4X−1

有向グラフのノードのクラスタリングにおけるノード間類似度の考察

千村 浩靖
日本電気(株) C&C情報研究所

1．はじめに

システム構造解析においては、対象システムを構成するサブシステムをノードに、サブシステム間の関係をアークにそれぞれ対応させたグラフとして全体構造を表現し、構造把握・分析することが行われる（例えば、ISM法[1]）。この時、作成したグラフのノードを何らかの観点からクラスタリングすることができれば、全体構造をより把握しやすくなる。

本稿では、有向グラフのノードクラスタリングを行う場合の、ノード間類似度定義方法についての基礎的検討を行う。

2．問題の所在と既報告手法および本研究の立場

通常のクラスタリング手法は、各サンプルに対して与えられた何らかの特性値データからサンプル間の類似度（または距離）を算出し、その類似度に基づいてクラスタリングを行うものである。しかし、上記のような特性値データが存在せず、単にサンプル間の関係のあり／なしの情報のみが与えられた場合、すなわち、ノードとアークからなるグラフのみが与えられた場合には、他のクラスタリング手法が必要となる。

上記の課題に対して、以下の報告がある。

Botafogoは、ハイパーテキスト構造解析の領域において、テキストをノードに、リンクをアークに対応させた有向グラフを対象として、このグラフから2連結成分を漸次切り出す事により、ノードのクラスターを求める手法を提案している[2]。

原は、同じくハイパーテキスト構造解析において、有向グラフ表現から単純でロバストな高次構造を抽出するACEクラスタリング手法およびその実行のための発見的アルゴリズムを提案している[3]。

佐古は、ドキュメンテーションエンジニアリングの領域において、マニュアルモジュールをノードに、モジュール間の参照関係をアークに対応させたグラフを考え、参照関係に応じた引力・斥力をノード間に定義し、ノードを自由運動させることによりノードのクラスターを求める手法を提案している[4]。

上記の手法は、いずれも対象システムが比較的大規模（ノード数が数百〜数千）であり、行列ベースでは計算困難な場合に効果的である。

一方、筆者が本稿において対象にしているのは、比較的小規模なグラフであり、与えられたグラフの隣接行列情報のみからノード間の類似度を定義し、それ以降は通常の行列ベースの手続きによりクラスタリングをおこなおうとするものである。将来的には、筆者らが以前、教育情報工学の分野において開発した、グラフのノードの見やすい配置法（CS系列法）[5]との結合を目指している。

A Study on Defining Similarities among Nodes of Digraph
Hiroyasu CHIMURA
C&C Information Technology Research Laboratories,
NEC Corporation

3．可到達行列の1／0パターンに着目したノード間類似度

3.1 対象グラフ

今、対象とする有向グラフを $G=\langle V,E\rangle$ とする。ただし、Vはノードの集合、Eはアークの集合である。また、

n…グラフGのノード数．従って、$V=\{v_1,v_2,\cdots v_n\}$．

$A=\{a_{ij}\}$…グラフGの隣接行列．

$a_{ij}=0$または1 $(i,j=1,2,\cdots,n)$

$R=\{r_{ij}\}$…グラフGの可到達行列．

$r_{ij}=0$または1 $(i,j=1,2,\cdots,n)$

$(A^{k+1}=A^k=R$ である）

とする。

3.2 基本的考え方

有向グラフの可到達行列は、個々のノードの全体の中での位置づけを潜在的に表している。そこで、可到達行列を利用してノード間の類似度を定義することを考える。すなわち、任意の2つのノードの可到達行列の行ベクトルまたは列ベクトルの1／0パターンを比較し、パターン的類似性を数値化してこれを当該ノード対の類似度と定義するのである。行ベクトル、列ベクトルのいずれを採用するかは、その意味するところが異なるので別途考察する。

3.3 類似度の定義

ノードv_iの可到達行列の行ベクトルまたは列ベクトルを、v_iの反応パターンと呼び、

$v_i=(v_{i1},v_{i2},\cdots,v_{in})$，$v_{ij}=0$または1

とする(図1)。

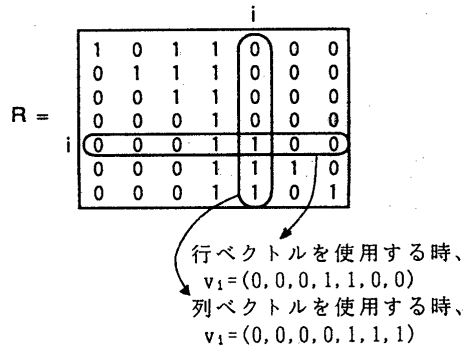

図1　ノードの反応パターン

また、反応パターンv_i、v_jのクロステーブルにおける各出現数a,b,c,dを次のようにする。

$$a=\sum_k v_{ik}\cdot v_{jk}$$
$$b=\sum_k v_{ik}\cdot(1-v_{jk})$$
$$c=\sum_k (1-v_{ik})\cdot v_{jk}$$
$$d=\sum_k (1-v_{ik})\cdot(1-v_{jk})$$
$$(a+b+c+d=n)$$

v_i \ v_j	1	0
1	a	b
0	c	d

0／1型データに対する類似度の定義はいくつか提案されており([6]など)、例えば以下のものがある。

① 一致係数 (the simple matching coefficient)
$M_{ij}=(a+d)/n \quad (0\leqq M_{ij}\leqq 1)$

