2007－CH－73 (5)
2007／1／27

グラフクラスタリングを用いたソシュールの概念ネットワーク解析

○赤間啓之，鄭在玲，*三宅真紀
akama@dp.hum.titech.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

現代言語学の祖ソシュールの最大のキーワード、「聴覚映像(ia)」とその「同義語」である「シニフィアン(st)」は、第三回言語学講義の展開の中でいかに造られていったのか。その過程を、観念連合のネットワークの組み替えという形で、カヴァチャというグラフ理論上の指標や、MCL(マルコフクラスタリング)の様々な手法を駆使して解明する。

Conceptual network analysis of Saussure using Markov Cluster Algorithms

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

The most important keywords of Saussure, father of the modern linguistics, are "image acoustique (ia)" and "signifiant (st)." By using the Markov cluster algorithm (MCL) and the coefficient related to the graph theory, we elucidate, from the viewpoint of restructuring of the network of ideas, how these words were invented in the course of the “Third Lecture of Linguistics”.

1 はじめに

グラフ理論とそれを応用したグラフクラスタリング技法は、多変量解析をベースにした手法ではしばしば発見できない「隠れキーワード」の所在をつきとめることを可能にする。

とくに、ひとつのドキュメントを通じて、文脈に大きな変化が生じたとき、内的に行われる知の組み換えを急速に促進した触媒的な語彙が関与していることが考えられる。それらは、頻度や共起を利用した単純に統計的な手法からは、抽出することは困難である。隠れキーワードは、単語間の意味ネットワークから、同文脈語のクラスターと

しての「概念」を網羅的に列挙した後、それら「概念」の中からグラフ構成上特徴的な単語を抽出することで獲得される。本研究では、現代言語学の祖ソシュールの最大のキーワード、「聴覚映像(ia)」、「シニフィアン(st)」がどのようにして造語されていったか、その急激な文脈変化の鍵を握る意外な隠れキーワードをこの手法で発見する。

2 方法

2.1 隠れキーワード

ドキュメントの意味ネットワークは、作成の前提として、確かに単語の共起情報を利用する。しかし共起行列を多変量解析に直接かけるのではなく、それを単語が点ノードに、共起関係が辺（エッジ）に見立てられるグラフの隣接行列の形にし、まず、大きな単語間のネットワークを構築する。これがドキュメントの意味ネットワークであり、グラフ理論に関する様々な手法をそれに適用することが可能である。

結論を先取りすると、隠れキーワードは、同次数でのカヴァチャ平均からある程度外れた値を取る単語で、文脈が変化する前後でマルコフクラスターへの帰属に大きな変更が見られるもの、という形で計算することができる。ここで次数とは、グラフの点ノードと接続した辺（エッジ）の数のことである。カヴァチャ（クラスター係数）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り、カヴァチャの低い単語ノードは一般に曖昧ないし多義的であると考えられる。

2.2 MCL

マルコフクラスターは、グラフクラスタリングの手法であり、Van Dongen (2000)により提案された、マルコフクラスター・アルゴリズム(MCL)と呼ばれるものから得られる。ドキュメントの語彙データの場合、MCL は意味ネットワークをいくつかのコヒーレントなサブグラフに分割し、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る。MCL では、グラフ全体が重複のない孤立したハード・クラスターに分割されるまで、マルコフ過程に基づくクラスタリング計算を繰り返す。MCL の様々な技法を利用した分析としては以下のようなものがある。1)キーワードあり２部グラフクラスタリング(Bipartite Graph Clustering、略して BPGC)、2)キーワードなし漸次進行ウィンドウベースグラフクラスタリング(Incremental Advancing Window Based Graph Clustering、略して IAWGC)、3) BranchingMCL（分枝マルコフクラスタリング、略して BMCL）である。

2.3 キーワードあり BPGC

キーワードあり２部グラフとは、二つの点集合に分割できるグラフで、各集合内の頂点間では隣接関係がなく、結線がないものを意味する。２部グラフ(Bipartite graph)に MCL および Recurrent MCL（鄭ら,2006）を施す２部グラフクラスタリング(BpGC)は、赤間ら(2006)により提案された。文脈変化が新しいキーワードの登場によって明示される場合、キーワードまわりの細かい文脈を追うために、キーワードのインスタンスを別々に単語ノードとして扱う。さらにキーワードインスタンス群とそれらと一定条件で共起する単語群との間で、2 部グラフを形成しグラフクラスタリングにかける。これにより、キーワードの意味の範囲を複合的なものとして分節化することができる。

