統計的テキスト解析（6）
～ 語のネットワーク分析 ～

同志社大学文化情報学部教授
金　明哲（Jin Mingzhe）

■中国生まれ。総合研究大学院大学数物研究科統計科学専攻博士後期課程修了。博士（学術）。1995年札幌学院大学社会情報学部、助教授、教授を経て、2005年４月より現職。E-mail：mjin@mail.doshisha.ac.jp

1. ネットワーク分析とは

ネットワーク分析は、社会学や通信ネットワークなどの分野で多く用いられており、数学のグラフ（Graph）理論に基礎を置いている。したがって、分野によってはグラフ分析とも呼ぶ。ネットワークは、頂点（V：Vertex）と辺（E：Edge）を基本構成要素とする。頂点を「点」「ノード」、辺を「線」とも呼ぶ。ネットワークは、線で点と点の関係を示す。線が方向性を持つグラフを有向グラフ（Directed Graph）、方向性を持たないグラフを無向グラフ（Undirected Graph）と呼ぶ。図１に、有向グラフと無向グラフの例を示す。

ネットワーク分析では、図１(a)の関連性を表１(a)のように、1, 0で示すデータ形式を用いるのが一般的である。図１(a)は有向グラフであるので、非対称である。無向グラフ図１(b)は、表１(b)のような行列で表すことができる。表１(b)は対称行列である。表１のデータを隣接行列と呼ぶ。

ネットワーク分析では、ネットワークマップ、グラフの構造に関する指標と統計量などを用いる。

ネットワーク分析のフリーツールとしては、Graphviz, Pajek, NetDrawなどがある。Rには、ネットワーク分析のパッケージとして、sna, network, graph, igraph, inetworkなどがある。

図１　グラフ例

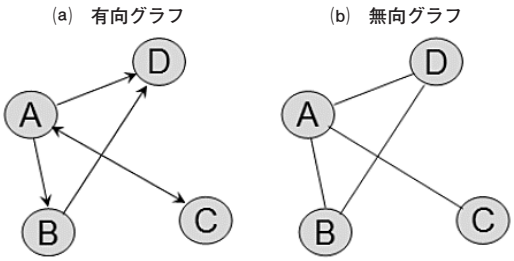


表１　隣接行列

(a)

	A	B	C	D
A	0	1	1	1
B	0	0	0	1
C	1	0	0	0
D	0	0	0	0

(b)

	A	B	C	D
A	0	1	1	1
B	1	0	0	1
C	1	0	0	0
D	1	1	0	0

本稿では、Rのパッケージの中で比較的にグラフ操作が便利であるigraphを用いることにする。パッケージigraphは、CRANミラーサイトからダウンロードすることができる。パッケージigraphの中には、表1のような隣接行列をネットワークマップ作成用のデータに変換する関数**graph.adjacency**がある。表1(a)のデータマトリクスを作成し、ネットワークマップ用のデータを作成するコマンドを次に示す。

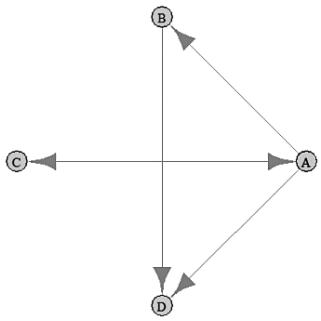
```
>library(igraph)
>test<-matrix(c(0,0,1,0, 1,0,0,0,
1,0,0,0, 1,1,0,0),4,4)
>(test.g<-graph.adjacency(test))
Vertices: 4
Edges: 5
Directed: TRUE
Edges:

[0] 0 -> 1
[1] 0 -> 2
[2] 0 -> 3
[3] 1 -> 3
[4] 2 -> 0
```

返された結果からわかるように、関数graph.adjacencyがデフォルトのまま作成するのは有向グラフのデータである。ノードにラベルを付け、関数**plot**でネットワークマップを作成するコマンドを次に示し、その結果を図2に示す。

関数plotは、plot.igraphの略である。ネットワークマップのレイアウトには、いくつかのオプションがある。ここでは、点（ノード）を円として配置するレイアウト（layout.circle）を用いる。

```
>V(test.g)$name<-c("A","B","C","D")
>plot(test.g,vertex.label=
V(test.g)$name,layout=layout.circle)
```

