2007－CH－75 (9)
2007／7／27

# グラフクラスタリングを用いた文献解析の諸技法に関して

## --カバニスとメスメールのテキストを例に--

○赤間啓之，*三宅真紀，鄭在玲
{akama.h.aa, jung.j.aa}@m.titech.ac.jp, *mmiyake@lang.osaka-u.ac.jp
東京工業大学社会理工学研究科 *大阪大学言語文化研究科

赤間ら(2007)は、フランス革命期の思想家、カバニスとメスメールの思想的類似性を論証するため、彼らのテキストをもとに、キーワード中心の意味ネットワークを作成し、グラフクラスタリングの技法であるMCLを適用した。本論考では、単語の共起データを取る際、単語インスタンスをすべて取り扱う漸進ウィンドウ法(IAW)を使用すると、それに基づく意味ネットワークのクラスタリング結果が、キーワード中心のものと比べどう異なるものになるか分析する。

# On the Techniques of Document Analysis Using Graph Clustering

## --Taking the Examples from the Texts of Cabanis and Mesmer--

○Hiroyuki Akama, Maki Miyake*, Jaeyoung Jung
akama@dp.hum.titech.ac.jp
Department of Human System Science, Tokyo Institute of Technology
* Graduate School of Language and Culture, Osaka University

By using the MCL (Markov Cluster Algorithm) known as a graph clustering method, Akama et al. (2001) measured the similarity of thinking between two contemporary thinkers, Cabanis and Mesmer. But the previous data under the form of semantic network were obtained by selecting beforehand the keywords as hubs around which the neighboring words were taken as dangling vertices. This study propose as an alternative to the keyword-based clustering a new windowing method called Incrementally Advancing Window (IAW) that generates co-occurring word pairs that can be used as inputs to the Incremental Routing Algorithm. Here we compare these two types of co-occurrence and/or adjacency data matrix by applying to each of them the indexes as weighted curvature, modularity Q and F measure.

### 1. これまでの研究

赤間ら(2007)は近代初頭、晩期啓蒙主義の時代のフランスにおける、ストア主義とメスメリスムの、目に見えない思想的類似性を論証するため、ストア主義の代表的著作として

Georges Cabanis, Lettre à M.F. sur les causes premières

（ジョルジュ・カバニス、『第一原因についての F 氏への手紙』）を、さらにメスメリスムの代表的著作として、

F.-A. Mesmer, Mémoires de F.-A. Mesmer, docteur en médecine, sur ses découvertes

（F.-A.メスメール、『医学博士 F.-A.メスメールによる、彼の発見についての論文』）

の二つを選択し、そこで用いられた単語の共起データをグラフクラスタリングにかけることで、両者の影響関係を推定し、それを因子分析に基づく潜在的な意味構成の検出結果(赤間、2001)と比較した。

そこでは、まず上記の 2 文書（以降カバニスは C,

メスメールはMと略記する）を、長さを規準化しつつ合併連結し、単語数にして長さ10482個のテキスト(ノイズワードは除去）C&Mを生成した。さらに、C、Mのそれぞれ最頻出名詞上位50位まで計77語のキーワードを抽出したが、それらキーワードの各々のインスタンスを中心に、ウィンドウ幅5内で共起するすべての語をオブザベーションとして、出現頻度をカウントしている。

さらにそのようにして得られた共起行列をもとに、両文書の統合的な意味ネットワークを形成するため、単語が点ノードに、共起関係が辺に見立てられるグラフの隣接行列を算出した。この隣接行列に対しては、マルコフクラスター・アルゴリズム(MCL)という、グラフクラスタリングの手法が適用可能である。MCLにより、意味ネットワークは、いくつかの一貫した意味をもつサブグラフに分割され、類似語・同一系統語のグループを一個のクラスター（概念）にまとめることができた。

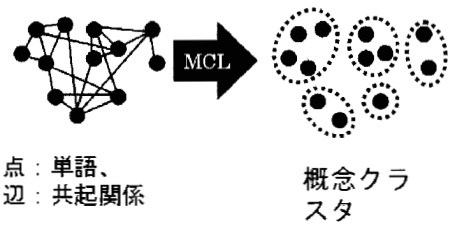


図1 MCLを意味ネットワークに適用する方法

その結果、Cを表す代表クラスター（サイズの大きいコアクラスター1）の中にMの最重要キーワードが見出され、さらに逆のケース、つまりMを表す代表クラスター（やはりサイズ大のコアクラスター2）の中にCの最重要キーワードが見出されるなど、両者がクロスする痕跡が明確に検出された。なお、MCLによる本データのクラスタリング結果は以下のとおりである。