2.4 キーワードなし IAWGC

一方、文脈変化前は、キーワードは探索的にしか得られないので、あらかじめ設定することはない。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 IAW（Incremental Advancing Window）を利用する。IAW において、幅左右 n 語ずつに固定されたウィンドウは、文書の先頭から末尾まで、1 回だけスライドする。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、共起関係を見ることになる。IAW を利用し、共起ペア頻度データを取得したうえで、入力データとして単語グラフの隣接行列を計算の上、意味ネットワークを構築し、それを MCL にかけることで、立体的なストーリー・マップのプロトタイプを作成できることがわかっている(三宅,2006)。

2.5 BMCL

最後に、分枝マルコフクラスタリング(BMCL)について述べる。ドキュメントデータのように、単語の次数分布がおおむね Zipf の法則に従う場合、MCL によって高頻度の語彙を中心としたサイズの巨大なコアクラスターができる。よってサイズのアンバランスを調整するため、それを再分割する必要がある。だが、MCL の結果残った大きなサイズのコアクラスターは、そのままの形では再分割できない。コアクラスター「のみ」をふたたび分割する（たとえば、潜在的な隣接関係を求めて MCL にかける）手法をまとめて BMCL と呼ぶ。

3 ソシュールのグラフ理論による分析

3.1 方法論概要

本研究では、現代言語学の祖と呼ばれるソシュールの『一般言語学講義』(エングラー版、仏語）を用いた。キーワードとしては、ソシュール理論で最も有名な "signifiant"（シニフィアン）(略して st)、およびその類義語とされる "image acoustique"（聴覚映像）(略して ia)を選択し、弟子コンスタンタンによる第三回講義ノートの中の出現インスタンスのみを対象とした。中尾浩は、このノートの断片が「講義」の再編成によりばらばらになっているものを、原ノートのページ順に戻して復元している。

ia と st は、中絶した第三回講義ノートの終盤に集中的に現われる。だがこれら最重要キーワードは、すでにある既存の観念連合に新たな名前を与えた「取ってつけたようなもの」なのだろうか？それともそれらが登場することによって、初めて新しい概念が「ひらめいて」形成されたものなのか？ソシュールの造語癖を考えると、きわめて興味深い問いである。

だが、ia, st のような中心語と同一範囲中で共起する惑星語を抜き出すとき、これら諸語のカヴァチャは最重要語の近傍ではかなり高く 1 に近いので、これらの単語の振る舞いを見ても意味がない。よって、ia, st のような中心語を含まぬ箇所全体に関し、キーワード共起語のそれぞれの出現データ、特にグラフ指標から次数、カヴァチャをみるとよい。その際、カヴァチャの高い語は、すでにそのまわりに（最重要キーワード抜きでも）観念連合ができあがっていることを示す。これらは、最重要キーワードによ

って情報が遮蔽されたことを意味するだろう。逆にカヴァチャの低い語は、もともと多義的に使われた曖昧語であり、それに一意性のフォーカスをあてる為、最重要キーワードが導入されたとも考えられる。

さらに、共起語のそれぞれが属するMCL・BMCLクラスター群を比較する。もし、共起語が共通のMCL・BMCLクラスターに現れる場合、これらは、最重要キーワードの存在しない所でも、強力な観念連合を形成していることを意味する。反対に、もし、キーワード共起語が他の場所では離れたMCL・BMCLクラスターに属する場合、それは、キーワードの出現にあたってキーワード共起語のうちいくつかの語の再帰属等による、知の再編成が行われたことになるだろう。こうした発想は、文献研究に予想外のテーマ発掘をもたらし、新しい解釈学的視点を導入できると考えられる。

3.2　キーワード出現以降:BPGC

キーワード出現箇所では、まず、iaとstの用例をすべて実際の講義に現れた順に抜きだす。iaの用例は1から15、stの用例は1から12で、そのうち、ia13,st4とia15,st7はオーバーラップするので、ia13st04、ia15st07と、ひとまとめにする。そしてこれらキーワードと「異なるページで2回以上にわたって共起する」、合計142個の単語を取り出し、それらとの、有無に関する1/0の共起行列を作成する。グラフの辺として採用する隣接関係は、iaもしくはstの少なくとも一方と、それら142個の単語のどれかとの間の共起関係であり、共起語どうしの共起は辺としては採用しない。なお、キーワードと20回共起するlangue(言語)という語はノイズワードとして外しておく。

