社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－ＮＬ－174（４）
2006／7／27

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

赤間 啓之† 三宅 真紀† 鄭 在玲†

† 東京工業大学大学院社会理工学研究科 〒152-8552 東京都目黒区大岡山 2-12-1

E-mail: † {akama, mmiyake, catherina}@dp.hum.titech.ac.jp

あらまし テキスト解析にマルコフクラスタリング(MCL)、およびそれを独自に改良したリカレント・マルコフクラスタリング(RMCL)を利用する場合、有効なデータ取得法として、キーワードと共起語のペアに基づく、２部グラフ化の方法を提案し、MCL-RMCL による２部グラフクラスタリングの計算結果を、従来のベクトル空間モデルに基づく多変量解析と比較、有効性を検証する.

キーワード マルコフクラスタリング，２部グラフ，多変量解析

Possibilities of the Bipartite Graph Clustering in Text Analysis

Hiroyuki AKAMA† Maki MIYAKE† and Jaeyoung JUNG†

† Graduate School of Decision Science and Technology, Tokyo Institute of Technology 2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8552 Japan

E-mail: † {akama, mmiyake, catherina}@dp.hum.titech.ac.jp

Abstract In the fields of text analysis, results of Markov Clustering and Recurrent Markov Clustering applied to a bi-partite graph made of pairs between key words and co-occurring words would be useful to be compared with those of multivariate analysis.

Keyword MCL, RMCL, Bipartite Graph, Multivariate Analysis

1. 研究の背景：MCL,RMCL

単語と単語の関係を表す意味ネットワークをひとつのグラフとしていくつかのクラスタ（サブグラフ）に自動分割するには、Van Dongen(2000)によるマルコフ・クラスタリング(MCL)が有効である。MCLは、きわめてシンプルなアルゴリズムであり、大規模なグラフデータであっても確実に収束し、グラフを重複ノードのないクラスタ（サブグラフ）に分割することが知られている。すなわち、グラフ上でマルコフ過程に従うランダムウォークを設定する時、Expansion と Inflation という二つの手続きを反復し、遷移行列自体を漸次修正することで、次第にランダムウォーカーがグラフの密なエリアに捉えられ、抜け出せなくなるよう仕向け、結果としてグラフ自体を非連結のクラスタに分割させるという手法である。

MCL の応用分野では、Enright et al. (2002)によるプロテイン分類のための Tribe-MCL、ノイズデータを加えクラスタリングをスムーズにした Gfeller の Synonymy Network(2005) 、Dorow et al. (2005) らによる単語の非曖昧化のための MCL のクラスターマージなどが知られている。また鄭 (2006)および三宅(2006)は MCL をカスタマイズして言語データに適用し、そこから様々な連想情報を抽出、連想作文支援システムなど教育工学等の分野に応用したり、新約聖書の意味ネットワークを構築して統計的語彙分析に適用したりした。その際、鄭らは MCL の過分類を修正しつつ、意味ネットワークの世界を自動的に要約するため、リカレント・マルコフクラスタリング(RMCL)という、独自の MCL カスタマイズ法を提案している。RMCL とは、MCL の収束状態におけるハードクラスタ（ノード間にオーバーラップがないクラスタ）間に、それ以前のクラスタリングステップ（クラスタステージ）におけるノード間のオーバーラップデータをもとにして、自動的に潜在的な隣接関係を復元し、さらにそれを再度 MCL に投入するというものである。

RMCL の飛び石(Stepping-stone)アルゴリズムを用いると、収束クラスタステージ ClusterStage*k* とそれ以前の各クラスタステージ ClusterStage*i* = {C*i*(1), C*i*(2), ..., C*i*(r)}(1,2,..,r はクラスタ番号)との間で、クラスタステージ間行列 ClusterStage*k*-ClusterStage*i* Matrix = Cluster-Word Matrix*k* × Tr (Cluster-Word Matrix*i*)を計算、さらに ClusterStage*i* のオーバーラップ情報を使って、ClusterStage*k* のハードクラスタを Cluster Matrix*i* = ClusterStage*k*-ClusterStage*i* Matrix × Tr (ClusterStage*k*-ClusterStage*i* Matrix*i*)により再連結した。なお Tr は行列の転置を表すとする。この Cluster

