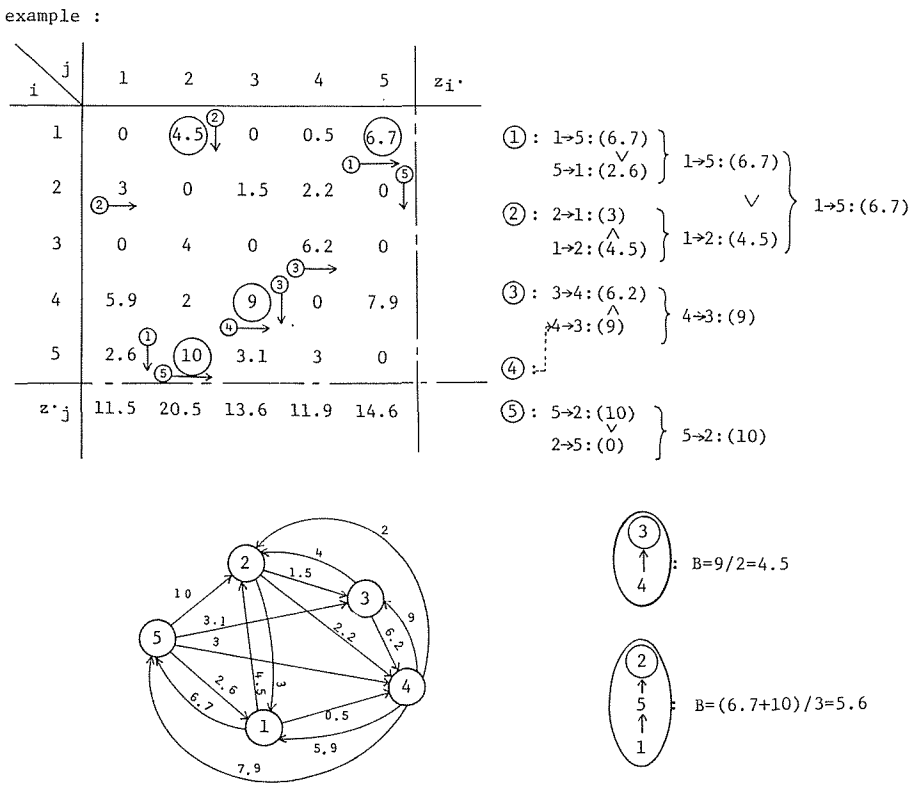

図 4　2 者間関係度依存法によるクラスタ分析法

［定義８］ 関係度行列を $Z=[z_{ij}]$ と定義する。ここで，$z_{ij}=\sum_{k=1}^{n} w_k \cdot y_{ij}^k$ とする。

［定義９］ 点 j における全関係度を $z_{\cdot j}=\sum_{i=1}^{t} z_{ij}$ と定義する。

クラスタリングアルゴリズムは２者間の関係度を基準とする２者関係度依存法と，自分以外の他者全体との関係度を基準とする全関係度依存法の２つを開発した。

４．２　２者関係度依存法によるクラスタ分析法

この方法では，関係度 $z_{ik} \geqq z_{ki} (i \neq k)$ なら点 i はクラスタ k にクラスタリングされる。つぎに，処理手順の概要を述べる。

［手順１］ 行方向に z_{ij} の最大値を探索しそれを z_{ik} とする。

［手順２］ z_{ik} と z_{ki} を比較して z_{ik} が大きければ点 i はクラスタ k にクラスタリングされ，z_{ki} が大きければ点 k がクラスタ i にクラスタリングされる。

図４は具体例を示す。このクラスタリング法の特徴はトリー状に階層が何層にも及ぶことである。

４．３　全関係度依存法によるクラスタ分析法

最終クラスタ個数，つまり核にすべき個数 u を決め，全関係度 $z_{\cdot j}$ の大きなものから順に核を u 個決める。残りの点はどの核との関係度が最大か調べ，最大関係度を与える核にクラスタリングされる。処理手順の概要は，

