近代文語のテキストマイニング

鈴木　努

2010 年 12 月 19 日
概念史・概念分析研究会

1　近代文語とテキストマイニング

昨今ではテキストマイニングツールの精度向上や普及が進み、実務・研究での実用化も進んでいる。概念史研究にそれらのテキストマイニング手法を応用する場合、現代語と異なる語彙や文法が用いられているテクストをどう扱うかが問題となる。幸い、近代文語（明治期の普通文）については、形態素解析ソフトで利用可能な辞書が公開されているので、今回はそれを用いたテキストマイニングの例を紹介したい。

近代文語は現代語と共通の部分もあり、名詞の抽出くらいなら現代語の辞書を用いてもできなくはない。しかしその精度は現代語を対象とする場合と比べ低下してしまう。形態素解析ソフト MeCab の Windows 版でデフォルトの IPA 辞書を用いると、例えば「欺くべからざるの確証...」の助動詞「ざる」が名詞と判定されたり、「即是れ人類の当に経過すべき階級なり」の接続詞「即」が接頭辞に、代名詞「是れ」の「れ」が助動詞に、副詞「当に」が名詞「当」と助詞「に」に判定されるということが起きてしまう[1]。

このようなことを避けるためにも、やはり近代文語には近代文語専用の辞書を用いるべきである。

2　近代文語 UniDic

近代文語 UniDic は、国立国語研究所が公開している形態素解析辞書 UniDic の近代文語版である [1]。辞書データとしても公開されているが、Windows 用[2]には解析用 GUI として近代茶まめ（図 1）というソフトが付いており、解析結果を得るならこれで十分である。

近代茶まめを利用するには別に形態素解析ソフト（ChaSen または MeCab）をインストールする必要がある[3]。今回は解析精度の高い MeCab を用いた。

[1] 文例は『文明論之概略』第二章より。
[2] Windows7 には対応していない模様。
[3] ChaSen は http://chasen-legacy.sourceforge.jp/から、MeCab は http://mecab.sourceforge.net/からダウンロードできる。

図 1: 近代茶まめのインターフェース

3 分析例

3.1 分析対象

分析例として、テキストデータが容易に準備できること、専門家による注釈書があり解析結果と比較できることから、福沢諭吉『文明論之概略』を用いる。ただし、今回は丸山真男『「文明論之概略」を読む 上』の第三講から第五講で論じられている第二章に限って分析する。テキストデータは上田修一・慶應義塾大学教授のウェブサイト[4]で公開されているものを使用した。

3.2 形態素解析

近代茶まめにテキストデータを入力する方法としては、1) テキストエリアに直接入力する、2) ファイルを指定する、3)URL を指定するの 3 つの方法がある。上記ウェブサイトからテキストデータを保存し（この段階で不要な部分は削除する）、テキストエリアにペーストする（今回はこの方法）か、ファイルを指定して解析する。

解析結果（表 1）をコピーし、テキストエディタか表計算ソフトにペースト、適当な名前をつけて保存する。今回の分析例ではエクセルにペーストして bunmeiron.ch2.csv というファイル名の CSV ファイルとして保存した。

[4]http://www.slis.keio.ac.jp/ ueda/

表 1: 近代茶まめの出力例

典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
chamame	B					空白				記号
chamame	I	前章	ゼンショー	ゼンショウ	前章	名詞-普通名詞-一般			ゼンショウ	漢
chamame	I	に	ニ	ニ	に	助詞-格助詞			ニ	和
chamame	I	事物	ジブツ	ジブツ	事物	名詞-普通名詞-一般			ジブツ	漢
chamame	I	の	ノ	ノ	の	助詞-格助詞			ノ	和
chamame	I	軽重	ケーチョー	ケイチョウ	軽重	名詞-普通名詞-一般			ケイチョウ	漢
chamame	I	是非	ゼヒ	ゼヒ	是非	名詞-普通名詞-サ変可能			ゼヒ	漢
chamame	I	は	ワ	ハ	は	助詞-係助詞			ハ	和
chamame	I	相	アイ	アイ	相	接頭辞			アイ	和
chamame	I	対し	タイシ	タイスル	対する	動詞-一般	文語サ行変格	連用形-一般	タイス	混
chamame	I	たる	タル	タリ	たり-完了	助動詞	文語助動詞-タリ-完了	連体形-一般	タリ	和
chamame	I	語	ゴ	ゴ	語	名詞-普通名詞-一般			ゴ	漢
chamame	I	なり	ナリ	ナリ	なり-断定	助動詞	文語助動詞-ナリ-断定	終止形-一般	ナリ	和
chamame	I	と	ト	ト	と	助詞-格助詞			ト	和
chamame	I	云え	イエ	イウ	言う	動詞-一般	文語四段-ハ行	命令形	イウ	和
chamame	I	り	リ	リ	り	助動詞	文語助動詞-リ	終止形-一般	リ	和
chamame	I	。			。	補助記号-句点				記号