Matrix$i$の対角成分を0、非対角成分のうち0でないものを1に置換することで、隣接行列 Adjacency Matrix$i$を生成し、そこから不要な過剰接続を排除した(鄭、2006)。さらにRMCLの計算結果をMCLに再入力し、MCLクラスタ自体のMCLクラスタリング(第2回MCL)も行っている。本研究でも鄭らのこの手法に従って計算するものとする。

2. 2部グラフクラスタリングとは

このように。グラフクラスタリングの言語データへの適用は、一定の制約のもとできわめて有効に機能する。ただし、全域的、あるいは局所的に密な結線率は、サブグラフへの分割という意味でクラスタ化を阻害するし、極端に次数の大きい点の影響によって、クラスタサイズに大きな偏りが出る場合がある。しかし、そうした制約を差し引いても、連想概念辞書に対するMCL,RMCLの適用例が示すとおり、かなりの精度で大規模な言語データを処理できる方法であると言える。

本稿では新たに、言語データ解析における多変量解析の利点を導入するため、グラフクラスタリングの一種である2部グラフクラスタリングを取り上げる。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,RMCLを施す2部グラフクラスタリング(BpGC)は、ベクトル空間モデルで使用される多変量解析の場合のように、変数(キーワード)、オブザベーション(共起語)を別々の集合として扱うので、その出力結果をベースラインとしての多変量解析と比較するのが容易である。

むろん、外的基準のない多変量解析は、他のデータ解析にとって精度や再現率などの点ではベースラインになりにくい。しかし、本研究では、言語で書かれた作品に対するコンピュータ解析支援を、適用される主なドメインとして想定している。そもそも、特に人文科学の研究には、検索の精度やヒット率よりも、マクロ(グローバル)な意味情報—抽象的全体—とミクロな(ローカルな)構成情報—具体的細部—の双方を保存しつつ両者の間で往復的な参照ができる作業方法論が期待されている。後で見るように、2部グラフクラスタリングは、計算結果から元情報を直接参照できない階層的クラスタ分析や因子分析に代わり、それらが明らかにできない「情報の源脈」や「情報の質差」を提示するという、feasibility(実施可能性)の観点から独自の有効性を持つものと考えられる。それは後で見るように、データの内部ネットワーク化によって、本質的な「概念」である因子の間に、解釈を押し拡げる源脈と質差を導入することになる。それを示すため、本研究では人工的なランダムデータと実際のテキストデータの双方を用い、グラフクラスタリングと多変量解析を行い、両者の結果を比較する。

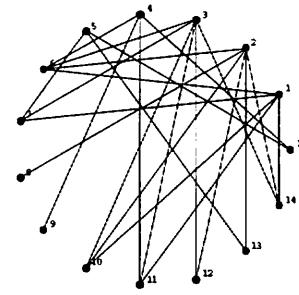

図1 典型的な2部グラフの図

3. BpGCの計算(1): ランダム人工データ

まずランダムな人工的2部グラフであるが、その2つの点集合をS、Lとし、集合の濃度(要素数)を#()であらわす。そして、完全2部グラフの隣接行列CompleteGraph(#(S),#(L))と、2項分布を用いたランダム確率pで結線する同サイズのランダムグラフの隣接行列 RandomGraph(#(S)+#(L),p)との間で、アダマール積(対応する要素間の積)を取ることによって生成させる。ただし、孤立点を避けるため、0のみからなる行、列は計算後すべて除去する。

BipartiteGraph(#(S),#(L))==Hadamard Power(CompleteGraph(#(S),#(L)),RandomGraph(#(S)+#(L),p))

その際、人工的2部グラフの隣接行列を対角上に並ぶ2つの零行列と残りの2つの転置関係にある2値の部分行列に4分割し、後者を変数-オブザベーションの関係にある多変量解析用データ行列とする。