［手順１］ 全関係度をみて大きい順に u 個核 $j_s (j_s=j_1,\ j_2,\ \cdots\cdots,\ j_u)$ とする。

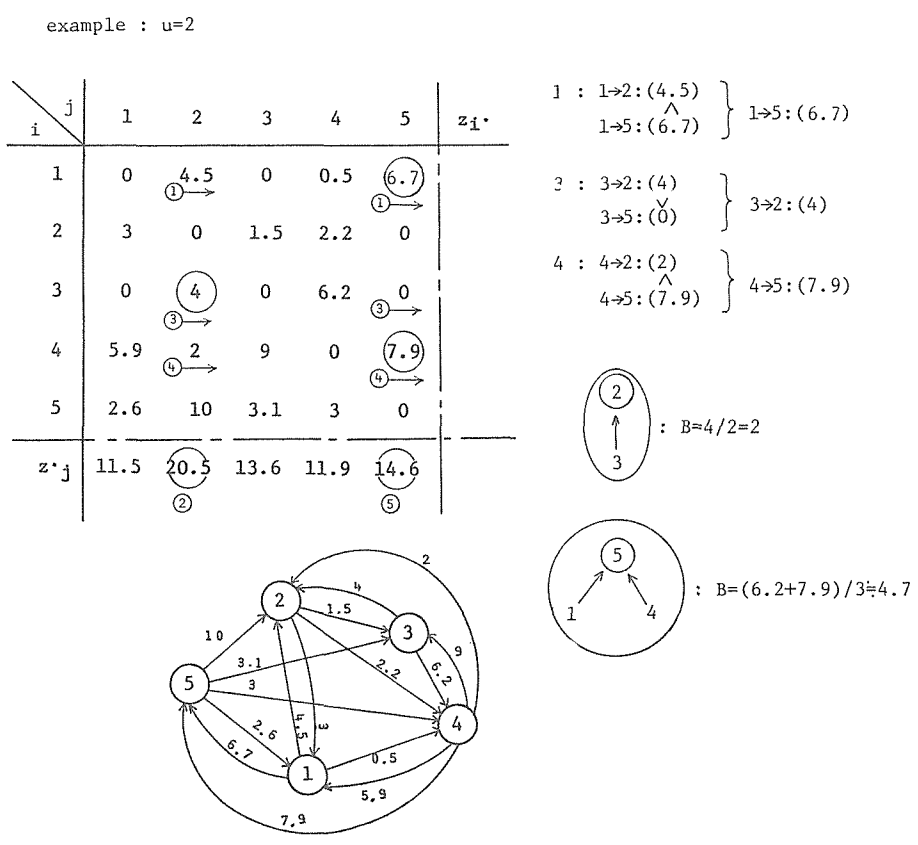

図５　全関係度依存法によるクラスタ分析法

［手順2］ 残りの点は一番大きな関係度を与える核にクラスタリングする。

図5に具体例を示す。このクラスタリング法の特徴は階層が2層になり，かつクラスタの個数が多くなり，またどのクラスタにも属さない個体ができる傾向がある。

5．実験結果と考察

実験に使用したデータはCAD/CAM関連の曲面創成理論を中心とした231文献である。また，本方法論が他の異なる性格の分野にも適合性があるかどうかを検証する目的で，原子核問題の140文献について実験した。表1はその分野の専門家が分類したものとの一致度である。必ずしも人間の分類と完全に一致させることが最終目標でなく，場合によっては人間の分類では期待できない特性をもったクラスタリングに意義があるとも考えられるが，ひとつの評価基準になりうるものとして比較してみた。表でCitation Ratioは1文献あたりの引用数のことである。原子核問題140文献の方がCAD/CAM 231文献に比して2倍の強さの引用の結びつきがあることを物語っている。以下に，結論を要約すると，

1．ISMモデルよりも最適関係構造化法の方が学術文献のクラスタ化に向いていることが確認された。

2．総合関係を考慮することが意味があることであると結論される。特に，CAD/CAMの分野には著しい（72.2%→86.6%）効果があった。

3．階層的なクラスタ分析法よりも本研究で提案した関係度依存法の方が良好な結果がえられることが確認できた。これは，学術文献情報空間は方向性をもった構造をしているためと考えられる。

4．本方法で原子核140文献をクラスタ分析した結果，全関係度の上位3つに入った文献はいずれも3つのクラスタの核となるべき中心的存在の文献であった。さらに，上位6つに入ってきた新たな3文献は，いずれも3つのクラスタの準核となるべき存在のものであることがわかった。

5．原子核の分野では有向遊歩道の長さkはせいぜい2ぐらいまで考慮すればよい反面，CAD/CAMの分野では到達可能な先端（$k=13$）まで考慮すべきであることが明らかになった。このことは，後者の特性として広範な研究と深く関わり合っていることがあげられよう。

表1　各専門家の分析結果との一致率

方法論等 ＼ 処理対象	CAD/CAM関連 231文献	原子核関連 140文献
Interpretive Structural Model	71.1%	
引用関係情報のみによる関係構造化法	72.2%	60.3%
最適関係構造化法	86.6%	78.5%
Citation Ratio	1.7	3.5

参考文献

1) Tatsuki Saito et al : A Modeling and Clustering Method by Relational Graph Model and an Exploratory Application for Cluster Analysis of Scientific Article Space, Proceedings of IFAC 8th World Congress, (1982), pp.XII62～XII67
2) 斉藤たつき：学術文献分類システムの一方法，日本行動計量学会大会発表論文抄録集，(昭57)
3) 斉藤たつき：学術文献情報のモデル化法と分類の一方法，計測自動制御学会システムシンポジュウム講演論文集，(昭57)
4) 斉藤たつき：研究論文空間の分類法，日本行動計量学会大会発表論文抄録集，(昭59)