3.3 R による分析

形態素解析の結果をもとに、名詞間の共起関係を分析する。分析にはデータ解析ソフト R を使用する。R は CRAN のサイト[5]からダウンロードできる。

R で先に保存した CSV ファイルを読み込む。あらかじめデータを保存したディレクトリに移動しておく。以下は C ドライブ直下に保存した例。

```
> setwd("C:/")
> data <- as.matrix(read.csv("bunmeiron.ch2.csv"))
```

データの行数を確認

```
> nrow(data)
[1] 12518
```

今回分析に用いるのは名詞だけなので、使いたい名詞の種類を列挙した meishi というオブジェクトを作って、それ以外の行を削除する。ただし、文境界の印として句点は残しておく。

```
> meishi <- c("名詞-固有名詞-一般","名詞-固有名詞-人名-一般","名詞-固有名詞-人名-姓",
+ "名詞-固有名詞-人名-名","名詞-固有名詞-地名-一般","名詞-固有名詞-地名-国",
+ "名詞-普通名詞-サ変可能","名詞-普通名詞-サ変形状詞可能","名詞-普通名詞-一般",
+ "名詞-普通名詞-形状詞可能")
> #分析対象の名詞と句点だけ残す
> data <- data[which(is.element(data[,"品詞"],c(meishi,"補助記号-句点"))),]
> (rn <- nrow(data))#data の行数
[1] 3315
> (n <- length(which(data[,"品詞"] == "補助記号-句点")))#句点の数＝文の数
[1] 411
```

411 文あることが分かったので、文ごとにどの名詞が生起したかというリスト（wordslist）

[5]http://cran.r-project.org/

を作る。また、それらの名詞から重複を取り除いた名詞一覧（words）を作る。

```
> wordslist <- vector("list",n)
> k <- 1
> for (i in 1:rn) {
+   if (data[i,"品詞"] == "補助記号-句点") {
+     k <- k + 1
+   }
+   else {
+   wordslist[[k]] <- c(wordslist[[k]],data[i,"語彙素"])
+   }
+ }
> words <- unique(unlist(wordslist))
> (wn <- length(words))#名詞の数
[1] 886
```

生起リストから生起行列を作成する。

```
> oc <- matrix(0,n,wn) #生起行列
> for (i in 1:n) {
+   for(j in 1:length(wordslist[[i]])) {
+     oc[i,which(wordslist[[i]][j] == words)] <- 1
+   }
+ }
```

語の生起頻度を求めて、ソートする。

```
> freq <- colSums(oc) #生起頻度
> FREQUENCY <- data.frame(NOUN <- words, FREQ <- freq)
> freq.order <- order(freq, decreasing = TRUE)
> FREQUENCY[freq.order,] #生起頻度のソート
（出力は省略）
```

不要な語や不適切な解析結果をチェック。ここでは頻度上位 10 語のみ例示。

```
> FREQUENCY[freq.order,][1:10,]
    NOUN....words FREQ....freq
9              物           98
37             事           75
6             文明           63
180           国体           49
34             者           44
106            人           39
29            人民           35
22             国           29
28            西洋           29
19            日本           25
```

「物」「事」などは特定の意味をもたないので、分析に用いても意味がない。これらの排除語を指定して生起行列、名詞一覧から取り除く。解析ミスにより生じた不要なものもここで削除。

```
> excep.words <- c("物","事","者","共","只","右","常","ション","方")
> oc <- oc[,-which(is.element(words,excep.words))] #排除語を除外
> words <- words[-which(is.element(words,excep.words))]
```

表記をチェックし、適宜置き換える。

```
> replace <- matrix(c(
+ "シコウ", "始皇",
+ "ポリチカル・レジチメ", "ポリチカル・レジチメーション",
+ "レジチメ", "レジチメーション",
+ "ショウトク", "聖徳",
+ "ホウジョウ", "北条",
+ "ア", "亜",
+ "カズヨシ", "和好",
+ "ソウ", "宋",
+ "フジワラ", "藤原",
+ "ゲンジ", "源氏",
+ "サンヨウ",  "山陽",
+ "マサシゲ",  "正成",
+ "タカウジ",  "尊氏",
+ "クスノキ",  "楠",
+ "アシカガ",  "足利",
+ "ヨリトモ",  "頼朝",
+ "カマクラ",  "鎌倉"),
+ ncol = 2,byrow = TRUE)
> for (i in 1:nrow(replace)) {
+ words[words == replace[i,1]] <- replace[i,2]
+ }
```