まず、コアクラスターに関しては、

Mクラスター(コアクラスター):

{action(活動),air(大気),animal(動物),corps(身体),degré(程度), éther(エーテル),feu(火),fluide(流体),impression(印象),influence(影響),lumière(光),magnétisme(磁気),matière(物質),mécanisme(機制),modification(変更),mouvement(運動),nerf(神経),ordre(秩序),organe(臓器),organisation(組織),propriété(特性),sens(感覚),sensation(感覚),sensibilité(感受性),substance(実質)}が挙げられる。

ハブ（最大次数）は、action(活動)である。またこのメンバーのうち、sens:（感覚）、sensibilité(感受性)が、案に相違してCの特徴語であった。

もうひとつのコアクラスターとしては、

Cクラスター(コアクラスター):

{analogie(類推),besoin(欲求),critique(批判),esprit(精神),état(状態),être[01](存在),examen(検討),existence(存在),faculté(能力),fait[01](事実},force(力),habitude(習慣),homme(人間),hypothèse(仮説),individu(個人),intelligence(知性),loi(法),magnétisme-animal(動物磁気),moi[01](自我),morale(道徳),moyen(方法),nature(自然),observation(観察),phénomène(現象),point[01](点),principe(原則),question(問題),raison(理性),rapport(関係),résultat(結果),sentiment(感覚),sommeil(睡眠),source(源),système(システム),temps(時間),univers(宇宙),volonté(意志)}が生成する。

ハブ(最大次数)は homme(人間)である。これらのうち、magnétisme-animal（動物磁気）が案に相違してMの特徴語であった。ふたつのコアクラスター間には、このように両者間の影響を暗示するクロス現象が垣間見られた。

他の小規模なマルコフクラスターについては、

{cause(原因),puissance(力)},{crise(発作),maladie(病気)},{effet(効果)},{effort(努力)},{erreur(誤謬)},{expérience(経験)},{idée(観念)},{objet(対象)},{opinion(意見)},{partie(部分)},{théorie(理論)},{vertu(徳)},{vie(生命)}([01]は同綴意義区別用タグ)が挙げられる。

以上が本研究のこれまでの進捗であるが、そこではいくつかの重要な問題がいまだ取り扱われてはいない。それは、多少とも恣意性が残るキーワードの選択とウィンドウ幅の決定の問題である。特に前者については、グラフに固有な情報として、カヴァチャのようなグラフ指標が、同時に利用されることが望ましい。そこで本論考では、グラフ指標の導入から始めて、前提的キーワードを使用しない漸進ウィンドウ法 (Incremental Advancing Window、略して、IAW)に基づく意味ネットワーク形成の方法を紹介する。IAWは同時にウィンドウ幅をあらかじめ決める必要がなく、後で述べるModularity Qやそれを精度情報とするF尺度を用いて、最適な単語の取捨とウィンドウ幅を自動決定することができる。最後にIAWに基づく解析結果を先のキーワードありの解析結果と、グラフ指標などの面から比較する。

## 2. グラフ指標の活用

### 2.1. カヴァチャ

まず、ここで重要なグラフ指標について簡単に解説する。次数(degree)とは、一般にある点に連結する辺の総数を言う。カヴァチャ(Dorow, 2005)とは、ある点ノードに関する隣接点ノードどうしの結線率のことである。次数 m の点 i と隣接する m 個の点の間に n 本の辺が見出される、言い換えれば、点 i を頂点とする n 個の三角形が存在するならば、点 i のカヴァチャは $c_i = \frac{n}{{}_mC_2}$ で表される。この値は点 i の周りに星形グラフができる時 0、完全グラフが見出される時 1 の値を取る。辺の有無に基づいてカヴァチャを計算する場合、それはある種の相関係数のように、0 と 1 の間の値を取り、0 に近いほど、意味が多義的で 1 に近いほど意味が凝集的であることが自ずと理解される。