このデータをMCLにかけたとき、BMCLの適用対象となるような大きなコアクラスターは生じない。最大、2番目に大きいMCLクラスターの要素数はそれぞれ21,20であり、個人が身心の関係の中で、自らを動かし発話行為を行うというテーマを表していると解釈できる。

3.3　キーワード出現以前:IAWGC

ia、stは中断した第三回講義の終盤に固まって出現するが、それ以前の未出の箇所で、単語の共起データを取得し、MCL、BMCLによるグラフクラスタリングを用いて先行文脈を抽出する。そのため、IAWを以下のようにして使用する。

単語を原形に戻し、ノイズワード、機能語を省いて、語順と節境界を保存したデータに、十分大きな幅のウィンドウを各語停止モードで走らせる。このウィンドウは節境界に至るとデータ収集を停止し、ウィンドウ内の単語対を重複なく一意的に採取する。その際、単語対の頻度データも取得するが、これは、その頻度に対し閾値を設け、意味ネットワークの辺として使用するか否かを決めるためのものである。

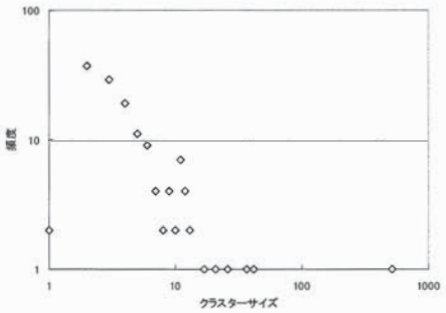

図1:IAWGCによるクラスターのサイズ分布

単語対データ(単語数 1151)から隣接行列を計算し、MCL にかけたところ、たとえば閾値 2 では、138 個の MCL クラスターが生成し、それらのサイズ分布は図 1 の通りであった。なお、コアクラスターが含む単語数は 511 で、これが後で BMCL による再分割の対象となる。

1151 語のうち、キーワードあり 2 部グラフクラスタリングの対象になった共起語は、105 個存在する。それらが、IAWGC(前)と BPGC(後)で、帰属変化があったか否かが、キーワードをめぐる文脈変化を分析する上で最も重要な視点となる。共起語のうち、多くは(73 語,全体の 70%)は、コアクラスターに帰属しており、大きく一貫した文脈を支える役割を担っている。しかも、BPGC による MCL クラスターは、ほぼ満遍なくこのコアクラスターにメンバーを送り込んでいる。ということで、これらの単語は、文脈変化の上で特徴が見つけにくい。

逆に共起語のうち、コアクラスターに属していない散在する単語のうち、後に IAWGC においても同じクラスターに属する単語はひとつもない。非コアクラスター所属語は 43 個あり、クラスター帰属がキーワード出現以前と以降では異なり、ia、st を生み出した文脈変化の大きさを物語っている。文脈変化に大きく関与しているこれらの語のうち、隠れキーワードをカヴァチャや、BMCL の使用を通じて絞ってゆく。

3.4 カヴァチャ分布による比較

共起語 142 語のうち 105 個が、st, ia の登場しない箇所（前半部）においても見られた。st、ia の登場しない箇所（前半部）において、102 個の共起語(大きいプロット)と 1151 個の非共起語(小さいプロット)が、いかなるカヴァチャの値を取るか、次数 1~26 という両者の重複範囲（低次数領域）で見てゆく。興味深いのは、各次数でカヴァチャの最大値、最小値を取る共起語の性格である。

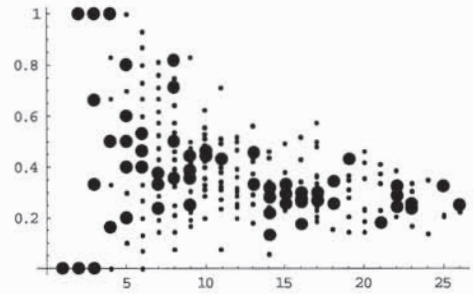

図 2：共起語と非共起語のカヴァチャ分布

次数 8 でのカヴァチャ最大値”fondamental(根本的)”(0.714)、次数 18 のカヴァチャ最大値”nature（自然）”(0.346)、次数 19 のカヴァチャ最大値”produit(産物)”(0.432)などは、一般的で漠然とした意味の単語にしては、カヴァチャの値が高く、コヒーレントな意味で使われていることがわかる。キーワードの登場しない箇所（前半部）においては、MCL,BMCL の双方にわたってコアクラスターを離れない。