同義語をまとめる。ここでは「イギリス」「英」を「英国」に、「亜」を「アジア」に含めることにする。それぞれの位置を確認する。

```
> which(words == "英国")
[1] 554
> which(words == "英")
[1] 511
> which(words == "イギリス")
[1] 841
> which(words == "アジア")
[1] 19
> which(words == "亜")
[1] 254
```

足し合わせてから、いらない行を削除。

```
> oc[,554] <- oc[,554] + oc[,511] + oc[,841]
> oc[,19] <- oc[,19] + oc[,254]
> oc <- oc[, -c(511,841,254)]
> words <- words[-c(511,841,254)]
```

生起行列から、単語間の共起頻度を表す行列を作成する。これは概念ネットワークにおける隣接行列となる。

```
> adj <- t(oc) %*% oc
> rownames(adj) <-  colnames(adj) <- words
> dim(adj) #隣接行列の次元
[1] 874 874
```

共起頻度が高い語の組み合わせを調べる。ここでは上位 10 組を示す。

```
> cooc <- matrix(NA,1,3)
> for (i in 1:(nrow(adj) - 1)) {
+   for (j in (i+1):nrow(adj)) {
+     if (adj[i,j] > 0) {
+       cooc <- rbind(cooc, c(rownames(adj)[i],rownames(adj)[j],adj[i,j]))
+     }
+   }
+ }
> cooc <- cooc[-1,]
> COOC <-
+ data.frame(N1 = cooc[,1], N2 = cooc[,2], FREQ = as.numeric(cooc[,3]))
> COOC[order(COOC$FREQ, decreasing = TRUE),][1:10,]
       N1          N2 FREQ
787  支那        日本   13
597  諸国        西洋   12
186  文明        西洋   11
248  文明        精神   11
208  文明          人    9
1047 半開        野蛮    9
6790 至尊        至強    9
169  文明 ヨーロッパ    8
180  文明          国    8
1110   国        国体    8
```

ここでもし 1 語として扱うべき語が分離しているのを見つけたら、統合するなどの処理をする。例えば、上では「諸国」と「西洋」の共起頻度が高いのは「西洋諸国」という熟語で用いられているためと考えられる。もとのテキストに戻ってみると、「西洋」単独の用法も見られるので、ここでは統合しないことにする。

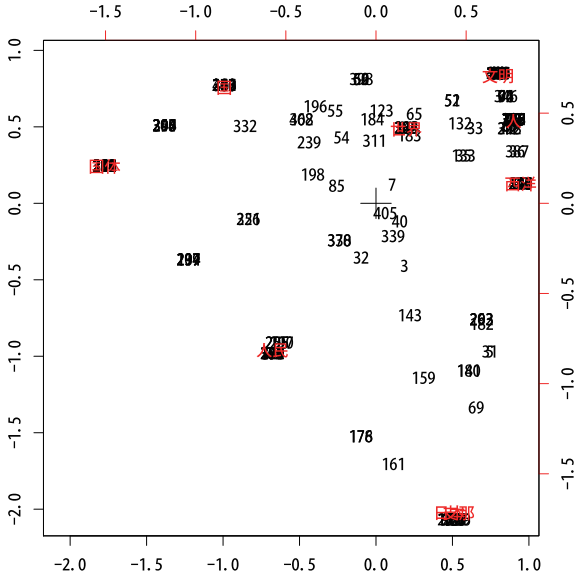

図 2: コレスポンデンス分析によるプロット

3.4 生起行列のコレスポンデンス分析

先に文を行、語（名詞）を列の項目とする生起行列（要素は 0 か 1 の二値）を作成した。この生起行列を用いてコレスポンデンス分析をすると、どのような語がいっしょに用いられる傾向があるのか、またそれはどの文においてか、といったことを可視化することができる。ただし、生起行列そのままでは行列が大きく、少ない次元にまとめるのが難しい。

```
> dim(oc)
[1] 411 874
```

411 文、874 語の生起行列のうち、生起頻度 20 以上の語と、それらの語が生起した文のみ用いる（本来はこれも探索的に決めることである）。

```
> oc20 <- oc
> oc20 <- oc20[,-which(colSums(oc20) < 20)]
> oc20 <- oc20[-which(rowSums(oc20) == 0),]
> dim(oc20)
[1] 197   9
> library(MASS)
> result <- corresp(oc20, nf = min(dim(oc20) - 1))
> biplot(result)#図 2
> eig <- result$cor^2
> round(100 * eig / sum(eig), 2)#寄与率
[1] 16.02 15.36 14.21 13.89 12.62 11.84 10.35  5.71
```