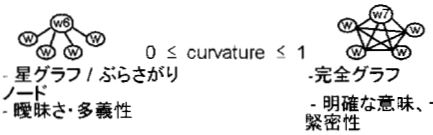


図 2 カヴァチャ

カヴァチャの分布は次数に依存するので、ある程度次数を考慮しないとカヴァチャ間の比較は難しい。特にキーワードを前提としたグラフの隣接行列の場合、キーワードを介さない共起語間の隣接はデータとして取らないので、非キーワード語の次数はひじょうに低くなり、カヴァチャの取りうる可能な値は限定され、高次数のキーワードのカヴァチャと単純に比較することはできない。一般に次数とカヴァチャの積の値はさほどのばらつきがなく、この積の平均からの偏差が大きな意味を持つ場合がある。

### 2.2. 重みつきカヴァチャ

このように、先の研究では、過去に行った因子分析とグラフクラスタリングの条件をそろえるため、キーワードの決定法としてグラフ指標をじゅうぶん活用しているとは言いがたい。さらにまた、ここでは単純なカヴァチャ値の導入だけではうまく意味ネットワークを解析できない、ひとつの制約条件が残されている。それはこの意味ネットワークが、文書の長さで基準化された頻度データをもとにした、「重みつきグラフ」であるということである。

先のカヴァチャの定義は、重みなしグラフにおける辺の有無をふまえたもので、辺の重みを考慮に入れる場合は、「重みつきカヴァチャ」を新たに定義する必要がある。定義の仕方はいろいろと考えられるが、ここでは、存在しない辺が取りうる潜在的な重みを 1 として計算することにする。すなわち、点 i に隣接する点の間に存在する辺の重み合計を $\sum w$ とすると、点 i の重み付きカヴァチャは $c_i = \frac{\sum w}{{}_mC_2}$ で表される。この値は、当然のことながら、1 より大きな値を取ることがありうる。さらに、存在しない辺が取りうる潜在的な重みは、恣意的に決定されるので、扱いに厳密さを欠いていないとは言い切れない。

## 3. 漸進ウィンドウ Incremental Advancing Window(IAW)

### 3.1. 方法

次に、本研究では、キーワードをあらかじめ選択せず、様々な条件下で意味関与語を自動的に取捨する方法として、漸進ウィンドウ Incremental Advancing Window(略して IAW)を提案する。

IAW は、ノイズワード、機能語のみを取り除いた単語インスタンスすべてをウィンドウの停止語として同等に取り扱う手法である。

これは通常のウィンドウィング法と同様、幅左右 n 語ずつに固定されたウィンドウを、文書の先頭から末尾まで、1 回だけスライドさせる。しかし共起情報に関しては、ウィンドウを、1 単語ずつ右にずらして（インクリメントして）ゆき、すべての単語インスタンスを 1 回だけ中心語として扱い、共起関係を見る。ウィンドウ停止状態の履歴からすでにカウントされた共起ペアを除外するため、以下のような数え上げ方をする。

つまりウィンドウ右端の単語からウィンドウ内の他の単語にそれぞれ伸ばしたパスのみを単語ペアとして追加する。ウィンドウ内の単語をそれぞれノードとする「完全グラフ」のうち、新しく捉えた右端の単語を root とする tree だけが新規に出現した共起ペアに対応するからである。

すなわち、幅左右 n ずつの時は、中心語 w(i)としてウィンドウの中の

[w(i-n),w(i-n-1),...,w(i),...,w(i+n-1),w(i+n)]

のうち、新規に

w(i-n)w(i+n),w(i-n-1)w(i+n),...,w(i+n-1)w(i+n)

をカウントする。そしてウィンドウの右端が文書の末端に到達したとき、ウィンドウィングを終了とする。

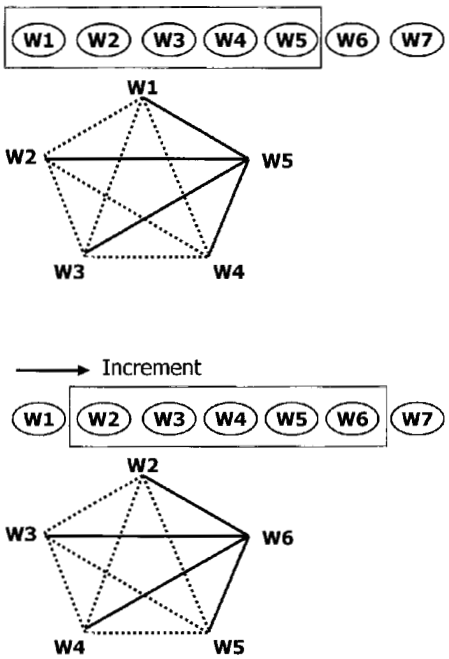


図 3. IAW におけるカウント法

この方法を使用する場合、同時にウィンドウ幅を様々に変えながら、グラフの元になる単語の共起ペアをその出現頻度とともに記録するとよい。

### 3.2. IAW の適用結果

キーワードにウィンドウの中心語を限定せず、すべての語を中心に置く場合は、ウィンドウ幅のみならず、ペア頻度閾値というもう一つのパラメーターを設ける必要が出てくる。それはウィンドウ幅によっては、低いペア頻度もグラフの辺として採用したとき、結線率を高くし、サブグラフ分割を困難にしてしまう結果になるからである。