ここで人工的な2部グラフの2つの点集合S,Lの濃度を、「多変量解析における変数の数」<<「オブザベーションの数」のバランスにならった形で、#(S)<<#(L)とする。Sは変数の集合、Lはオブザベーションの集合となる。<<は多変量解析に主に必要となる十分に大きな個数差のことを表すとする。そのような人工的な2部グラフにMCLを適用した場合、偏った次数分布から自明なこととして、S,Lのうち十分小さい方Sの要素がすべて、MCLクラスタの各々に1個ずつ分散して代表的なノードとなり、ほぼSの濃度と同じ個数のクラスタが生成する。後で見るように実データでは、互いに隣接関係を持たない点集合内でMCLクラスタが生成する場合もある。しかし、後出の実データの場合でもわかるとおり、変数1クラスタ1の一意対応関

係は原則として維持される。

#(MarcovClusters(BipartiteGraph(#(S),#(L))))==#(S)

for i=1,2,..,#(S),
Intersection(MarkovCluster(i), S)==S(i)

すなわち以上のような条件では、2部グラフのMCLは、対応するクラスタ分析の場合、1変数1クラスタの状態とパラレルになる。第1回MCL収束時に、変数ノードは、すべて1個ずつ別々にMCLクラスタに分配されるので、変数のノード番号と同時に第1回MCLクラスタ（クラスタノード）番号として読めるように設定できる。言い換えると、すべてのオブザベーションは一番結びつきの強い変数と一意的に結びつく。1変数1クラスタでオブザベーションのクリスプな分類ができるのはBpGCの大きな利点である。言い換えると、その条件下で複数の変数が属するMCLのクラスタが出現することには、大きな意味があることになる(L,Sのサイズ差のない人工ランダムデータや実データでは、ありうるケースであり、http://dl.dp.hum.titech.ac.jp/wiki/?MCL#content_1_3 のMCL-based connection inside a vertices set of bipartite graph にデータあり)。さらにMCL結果に飛び石タイプのRMCLを施す、すなわちMCLのクラスタを今度は点として扱い、そこに隣接関係を復元して再度MCL（第2回MCL）を施すと、今度は新たに変数間にグループができる。ちょうどたとえばバイナリーなWARD法などのクラスタ分析の出力とパラレルになるが、結線率が0.05未満程度のごく小さい値を取るときに限り、2回目のMCLのループ回数が最大で収束したクラスタステージ(だいたい8,9クラスタステージ目)においてのみ、ある程度類似点のある計算結果が得られた。双方の計算結果について、詳細はhttp://dl.dp.hum.titech.ac.jp/wiki/?MCL#content_1_3 のBpGC Compared with Cluster Analysis (Ward Method)を参照されたい。

ここでは、結線率0.05の条件で、人工ランダム2部グラフの2つの点集合S,Lのサイズをだんだん近づけていった場合（すなわちオブザベーションに比べ、変数の個数を増やして行った場合)、1変数1クラスタの条件がどこまで維持できるかを示す。1変数1MCLクラスタの原則とその例外の生ずる条件について、この表はBpGCと多変量解析の比較が可能な基本的背景を示している。図2は、変数によって代表されるMCLクラスタ内にオブザベーションが含まれる(オブザベーション1個だけのクラスタにはならない)割合である。thetaはMCLの計算でノードを拾うことが十分可能な閾値である。このように、変数の設定可能な個数内では変数1MCLクラスタ1の原則が維持できると見なしうる。

http://dl.dp.hum.titech.ac.jp/wiki/?MCL#content_1_3 のLimits of the principle: "1 variable in 1 cluster"を同時に参照されたい。

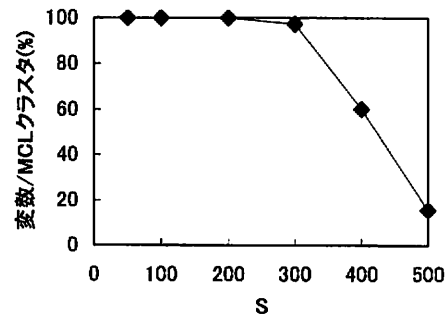

図2 変数1MCLクラスタ1の原則

4. BpGCの計算(2): 実テキストデータ

4.1. MCL,RMCL

一方BpGCと多変量解析を同時に適用する実際のテキストデータであるが、例として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）を用いて生成した。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のを原ノートのページ順に戻して復元している。

ia, stの用例を、原ノートのページ単位で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まで、stの用例は1から12まであり、出現順位をタグとして添えるが(そのうち、ia13, st4とia15, st7は同一ページでオーバーラップするので、ia13st04、ia15st07のように、ひとまとめにする)、これらのキーワードと、「異なるページで2回以上にわたって共起する」、合計142個の有意味な単語（前もって設定したストップリストの中にあるノイズワードは除く）との間で、有無に関する1/0の共起行列を作成する。この共起行列をそのまま隣接行列に置換するが、その際、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。