197 文、9 語による分析だが、軸の寄与率にもあまり差がなく、図も解釈しにくい。

3.5　概念ネットワーク

先の共起頻度行列を隣接行列として概念ネットワークを可視化する。ネットワークの作図には sna パッケージを用いる。sna パッケージはデフォルトではインストールされないので、別途インストールが必要である。

```
> install.packages("sna")
```

ここでも隣接行列をそのまま用いると煩雑になるので、共起頻度によるフィルタリングをした方がよい。以下は共起頻度 3 以上の関係のうち最大連結成分のみを示す例である。

```
> library(sna)
> adj3 <- adj
> adj3[adj3 < 3] <- 0
> lc.adj3 <- component.largest(adj3, result = "graph")
> gplot(lc.adj3, gmode = "graph",  displaylabels = TRUE)
```

結果は図 3 のようになる。ただし、見やすいように Illustrator で加工している。同様にしてフィルタリングの基準を変化させたのが図 4 から図 7 である。

3.6　『「文明論之概略」を読む』との比較

丸山真男の『「文明論之概略」を読む』では次の 3 つの章が『文明論之概略』第 2 章の論考に充てられている。

- 第三講　西洋文明の進歩とは何か
- 第四講　自由は多事争論の間に生ず
- 第五講　国体・政統・血統

それぞれを、文明論、自由論、国体論とすると丸山はこの 3 点を第二章の要点と捉えたことになる。図 3 を見ると、「文明」と「国体」が中心的な概念となっている。これは図 5 を見るとより顕著である。では図 3 で「自由論」に当たるのはどこかというと、「人民」から 1 本の線のように伸びた「気風」「自由」「争論」「異説」という連なりである。これは「異説」の「争論」があるところに「自由」の「気風」が生じるという内容を表している。しかしこの連なりは図 4 以降ではなくなってしまう。これは丸山が第二章の中では、挿入的に論じられるにすぎない「自由」に、強い思い入れをもって論じたことの現れだろうか。その他の論点についても研究会で議論したい。