本研究では、IAW を適用するパラメーター条件として、ウィンドウ幅 1~9、ペア閾値 1~9 とし、計 81 通りの条件を取りそろえた。IAW の適用により生成した 81 種の隣接行列をすべて個別に MCL にかけ、結果を各ペア頻度閾値ごとに、横軸をウィンドウ幅、縦軸を MCL クラスター数として折れ線グラフに表したものが、図４である。

このように、グラフクラスタリングに必要な、結線率というパラメーターの調整を、ウィンドウ幅や単語ペア頻度の閾値の調整と連動して行うのがこの技法の特徴である。これによって、統制された複数のローデータを取得し、比喩的に言うと、さまざまな視角と焦点距離で対象を観察することが可能になる。

ここで CM 文書の IAW データによる MCL の結果について、おおよその傾向をかいつまんで述べるならば、ペア頻度閾値が高いとすでに取り上げたキーワード中心の結果になることがわかる。しかし、ウィンドウ幅を広げていくと、準キーワードと言えるもの、たとえば動詞や形容詞も含むようになり、重要な骨組に実質的な肉付きが行われたという印象が生じる。逆にウィンドウ幅が狭いと、コロケーションの様な単語間の固定した緊密な関係が重点的に採取される。

またペア頻度閾値が低いとグラフの結線率が高くなり、その結果 MCL はクラスター分割を行えない場合があるとわかった。閾値 1 においては単語の最大異なり数 2262 を記録し、すでにウィンドウ幅 4 の段階で、MCL の結果は常に 1 クラスターのみであって分割されない。しかし、閾値が上昇すると、登場頻度の高い重要なペアがクローズアップされる。それらをもとに、やはり、キーワードつき MCL の結果が示すような、２大クラスター構成に近づいてゆくことがわかる。

ここでペア頻度閾値別に、ウィンドウ幅を変えたときの MCL クラスター数の推移を図 4 に示す。各折れ線は、ペア頻度閾値、横軸は、ウィンドウ幅、縦軸は、MCL クラスター数である。またウィンドウ幅別に、ペア頻度閾値を変えたときの MCL クラスター数の推移を図 5 に示す。各折れ線は、ウィンドウ幅、横軸は、ペア頻度閾値、縦軸は、MCL クラスター数である。

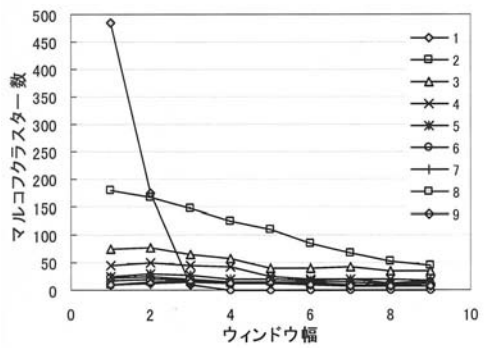


図 4. ウィンドウ幅-ペア頻度閾値によるクラスター数(1)

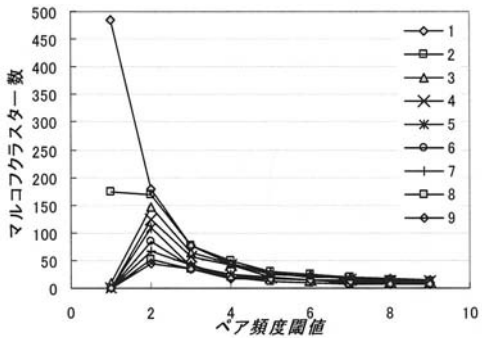


図 5. ウィンドウ幅-ペア頻度閾値によるクラスター数(1)