次数 25 でのカヴァチャ最大値:”société(社会)”(0.33)は、次数の割には、カヴァチャの値が高く、st、ia の登場しない箇所（前半部）ではコヒーレントな意味で使われていることがわかる。しかし、コヒーレントとは言っても、コアクラスターではなく要素数 10 という比較的小さなクラスターに所属している。このクラスターでは他に"humain(人間の)"が共起語だが、次数 27 でカヴァチャ 0.282 と若干高めであ

る。この緊密さにもかかわらず、st、ia の登場する箇所では両者は別のクラスターに別れて所属する。"humain(人間の)"はまるで反対の指向性を示すように"société(社会)"を離れ、”liberté(自由)”とペアを組む。この移動は特徴的である。

次に、次数 21 でのカヴァチャ最小値:”acoustique(聴覚的)” (0.186)だが、st、ia の登場する箇所では” image acoustique(聴覚映像、すなわち ia)”というコロケーションを取らないインスタンスのみを取り上げている。この語は、st、ia の登場しない箇所（前半部）においても、次数の割にはカヴァチャが著しく低い。よってこの語は、もとから多義的に使われ、いろいろな用法を受け入れる下地を持っていたといえる。

次数 26 でのカヴァチャ最大値:”individuel(個人の)” (0.252)は、次数の割には、カヴァチャの値が高く、st、ia の登場しない箇所（前半部）ではコヒーレントな意味で使われていることがわかる。しかもこれは” société (社会)”の対義語であると解釈され、MCL、BMCL の双方にわたってコアクラスターにも属している。ところが st、ia の登場する箇所ではコアクラスターを離れ、{ “action(活動)”,”changement(変化)”,” immense(壮大な)” }と小さいクラスターを組む。ここに「個人による行動と変化」という新しい文脈が生じていることがわかる。

前半で特定の緊密な意味関係にあった単語は、問題の造語を通じた知の組み換えの中で、その緊密な意味関係を脱し、異なる文脈で（違った意味で）使用されることで、生き延びる（全体にわたり分布する）。なお、キーワード出現以前に次数 27 以上を記録した語はすべて、共起語となっており、比較的きれいなカヴァチャ分布をする。

4　ＢＭＣＬを利用したコア分析

BMCL は、キーワードなし IAWGC が生み出したコアクラスターの再分割のために用いられる。これにより、キーワード出現以前の最も安定一貫した文脈を限定的に抽出することができ、キーワードとの意味上の乖離を明確にとらえることができる。

ここで用いる BMCL の手法は以下のようなものである。まず、元の隣接行列から、コアクラスターの内部ノード cn と外部の諸代表ノード rn との間の隣接部分のみを行列 Acn*rn として抽出し、$Acn*cn=Acn*rn\times(Acn*rn)^t$ を計算する。つまり、外部の諸代表ノードを媒介にした内部ノード間の結合強度を、内部ノードどうしの生成可能な辺の潜在的な重みとして計算する。ここで結合強度とは、外部代表ノードを介した距離 2 のパスの個数であり、それが一定の閾値を越えたとき、直接に結合する辺が現に生成されると考える。これを潜在隣接と呼び、これにより再隣接化する点どうしは潜在隣接行列の値が 1 であると言う。この潜在隣接行列を MCL にかければ、コアクラスター自体のグラフクラスタリング、BMCL が可能になる。

ここでは、潜在隣接を設置する外部代表ノードが媒介する結線数の閾値が 3 の場合を例に挙げる。BMCL の後にも、さらに小規模ではあるがコアクラスターは「芯のまた芯」のような形で残る。このコアコアクラスターのメンバーを以下に列挙する。BPGC の項目では、ia、st の共起語にかぎり、所属クラスターを示す。

単語	意味	次数	カヴァチャ	BPGC
aborder	近づく	12	0.515152	-
altération	変遷	23	0.241107	6
arbitraire	恣意的	30	0.282759	8
âge	年代	9	0.5	-
écriture	記法	57	0.130952	16
époque	時代	33	0.293561	-
établir	確立する	24	0.347826	-
changer	変化する	23	0.268775	-
coexistant	共存する	12	0.5	-
concevoir	把握する	10	0.555556	-
créer	創造する	19	0.280702	-
dépendre	依存する	11	0.345455	-
diachronique	通時的	44	0.204017	-
fixer	固定する	7	0.571429	-
moyen	方法	31	0.195699	-
offrir	呈する	17	0.5	-
phonétique	音素的	42	0.239257	6
produire	産出する	23	0.3083	-
synchronique	共時的	32	0.245968	-
venir	来る	30	0.181609	-