4　研究会での議論から

- 「ヨーロッパ」と「西洋」を統合したらどうなるか。

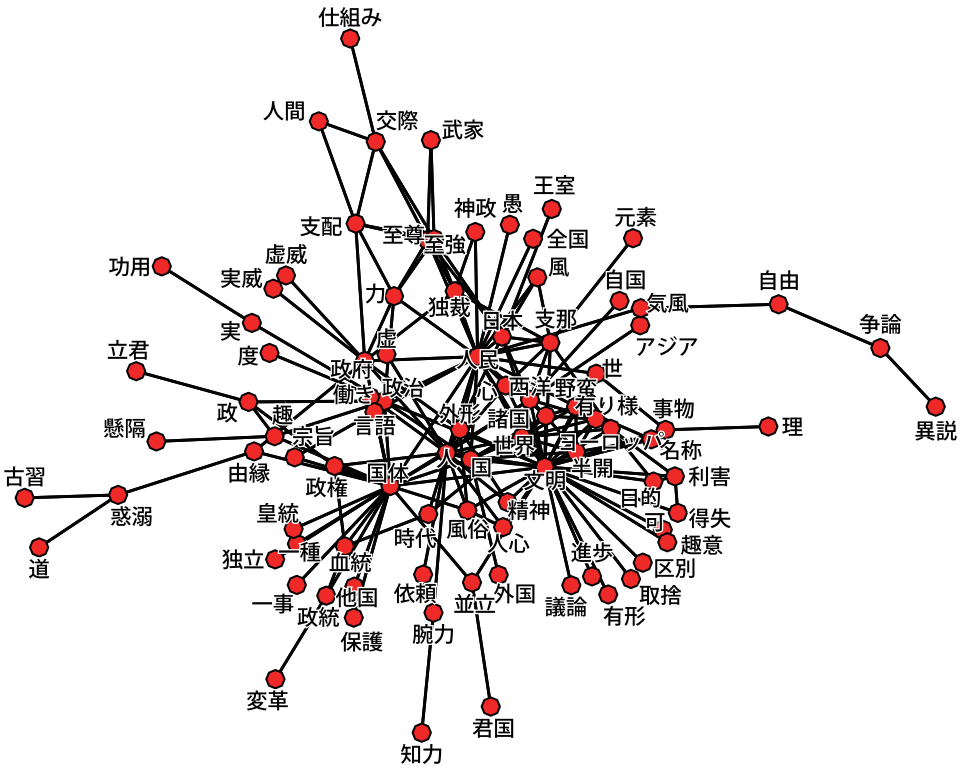

図 3: 共起頻度 3 以上の関係の最大連結成分

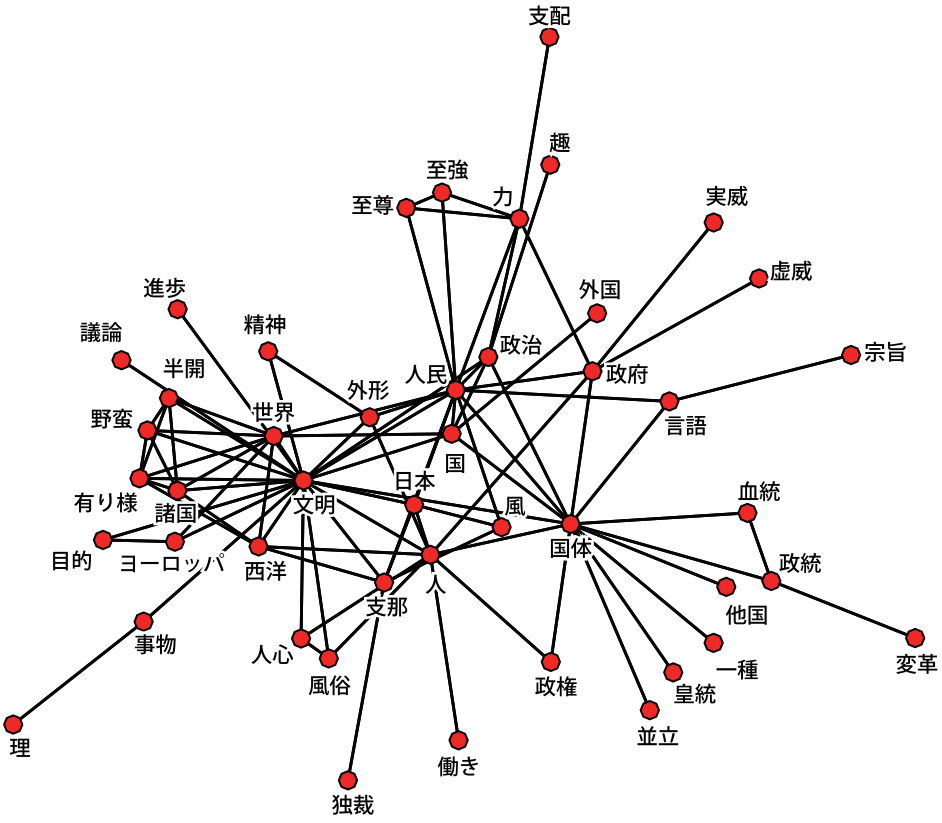

図 4: 共起頻度 4 以上の関係の最大連結成分

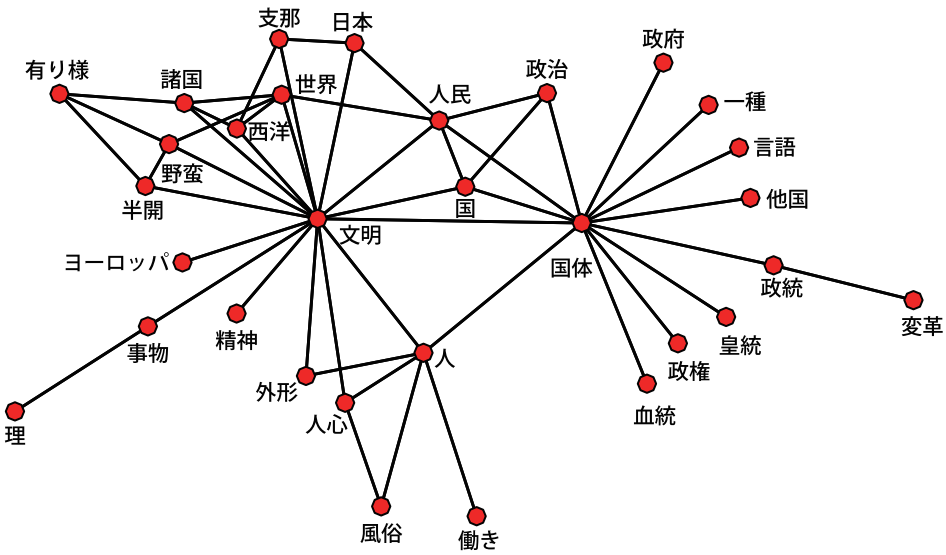

図 5: 共起頻度 5 以上の関係の最大連結成分

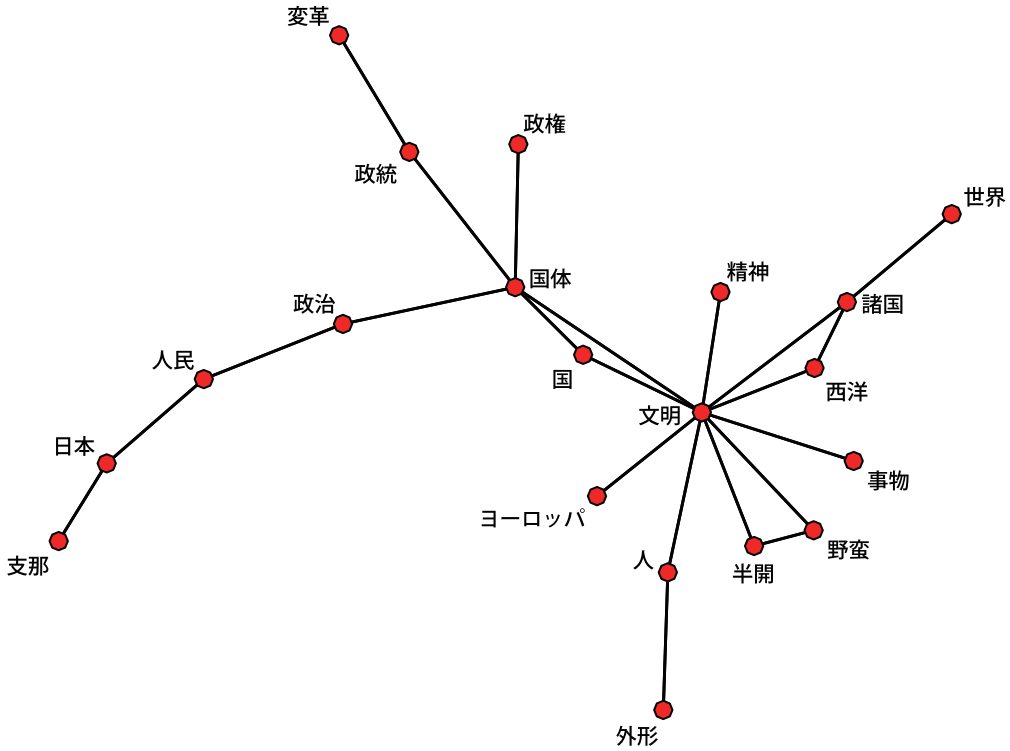