これらのグラフから興味深いのは以下のような点である。たとえば、ペア頻度閾値を 1 として、事実上データの制限(フィルタリング)をしないと、単語の異なり数は最大になるが、ウィンドウ幅が小さい場合は、取得できる単語ペアは種類が少なく、意味ネットワークの結線率も低くなるので、クラスター数も相当多くなる。ところが閾値 1 の場合、ウィンドウ幅を大きくしていくと、急激に結線率が上がり、ウィンドウ幅 4 でもはやグラフは密になりすぎ、これ以上、クラスターには分割できなくなる。

このようにペア頻度閾値をあげていくと、ペア頻度閾値が 1,2 の時はウィンドウ幅が 1 の場合、閾値が 3~6 の時は幅が 2 の場合、それぞれクラスター数が最大になるが、クラスター数の差はだんだん縮まっていき、閾値 7 を越えると、幅 3 の場合クラスター数が最大になるとはいえ、幅ごとの差はほとんどなく、閾値ごとの折れ線は、ほとんど特徴のないフラットなものとなる。

### 3.3. Modularity Q と F 尺度

さてそれらの結果の評価であるが、赤間ら(2007)は、MCL の結果精度にも一般に modularity Q として知られている評価値を導入している。

Modularity Q とは、同じ条件(点の総数、結線総数)のランダムグラフと比較し、結線分布が各クラスター内にどの程度偏っているかを見ることで、グラフクラスタリングの精度を与える指標である。Newman(2004)らの定義によれば

$$Q=\sum_i (e_{ii}-a_i^2)$$

であり、ここで $i$ はクラスター $c_i$ の番号、$e_{ii}$ は、グラフ全体に対するクラスター内部リンクの割合、$a_i$ は $c_i$ 内の点がもつ辺数のグラフ全体の辺数に対する割合である。Modularity Q が大きいほど、クラスタリングは精度が高いということが言われている。

先の研究で扱ったキーワード付きグラフの MCL の場合、Q の値は対角重み 1 で 0.3059、対角重み最大で 0.2878 であった。この値は、キーワード、非キーワードを全て含めたグラフ全体に対するクラスタリングの評価値になるが、非キーワード間の共起を辺として採用していないため、キーワードがハブとしてその周囲に辺を集中させているので、Modularity Q の値はおのずと高くなる構成になっている。

それに対し、IAW はあらかじめ大きな重みをハブに与える構成を取らないので、クラスタリングがいつもきれいな結果になるとはかぎらない。

また、パラメーター値によって、ウィンドウが採取できる単語の異なり数が異なり、最小 23(ウィンドウ幅 1、ペア頻度閾値 9)から最大 2262(ウィンドウ幅 4 以上、ペア頻度閾値 1)まで様々なので、結果は Modularity Q による精度 precision ばかりでなく、再現率 recall まで考慮しないといけない。再現率 R はパラメーターの各条件での単語(点)の数を最大単語取得数 2262 で割った値とする。精度 P は、Modularity Q として、両者の均衡点を探るべく、R と P の間で F 尺度を計算する。F 尺度は精度 precision と再現率 recall のトレードオフの関係の中から最適な条件を選択するのに利用される。F 尺度の公式は、一般に $0<\alpha<1$ なる重み $\alpha$ として $\frac{1}{\frac{1-\alpha}{P}+\frac{\alpha}{R}}$ であり、$\alpha=0.5$(両者の重みを均等にする)ならば、調和平均 $\frac{2PR}{P+R}$ になる。

このようにして、各パラメーター条件で Modularity Q と F 値を計算したところ、以下に示すような結果になった。

ウィンドウ幅 4 以上の場合、ペア頻度閾値が 1 では MCL はペアデータによるグラフをもはや分割できないので、modularity Q も F 尺度もほぼ 0 として扱う。むろんその場合の再現率 recall は 1 で最大値を取るがこれは考慮から外す。