② 類似比 (the coefficient of Jaccard)
$J_{ij}=a/(a+b+c) \quad (0\leqq J_{ij}\leqq 1)$

③ ϕ-係数
$\phi_{ij}=(ad-bc)/\sqrt{(a+b)(c+d)(a+c)(b+d)}$
$(-1\leqq \phi_{ij}\leqq 1)$

④ Loevinger の等質性係数
$H_{ij}=(ad-bc)/\{(a+c)(c+d)\}$

以上の４種の適用を予定しているが、本稿では①②の係数について検討する。

3.4 一致係数M_{ij}と類似比J_{ij}の検討

有向グラフのノード間類似度が最低限満たすべきと考えられる下記の４条件について、M_{ij},J_{ij}がこれら条件を満たすか否かを検討する。なお、ここでは、M_{ij}とJ_{ij}を共通にS_{ij}と記す。

【条件A】強連結ノードの一致性
強連結の関係にあるノード対 v_i,v_jについて、$S_{ij}=1$ となること。

【条件B】非連結グラフの分離性
異なる構造に含まれるノード対 v_i,v_jについて、$S_{ij}=0$ となること（図２）。

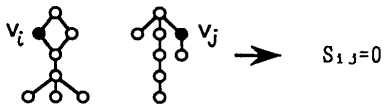

図２　非連結グラフの分離性

【条件C】最短経路長との整合性
任意のノード v_i,v_jの最短経路長を$L(v_i,v_j)$とした時、$L(v_i,v_j)<L(v_i,v_k)$ ならば $S_{ij}>S_{ik}$ が成り立つこと（図３）。

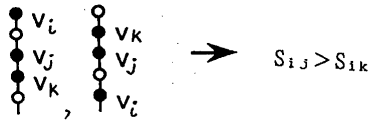

図３　最短経路長との整合性

【条件D】同一条件下の無矛盾性
ノード対v_i,v_jと別のノード対v_k,v_ℓとがグラフ上で同一の条件下にある時、$S_{ij}=S_{k\ell}$ となること（図４）。

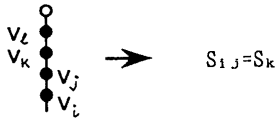

図４　同一条件下の無矛盾性

検討結果を表１にまとめる。M_{ij},J_{ij}とも一長一短であり、今後の検討を要す。

表１　検討結果
（○：条件を満たす　×：満たさない）

	一致係数 M_{ij}		類似比 J_{ij}	
	行ﾍﾞｸﾄﾙ	列ﾍﾞｸﾄﾙ	行ﾍﾞｸﾄﾙ	列ﾍﾞｸﾄﾙ
条件A	○	○	○	○
条件B	×(注1)	×(注2)	○	○
条件C	○	○	○	○
条件D	○	○	×	×

(注1)vi,vjのベクトル要素中、0,0でマッチしてしまう部分があるため、Mij=0とはならない。　特に、vi,vjが階層の上のノードである程Mijは大きくなる。
(注2)同上。特に、vi,vjが階層の下のノードである程Mijは大きくなる。

４．おわりに

本稿では、有向グラフのノードクラスタリングのための、可到達行列を利用したノード間類似度定義に関する基礎的検討を行った。今後は、3.4で述べた条件以外の観点からも検討を進める予定である。

最後に、しばらく保留となっていた本研究を再開するにあたって、様々な刺激を与えていただいた、日本電気C&C情報研究所情報応用研究部 原良憲氏、Rodrigo Botafogo氏、小川隆一氏を始めとする諸氏に感謝します。

参考文献

[1]Warfield,J.N.(1976) “Societal Systems” John Wiley & Sons, 204-284.
[2]Botafogo,R.A. & Shneiderman,B.(1991) “Identifying Aggregates in Hypertext Structures” Proceedings of the Hypertext91 Conf.,63-74.
[3]Hara,Y., Keller,A.M. & Wiederhold,G.(1991) “Implementing Hypertext Database Relationships through Aggregations and Exceptions” Proceedings of the Hypertext91 Conf.,75-90.
[4]佐古･原･千村(1992) “参照関係を用いた大規模マニュアルの自動構造化” 情報処理学会 情報学基礎研究会技報、FI-26-2、9-16.
[5]千村･佐藤(1980) “ＩＳＭ教材構造化法における教材要素の配置アルゴリズム” 電子通信学会 教育技術研究会技報、ET80-9、5-10.
[6]奥野忠一･他(1971) “多変量解析法” 日科技連、393-395.