図 6: 共起頻度 6 以上の関係の最大連結成分

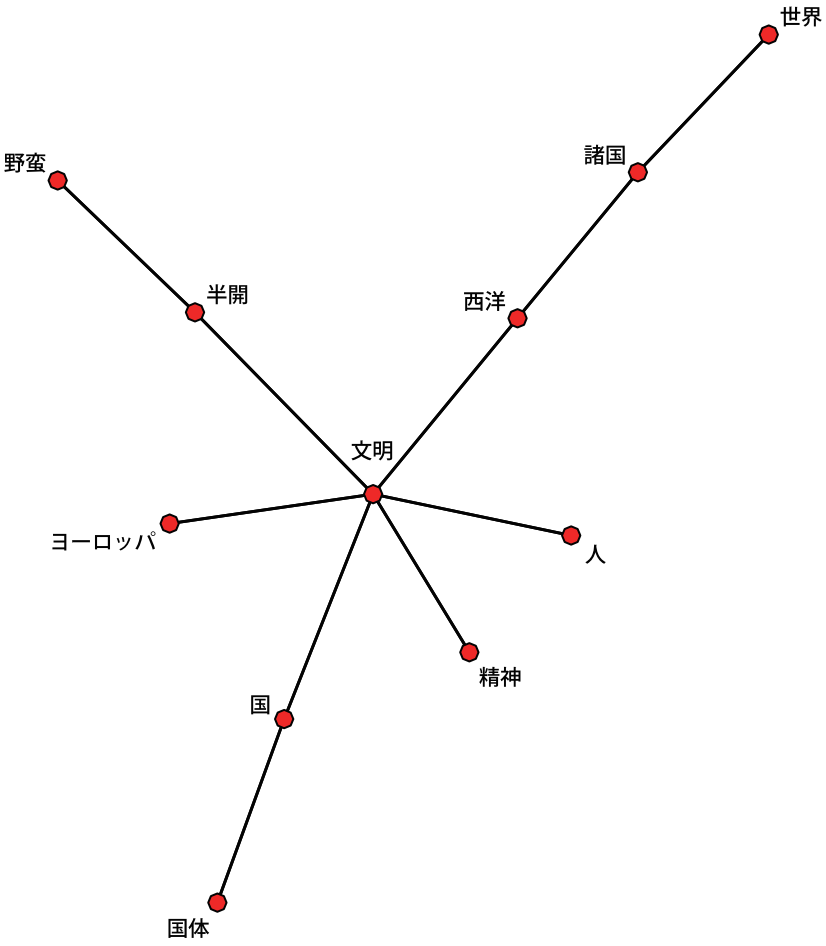

図 7: 共起頻度 7 以上の関係の最大連結成分

[補足] 二章末に「書中西洋と云い欧羅巴と云うもその義一なり。地理を記すには欧羅巴と亜米利加と区別あれども，文明を論ずるときは亜米利加の文明もその源は欧羅巴より移したるものなれば，欧羅巴の文明とは欧羅巴風の文明と云うの義のみ。西洋と云うもこれに同じ。」とあるので統合してもよいかもしれない。実際に「西洋」に統合すると、「西洋」の生起頻度や、他の語との共起頻度は当然大きくなるが図はほとんど変化しなかった。