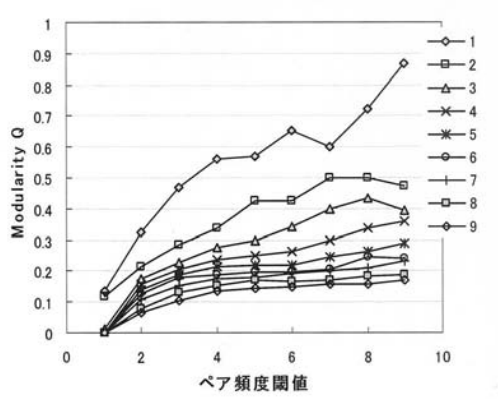


図 6. ウィンドウ幅-ペア頻度閾値による Modularity Q

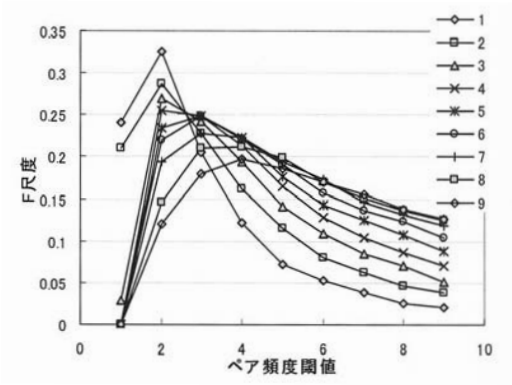


図 7. ウィンドウ幅-ペア頻度閾値による F 尺度

図 6 でわかるように、ウィンドウ幅が狭いほど、そしてペア頻度閾値が高いほど Modularity Q は高い。特に注目すべきは、ウィンドウ幅 1、ペア頻度閾値 9 の時、Modularity Q が非常に高い最大値 0.8673 を取ることである。しかし、一方でこの時の関与単語数はわずか 23 語である。確かにその分類は明解であり、CM 思想の精髄を端的に示していると言えるが、あまりに取り上げられた単語の異なり数が少なすぎる。その結果の内訳を以下に示す。各リストの先頭の数字はクラスター番号、括弧の中の数字は要素数である。

{"1(2)", "animal(動物)", "corps(身体)"}
{"2(2)", "état(状態)", "homme(人間)"}
{"3(3)", "cause(原因)", "effet(結果)", "produire(生み出す)"}
{"4(2)", "crise(発作)", "maladie(病気)"}
{"5(3)", "esprit(精神)", "humain(人間の)", "nature(自然)"}
{"6(4)", "externe(外的な)", "sens(感覚)", "interne(内的な)", "organe(臓器)"}
{"7(2)", "impression(印象)", "recevoir(受ける)"}
{"8(2)", "intelligence(知性)", "volonté(意志)"}
{"9(3)", "matière(物質)", "mouvement(運動)", "partie(部分)"}

一方、ウィンドウ幅 1、ペア頻度閾値 2 の時は、Q は 0.3258 にとどまるが、736 個の単語を拾うので、調和平均による F 値が最大値 0.3256 を記録した。図 6 でわかるように、F 尺度に関しては、狭いウィンドウ幅では、ペア頻度閾値 2 の時、ピークが大きく出るが、ウィンドウ幅が大きくなるにつれ、ペア頻度閾値 3 の方が安定して最大値を取ることがわかる。Q による精度 precision が少々落ちても、この幅ならある程度じゅうぶんな単語ペアインスタンスを拾えるからであろう。一般にウィンドウ幅が狭いほど、ペア頻度閾値を変化させたときの F 値のグラフは尖ったものになり、広いほど緩やかなものになる。

むろん以上の結果は、$\alpha$=0.5 の場合なので、精度 precision と再現率 recall のどちらかを重視するかで、$\alpha$の重みも変化させながら、バランスの取れたと思われるいくつかの結果を同時に考慮するのが良いと考えられる。

## 4. ウィンドウィングにおけるキーワードの有無

### 4.1. キーワードのカヴァチャ

ここでは、先の研究で扱ったキーワード 77 語にもとづくグラフクラスタリングと、本研究における IAW に基づくグラフクラスタリングの結果を、カヴァチャを用いて比較することにする。

まずキーワード 77 語をハブとした全部の点に関する隣接行列をもとに、重み付きカヴァチャを計算した。重み付きカヴァチャは、重みなしのもの(あまり差がつかない)にくらべ、値の差にメリハリがつき、その単語の意味のコヒーレンス(カヴァチャ値が高い場合)やポリセミー(カヴァチャ値が低い場合)を明確に反映していることがわかった。

カバニス&メスメールの例では、動物磁気などの明確な概念を構成する単語の重み付きカヴァチャがとても高くなった。一方、見方を変えれば、読み手に予備知識がない場合、普通の意味で使われる高頻度・多義語が、対象となるテキストでは、文脈依存で特殊な意