表 1：BMCL 後のコアクラスター

一言で言って、これらの単語は、時間的変化と空間的均衡、あるいはマクロレベルのダイナミクスとスタティクスを表しており、こうした概念系統が ia、st というキーワード出現以前ではコアをなしていたことが BMCL の結果からわかる。

5　隠れキーワード"masse"について

カヴァチャ分析や BMCL によるコア分析からわかるように、ia,、st というキーワード出現以前においてきわめてコヒーレントな文脈を形成していた語は、人間個人に対してマクロレベルで拘束力をもつ実体の動静にかかわる、たとえば、社会や集団に関するものである。

さて、ここで特に注目する" masse"という語は、たしかにそうしたコア部分に直接属するものではないが、キーワード以前と以降で属するクラスターが趣をまったく異にする。キーワード以前は、{aboutir(達する), colonie(コロニー), conceptual(概念的), discontinuité(不連続性), informe(不定形), masse, nébuleux(不明確な), partiel(部分的), regard(視線)}というクラスターであり、後で見るように、カヴァチャ特徴語と関係する。しかしキーワード出現以降に属するのは、{complexe(複合的な),ia12(ia の 12 番目のインスタンス),intermédiaire(媒介), masse, ordre(秩序), phénomène(現象), positif(実定的な)}というクラスターである。

このように、"masse"という語は、ia,、st 以前は、集団的様態のなかで、「未分化」という観点で用いられていることがわかる。次数がある程度大きい(34)わりには、カヴァチャも大きく(0.203)、この語の属している文脈のコヒーレンスの高さをうかがわせる。ところが以後の BPGC では、キーワード出現以前の文脈を離れ、一般的に「関係性」を示す語と MCL クラスターを形成していることが見て取れる。

実は"masse"こそ、隠れキーワードのひとつであると考えられるのである。この語は、まず人間が寄り集まった「社会」(カヴァチャ特徴語)、「集団」を指すとともに、弁別可能なものがそこから出来するような、不定形で連続的な、抽象的「塊」を意味する。この「集団」と「塊」の両義性が、心身性と対立性を、そしてけっきょくは「聴覚映像(ia)」と「シニフィアン(st)」を媒介する

ことになるのである。ただし、ここでは、具体的な引用に基づく解釈は本研究の範囲を超えているので深入りしない。

ただ、少なくとも言えることは、この語は抽象的・操作的な意味も含め無定義に近い豊かな多義性をもち、キーワードの展開において、社会集団における言語というものの統一されたイメージを、個人（これもカヴァチャ特徴語）が担うさまざまな（聴覚）映像の同一性のイメージへと、マクロレベルからミクロレベルに対し投影するために利用されているということである。この語は、分節言語が差異性の原理に従って分節されるにあたって、そのもともとの背景となるもの--集団と個人双方の場--の包括的な表象として使われていると言える。

6　まとめ

“masse”のように、ドキュメントの展開の中で、連続性と不連続性を、コヒーレンスとフラクチャーの矛盾を媒介する単語は、それ自身の自己主張的な意味を前面に出すことはなく、むしろ黒子的な機能に徹するので、しばしば読者の目にとまらない隠れキーワードとしてその存在が忘れられてしまう。こうした「頻度の小さい機能語」は、概念マップの産出にあたってまず消去されてしまうのが常であろう。

それに対し、グラフ理論の指標とグラフクラスタリングは、ドキュメントの概念ネットワーク中の目に見えない重要なポイントを自動的に抽出することで、鋭い読み手や解釈の達人のみが可能な深い「読み」をシミュレートすることができると言える。

そもそもグラフ理論の利用に関しては、1)マップ上を移動できるルート（ネットワークフロー）の総検索が可能、2)グラフという形での、アナロジー的直観に合った形の 2 次元マップ、3)広い意味での主成分としてのクラスターの、分割統合の自在性が、強調できる材料である。さらに、大規模知識資源の処理に向いている点、MCL 系アルゴリズムは、大きく有利であることも特筆されよう。こうした方法を人文系ドキュメント解析に大々的に適用すれば、主観性と客観性の対立を超えた「計算解釈学」という新しい学問分野の創出も、夢ではないと考えられる。

【参考文献】

[1] 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

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

[3] 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

[4] Saussure, F. de., « Cours de linguistique generale, tome 1 », Edition par Engler, R., Otto Harrassowtz, Wiesbaden, 1989

[5] Van Dongen, S. “Graph Clustering by Flow Simulation”. PhD thesis, University of Utrecht, 2000

[6] 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

[7] 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.

[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、(2006)