- **テクストの解釈が複数あって論争があるようなテクストに適用してみたらどうか。同時代の同じようなテーマを論じた他の論者のテクスト（例えば田口卯吉の文明論など）と比較したらどうか。**

- **共起関係であっても、例えば「〜は文明である」と「〜は文明ではない」では意味が違うのではないか。**

  肯定的であれ否定的であれ、「文明」という概念の関わりで論じている点では同じであり、ここでの共起関係とはそのような「切り口」を表している。

- **図の頂点の位置に意味はあるのか。中心的な語が中心にくるようになっているのか。**

  sna パッケージの `gplot()` 関数はデフォルトでは、力学的アルゴリズムを用いて頂点を配置する。辺で直接つながった頂点どうしは近くに、直接つながっていない頂点どうしは離れて配置されるようになっている（ちょうどバネで引きあったり、電気力で斥けあったりするように)。結果として中心的な語が中心部にくる傾向が生じる。なお、力学的モデルの結果は一意ではない（つまり作図するたびに位置が変わる可能性がある。ただし、`gplot()` 関数で繰り返し同じ図を描くと位置が変わっていくのは、異なる力学的モデルを順に適用しているため)。

- **頂点の円を大きくしたりできないか。**

  例えば、以下のコードは共起頻度 5 以上のグラフで、頂点の大きさを生起頻度に比例させ、辺の太さを共起頻度に比例させる場合（図 8)。

```
> adj5 <- adj
> adj5[adj5 < 5] <- 0
> lc.adj5 <- component.largest(adj5, result = "graph")
> #語の頻度を計算
> vertex.freq <-
+   colSums(oc)[component.largest(adj5, result = "membership")]
> gplot(lc.adj5, gmode = "graph",displaylabels = TRUE,
+ label.pos = 5,vertex.sides = 20,
+ vertex.cex = vertex.freq/10,#頂点の大きさを頻度に比例させる
+ edge.lwd = lc.adj5)#辺の太さを共起頻度に比例させる
```