味をもつことを自動的に検出することができる。つまり頻度では決定できないキーワードの自動抽出が可能になるわけである。

それに対し、低カヴァチャの語は一般的な広い意味で使われ、これがドキュメントのハブである場合、MCL クラスターでは主にサイズの小さいクラスターのハブになる傾向がある。ただし、コアクラスターのハブについても、重み付きカヴァチャの値は小さく、そのうちのひとつ homme(人間)は重み付きカヴァチャ値が最低である。

以下、○は小 MCL クラスターのハブを意味する。◎はコアクラスター2 個のそれぞれのハブを意味する。

重み付き高カヴァチャ語トップ 10 は以下の通りである。なお、存在しない辺の潜在的な重みは 1 として計算している。このように、ほとんどの高カヴァチャキーワードは、コアクラスターのメンバーになっていることがわかる。

modification(変異) 0.4871
substance(実質) 0.4830
éther(エーテル) 0.4776
air(大気) 0.4578
feu(火) 0.4042
mécanisme(機制) 0.3927
sensibilité(感受性) 0.3893
crise(発作) 0.3776 ○
magnétisme(磁気) 0.3770
nerf(神経) 0.3677

一方、重み付き低カヴァチャ語トップ 10 は以下の通りである。やはり、存在しない辺の潜在的な重みは 1 として計算している。このようにほとんどの低カヴァチャ語は各マルコフクラスターのハブである。

homme(人間) 0.0599 ◎
cause(原因) 0.0894 ○
esprit(精神) 0.0961（重み付き MCL では ○）
nature(自然) 0.0980
idée(観念) 0.0982 ○
partie(部分) 0.1010 ○
erreur(誤謬) 0.1109 ○
effet(効果) 0.1116 ○
mouvement(運動) 0.1137
état(状態) 0.1176

その他小 MCL クラスターのハブの重み付きカヴァチャは

effort(努力) 0.2563 ○
expérience(経験) 0.1932 ○
objet(対象) 0.1580 ○
opinion(意見) 0.1221 ○
théorie(理論) 0.3509 ○
vertu(徳) 0.1224 ○
vie(生) 0.1405 ○

と、théorie(理論)を除くと概して低いほうである。なおもうひとつのコアクラスターのハブは

action(活動) 0.1234 ◎

と低いほうであった。

ただし、重み付きカヴァチャは、このケースでは、キーワードの性格付けに関して非常に有効だが、キーワードを介さない共起語間共起はデータを取らないため、キーワード以外は 1 以上の大きな値となり、あまり解釈上有効とはいえない。

キーワード中心データの場合、キーワードは最初からハブとして設定されているので、それらを中心としたクラスターを生成しやすく、安定した解が得られる。しかし、そこから漏れた単語は、ひとつもしくは複数のハブにぶら下がるだけで、それらの点のグラフ指標はあくまで、ハブのグラフ指標を形成するデータの中に吸収され、単独では積極的な意味を持たない。このことは、準キーワードの場合、問題を残す結果となる。一方、ひとつのハブに隣接する非ハブ共起語間は結線がないので、カヴァチャの値は、おおよその目安にしかならない。

### 4.2. IAW による単語のカヴァチャ

それでは、キーワードの選択から漏れた準キーワードとしてどのような単語が得られるか？たとえばIAWによる方法で、名詞キーワードをあらかじめ選択した場合とほぼ同じ数の単語が拾えるのは、ウィンドウ幅 2、ペア頻度閾値 7 のときで、そのときの単語の異なり数は 75 個になる。そのうち、名詞キーワードとしても選択されたものは 40 個、残りの 35 個は動詞が 11 個、形容詞が 11 個、非キーワード名詞が 13 個であった。つまり、もともと名詞のみをキーワードとして選んだ場合に比べ、IAW を利用すると名詞はほぼ半分になり、動詞・形容詞が残りの部分を占めるようになったわけである。

興味深い点は、そのとき取り上げられた単語ペアに基づく隣接行列から、単語のカヴァチャを計算した結果である。名詞キーワードのカヴァチャは、40 個のうちの 19 個が値 0、18 個が 0 より大きく 1 より小さい間

の実数値を取り、1は3個にとどまるということであった。一方、それ以外の単語(名詞キーワード群に含まれていないもの)のカヴァチャは、35個のうち、大多数の31個が値0、1個が0より大きく1より小さい間の実数値を取り、1は3個であった。すなわち、カヴァチャのほぼ一律0という結果から、非キーワードはほとんどが、その周囲に明確な文脈依存の意味を形成しているわけではないということになる。これは、動詞・形容詞という品詞がもともと有する機能的性格によるものであろう。キーワード中心のアプローチの有効性に変わりはないということが、ここでも言うことができる。