図2　igraphの有向グラフ例



ネットワーク分析は、多くの統計量と指標が考案されている。それを詳細に述べる誌面がないので、最も基本となる密度と中心性の指標について簡潔に紹介する。

密度（Density）は、ネットワークに含まれる関係の複雑さを示す度合であり、有向グラフでは $density = \frac{m}{n(n-1)}$、無向グラフでは $density = \frac{2m}{n(n-1)}$ と定義されている。式の中の n は点の数、m は辺の数である。式からわかるように、辺の数が多いほど密度の値が大きくなるので、グラフが複雑になる。

中心性に関しては、次数中心性、接近中心性、固有ベクトル中心性、媒介中心性などがある。

次数（Degree）は、ある点に接している辺の数である。接近中心性（Closeness Centrality）は、グラフの中のある点が他の点に対する近さに関する指標であり、次の式で定義されている。式の中の $d(n_i, n_j)$ は、点 n_i と点 n_j の最短距離であり、g はグラフに含まれた点の数である。接近中心性は、値が小さいほど中心性が高い。

$$closeness = \frac{g-1}{\sum_{all\ i,j}^{i \neq j} d(n_i, n_j)}$$

固有ベクトル中心性(Eigenvector Centrality)は、隣接行列の第1固有ベクトルを用いて、隣接する点の中心性を表す指標である。固有ベクトル中心性は、値が大きいほど中心性が高いことを意味する。ただし、隣接行列が非対称の場合は、注意が必要である。

媒介中心性（Betweenness Centrality）は、そのノードを通過しないと他のノードに到達できない度合、つまり、ある点がその他の2点を結ぶ最短経路である度合であり、値が大きいほど中心性が高い。

表2にパッケージigraphに用意された上記の指標の関数を示す。

表2　igraphに含まれる指標の関数

指標	関数
密度	graph.density
次数	degree
接近性	closeness
固有ベクトル	evcent
媒介性	betweenness

2. 語のネットワーク分析

テキストマイニングを行う際は、語の共起（共出現）パターンは、1つの重要な情報となる。語の共起とは、n-gramを含む広い意味での、語が文あるいはテキストの中に、同時に用いられていることを指す。

語のネットワークマップとは、基本的には、文あるいはテキストの中で用いられた語をノードとし、同時に用いられた場合は、語と語を線で（辺として）リンクしたグラフである。共起パターンに前後の関係がある場合は有向グラフ、そうではない場合は無向グラフになる。また、共起パターンの多少に関する重みを線の太さや長さで示す工夫も行われている。

言語学においては、コーパスから共起パターンを抽出し、語学教育に取り組む試みが行われ、良い反響が現れている。その例としては、NHKテレビが、2008年4月から再放送を行っている「アンコール㊐3か月トピック英会話 英単語ネットワーク」がある。本番組は、コーパスの中の単語が共起するネットワークマップを用いて、語彙の全体像から単語を覚えることを目指している。

近年、市販のテキストマイニングツールは、語の共起関係をネットワークマップで示す機能を備えるようになっている。

福田総理の所信表明演説文を形態素解析し、名詞のbigramを出現頻度が高い順に並べ替えた共起データを表3に示す。bigramは特殊な共起パターンである。

このデータは、「茶筌」で形態素解析を行い、フリーソフトMLTPを用いて、名詞に限定してbigramを求め、その結果を整形したものである。

パッケージigraphの中には、表3のデータ

表3　名詞のbigram

前の語	後の語	度数
国民	皆様	8
安全	安心	7
国際	社会	7
持続	可能	5
⋮	⋮	⋮
可能	社会	4
環境	問題	3
行政	信頼	3
政治	資金	3
⋮	⋮	⋮

をネットワークマップデータに置き換える関数**graph.data.frame**がある。

表３のデータがfukudabi.csvというファイル名でｃドライブの中に保存されたとする。

```
>fukudaNbi<-read.csv("c:/fukudabi.csv", head=F)
>fukudaNbi[86:87,]
        V1       V2   V3
86    信頼       関係    2
87    強化    アジア     1
```

ここでは共起頻度が２以上のものを用いることにする。まず、関数graph.data.frameを用いてネットワークマップデータを作成し、次に、関数**tkplot**を用いてネットワークマップを作成する。コマンドの書式を次に示す。

```
>library(igraph)
>wng<-graph.data.frame(fukudaNbi[1:86,])
>tkplot(wng,vertex.label=V(fwng)$name)
<図は省略>
```

パッケージRMeCabの作成者石田基広氏（徳島大学総合科学部）が本稿に間に合うように作成した関数**NgramDF**を用いると、テキストを形態素解析し、表３のように、隣接している語の共起パターンを集計することができる。その使用例を次に示す。