  ラベルの大きさを指定する引数は `label.cex`。詳しくは拙書 [4] を参照。

- **共起頻度によるフィルタリングを頂点の座標を変えずにできないか。**

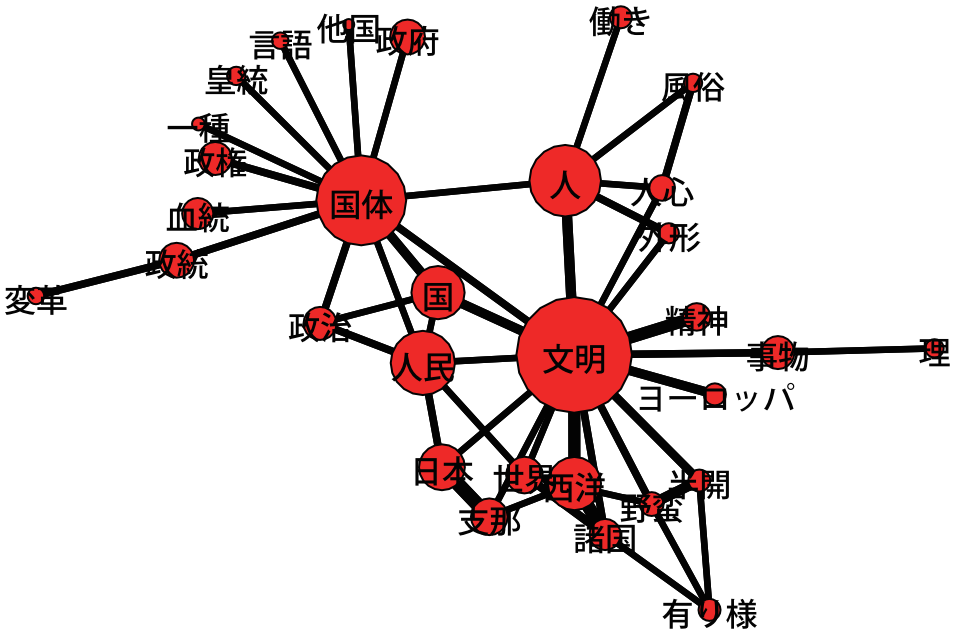

図 8: 頂点の大きさや辺の太さの変更

各頂点の座標が共通になるように指定することでできる。以下の例では、共起頻度によるフィルタリングの基準を 3 から 6 まで変化させる場合（図 9)。最初に共起頻度 3 以上のグラフで座標を決めておき、以後はその座標を使う。見やすいようにラベルは省略した。

```
adj3 <- adj; adj3[adj3 < 3] <- 0
lc.adj3 <- component.largest(adj3, result = "graph")
lc.adj4 <- lc.adj3; lc.adj4[lc.adj4 < 4] <- 0
lc.adj5 <- lc.adj3; lc.adj5[lc.adj5 < 5] <- 0
lc.adj6 <- lc.adj3; lc.adj6[lc.adj4 < 6] <- 0
zahyo <- gplot(lc.adj3)
oldpar <- par(no.readonly = TRUE)
par(mfrow=c(2, 2))
gplot(lc.adj3, gmode = "graph", coord = zahyo);box(col = "grey")
gplot(lc.adj4, gmode = "graph", coord = zahyo);box(col = "grey")
gplot(lc.adj5, gmode = "graph", coord = zahyo);box(col = "grey")
gplot(lc.adj6, gmode = "graph", coord = zahyo);box(col = "grey")
par(oldpar)
```

R によるテキストマイニングの参考書としては参考文献 [5, 6] もある。

参考文献

[1] 小木曽智信・小椋秀樹・近藤明日子，2008，「近代文語文を対象とした形態素解析辞書の開発」『言語処理学会第 14 回年次大会発表論文集』，225–228.（http://www.gakkai.ne.jp/jss/bulletin/guide4.php#sh4-4）

[2] 丸山真男，1986,『「文明論之概略」を読む 上』岩波書店.

[3] 石田基広，2008,『R によるテキストマイニング入門』森北出版.

[4] 鈴木努，2009,『R で学ぶデータサイエンス 8　ネットワーク分析』共立出版.

[5] 金明哲，2009,『テキストデータの統計科学入門』岩波書店.

[6] 松村真宏・三浦麻子，2009,『人文・社会科学のためのテキスト・マイニング』誠信書房.

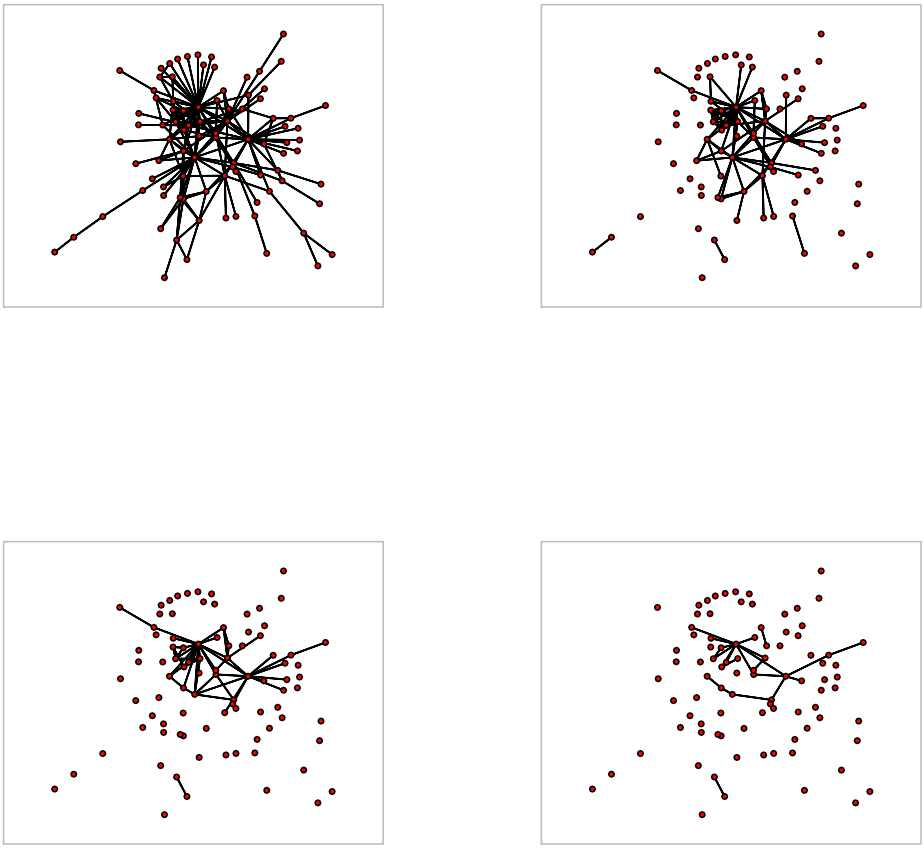

図 9: 頂点の座標を固定してフィルタリング