## 5. まとめ

本研究では、カバニスとメスメールの代表テキストを用い、単語共起に基づくテキストの特徴抽出について、キーワードをあらかじめ設定する場合としない場合とで、グラフクラスタリングの結果がどのように異なるかを示した。前者は先の研究における頻出名詞77個によるもの、後者は漸進ウィンドウ法(IAW)を利用し、共起ウィンドウが捉える単語ペアの全インスタンスを利用するものである。

後者では、ウィンドウ幅やペア頻度閾値といった様々なパラメーターの値を変えることで、それぞれの視点に応じて最適な条件を探索することができる。たとえばグラフクラスタリングの精度を表すModularity Qと、データの再現率まで加味したF尺度が最大になる条件を選択する方法がある。また、単語の文脈依存性や多義性の指標となるカヴァチャといったグラフ指標は、IAWに基づくネットワークデータにおいてのみ、全単語にわたって解釈に用いることが可能である。むろん、IAWベースのグラフクラスタリングであっても、先の研究におけるキーワード選択の妥当性は、カヴァチャの観点から示すことができる。

IAWデータにもとづくMCLの最大の問題は、出力結果に見られるクラスターサイズ間の著しい不均斉である。大きな次数のハブの周りには、極端に多くの点を含むコアクラスターが生成されやすい。このコアクラスターをどう再分割するかが次の課題となる。筆者らはその目的でBranching MCL(BMCL)というマルコフクラスターの再分割の技法を2種類(潜在隣接法、分岐分類法)提案しており、このうちのひとつがIAWデータに基づくグラフクラスタリングにも適用可能であるが、これは稿を改めて論じることにする。

## 6. 謝辞

本研究は、21世紀COEプログラム(研究拠点形成補助金)「大規模知識資源の体系化と活用基盤構築」の言語・文献、知識資源分野に関する研究の一環として行われたものである。

## 文　献

[1] 赤間啓之、ベクトル空間モデルに則った、近代ストア主義とメスメリスムの類似性に関する計量文体論的分析、情報処理学会報告, Vol.2001 No.51 1-8, 2001

[2] 赤間啓之、鄭在玲、三宅真紀、近代ストア主義とメスメール主義の思想的類似性に関するグラフ言語学的分析、情報処理学会研究報告、Vol.2007, No.49, pp.49~56

[3] 赤間啓之、三宅真紀、鄭在玲、テキスト分析における2部グラフクラスタリングの可能性、電子情報通信学会研究会、言語理解とコミュニケーション研究会、情報処理学会研究報告、2006-NL-174, pp.19~24,2006

[4] P-J-G. Cabanis, Oeuvres philosophiques de Cabanis, Edited by Claude Lehec and Jean Cazeneuve, 2 vold. Paris, 1956

[5] Dorow, B. et al. Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sence Discrimination, MEANING-2005,2nd Workshop organized by the MEANING Project, February,3rd-4th.2005

[6] Jung, J., Miyake, M., Akama, A., "Recurrent Markov Cluster (RMCL) Algorithm for the Refinement of the Semantic Network", LREC2006, pp.1428~1432,2006

[7] Jung, J., Miyake, M., Akama, A., "Markov Cluster Shortest Path Founded upon the Alibi-breaking Algorithm", CICLing-2006, LNCS 3878, Springer Verlag Berlin Heidelberg, pp55-58, (http://dx.doi.org/10.1007/11671299_6) , 2006

[8] 鄭在玲、三宅真紀、赤間啓之、再帰的なグラフクラスタリングを利用した言語連想データの処理について、人工知能学会大会、CDROM、2006

[9] F.-A.Mesmer, Le magnétisme animal, Payot, Paris, 1971

[10] 三宅真紀、グラフクラスタリングに基づく共観福音書意味ネットワークの実装、じんもんこん2006、人文科学とコンピュータシンポジウム、pp.161~165、2006

[11] 三宅真紀、鄭在玲、赤間啓之、グラフクラスタリングとパターン分類を併用したストーリー・マップ生成の試み、言語処理学会第12回年次大会(NLP2006)、pp.644-647、2006

[12] J Newman M. E. J. and Girvan M., Finding and evaluating community structure in networks, Physical Review E 69. 026113, 2004

[13] Van Dongen, S. "Graph Clustering by Flow Simulation". PhD thesis, University of Utrecht, 2000