福田総理の所信表明演説文のテキストファイルをｃドライブの中のAFGenbunの中に保存したとする。まず、データの場所を指定し、次に、関数NgramDFを用いてMeCabによる形態素解析を行った結果から、名詞のbigramを集計する。関数NgramDFは、形態素解析とn-gramの集計を一体化したものである。引数type＝1は単語（形態素）の集計、N＝2はbigramの指定、posは品詞を指定する。

```
>targetText<- "c:/AFGenbun/福田所信.txt"
>kekkaDF <- NgramDF(targetText, type = 1, N = 2,
pos="名詞")
>head(kekkaDF)
  Ngram1  Ngram2  Freq
1      .      25     1
2      1       .     1
3      1      兆     1
4      1      万     1
5      1      目     1
6     10      月     1
```

集計した結果を度数（Freq）の降順に、次のように並べ替える。

```
>sortlist<-order(kekkaDF[,3],decreasing = TRUE)
>fwn<-kekkaDF[sortlist,]
>fwn[87:88,]
        Ngram1    Ngram2    Freq
1280      立場      行政       2
1          200        年       1
```

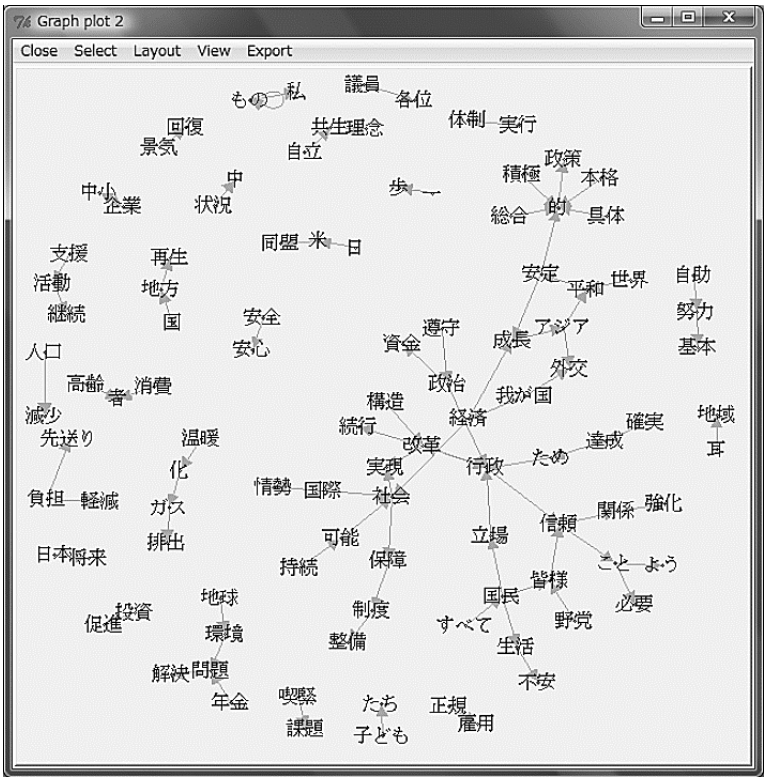
度数が２以上のものを用いてネットワークマップを作成することにする。そのコマンドを次に示し、作成されたネットワークマップを図３に示す。

```
>wng<-graph.data.frame(fwn[1:87,])
>tkplot(wng,vertex.label=V(fwng)$name, layout=
layout.fruchterman.reingold,vertex.size=1)
```

経験上、ネットワークマップの全体を見通すには、レイアウトのオプションをlayout.fruchterman.reingoldにすることをすすめる。

レイアウトのオプションReingold-Tilfordを用いると、ネットワークマップ上の最大サブネットワークのみを示すことができる。また、関数tkplotで作成したマップは、レイアウトを変更することによりノードを再配置することができる。次のコマンドのようにノードのサイズを大きくし（vertex.size＝20）、作成したネットワークマップのメニュー「Layout」からReingold-Tilfordを選択して、ノードを配置し直したグラフの画面コピーを図４に示す。

図３　福田総理の所信表明演説文における名詞の共起ネットワークマップ(1)




```
>tkplot(wng,vertex.label=V(fwng)$name, vertex.
size=20)
```

語のネットワークマップから、文章の中で強調された語句やそれらの関連性を推測することができる。意味分析を行う際は、形態素が長い単位を取るのがよいであろう。

このような方法は、アンケート調査の自由回答文の分析、商品のイメージ調査、コールセンターの質問データの自動分析などに広く用いられている。

語の共起関係のデータが表１のような隣接行列形式の場合は、関数graph.adjacencyを用いてネットワークマップ用のデータに変換することができる。

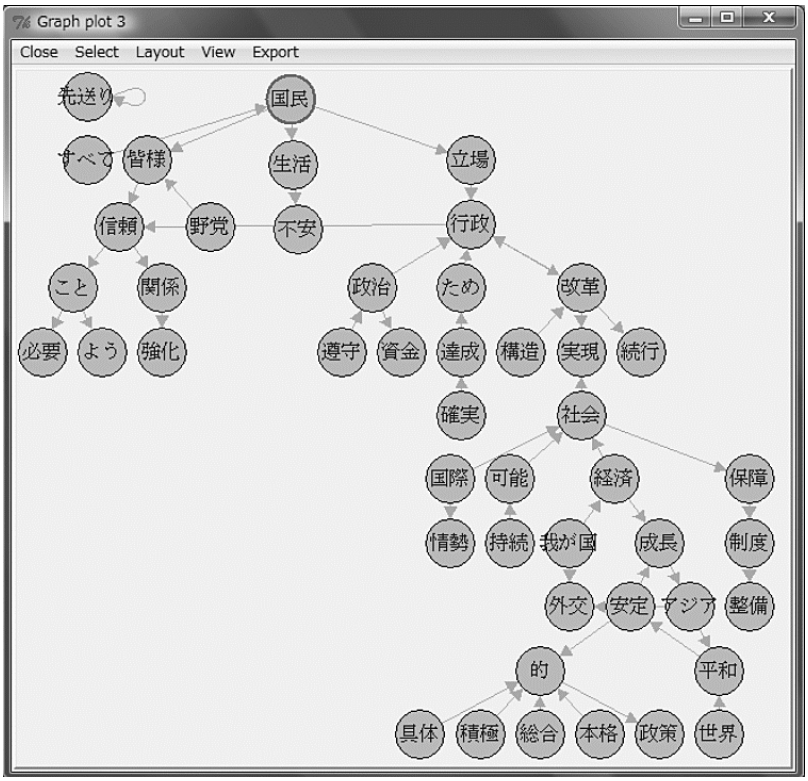
また、パッケージnetworkの中の関数**network**を用いると、表４のようなデータについてもネットワークマップを作成することができる。

表４　ドキュメントと語の行列

	国民	生活	安心	安全
d_1	1	0	1	1
d_2	1	1	1	0
d_3	0	1	1	1
d_4	0	1	1	1
d_5	0	1	1	0

表４のデータを用いてネットワークマップを作成するコマンドを次に示し、その結果を図５に示す。

図４　福田総理の所信表明演説文における名詞の共起ネットワークマップ(2)

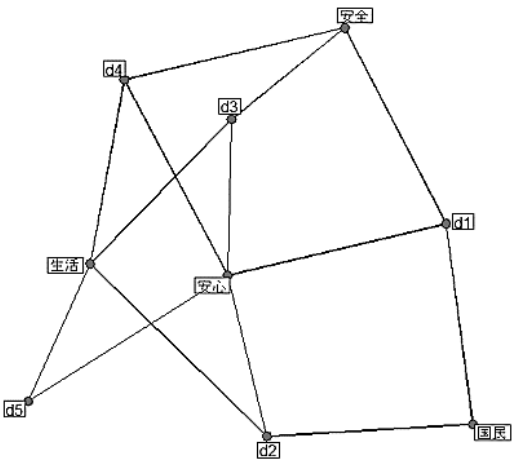



```
>test<-matrix(c(1,1,0,0,0, 0,1,1,1,1, 1,1,1,1,1,
1,0,1,1,0),5,4)
>colnames(test)<-c("国民","生活", + "安心","安全")
>rownames(test)<-c("d1","d2","d3", "d4","d5")
>library(network)
>test.ne<-network(test)
>plot(test.ne,displaylabels = TRUE)
```

TF-IDFや共起パターンの特徴の度合を算出し、語のネットワーク分析を行う試みも報告されている。

誌面の都合により、安倍元総理の所信表明演説文の分析、次数や中心性などの指標を用いた比較分析について議論ができないのが残念である。

図５　表４のネットワークマップ



謝辞
ご多忙の中、本稿に間に合わせてRMeCabの関数を作成していただいた徳島大学総合科学部の石田基広氏に心から感謝申し上げます。