GridMathematicaによる独自開発のMCLプログラム

にこの2部グラフをかけたところ、ほぼ13回で収束し、23個のクラスタが生成した。また、このように変数間、サンプル間でショートカットを一切設けない隣接関係の設定では、しばしば報告される大きなコアクラスタは生じなかった（最大クラスタサイズ 22 が、次点が 20 であった）。その結果をここでは共起語非表示で、｛"MCL クラスタ番号(クラスタ内のメンバー数、ただしキーワード、共起語すべて含む)"，" キーワードのインスタンス"｝のように簡略してあらわすと、{1(10), ia01}, {2(22), ia02}, {3(2), ia03}, {4(2), ia04}, {5(20), ia05}, {6(3), ia06}, {7(11), ia07}, {8(1), ia08}, {9(9), ia09}, {10(6), ia10, st09}, {11(11), ia11}, {12(6), ia12}, {13(7), ia13st04, st05}, {14(4), ia14}, {15(1), ia15st07}, {16(1), st01}, {17(7), st02}, {18(7), st03}, {19(1), st06}, {20(1), st08}, {21(17), st10, st12}, {22(4), st11}, {23(14), NUL(変数なし)}になる。ほぼ1変数1クラスタであるが、ローデータにはない、変数間に結合が新たに生じたもの3個（クラスタ番号 10, 13, 21）、オブザベーション間（だけ）に新たに生じたもの1個(クラスタ番号 23)が認められた。さらにこの結果を飛び石型 RMCL にかけたところ、ランダム2部グラフの場合と同じように MCL の第 8, 9 ステージ付近で興味深い再クラスタリング化が見られた。これを RMCL8, RMCL9 と名づけ、｛"RMCL クラスタ番号[第1回 MCL クラスタ番号]"｝のように表すと、RMCL8 は、{1[1, 18]}, {2[2, 5, 14]}, {3[3, 23]}, {4[4, 7, 11, 21]}, {5[6, 12, 17]}, {6[8]}, {7[9]}, {8[10]}, {9[13]}, {10[15]}, {11[16]}, {12[19]}, {13[20]}, {14[22]}、RMCL9 は、{1[1, 3, 23]}, {2[2, 5]}, {3[4]}, {5[6, 17]}, {6[7, 11]}, {7[8]}, {8[9]}, {9[10]}, {10[12]}, {11[13]}, {12[14]}, {13[15]}, {14[16]}, {15[18]}, {16[19]}, {17[20]}, {18[21]}, {19[22]}になる。それ以外のステージとの間では、さほどきれいなクラスタに分かれなかった。むろん過結合を除去するアルゴリズム次第で、飛び石型 RMCL の結果は微妙に変化することに注意する必要がある。

なお計算結果について、詳細は http://dl.dp.hum.titech.ac.jp/wiki/?MCL#content_1_3 の BpGC compared with Factor Analysis (from Saussure's Document)を参照されたい。

4.2. 因子分析

一方、それと比較される多変量解析には、単語の出現有無に関するバイナリーな共起行列に対するプロマックス回転付の因子分析を用いた。因子分析は変数となるキーワードの特徴が明瞭に抽出されることから、すでに聖書の計量文体論など計算人文科学の分野ではよく用いられる方法である(三宅, 2005, 2006)。データは隣接行列のもとになった共起行列であり、データセットに大きな違いはない。相関行列をベースに、SPSS(v. 13)で計算した所、固有値1以上で因子9個が抽出され、累積寄与率は 59.6%であった（表1)。これらの因子をひとつずつ、BpGC における MCL クラスタ、RMCL8 クラスタのそれぞれと比較してゆく。

比較の結果、9つ中5つの因子が BpGC のクラスタと類似していることが判明した。すなわち、因子1は MCL の {13(7),ia13st04,st05| と、因子2は MCL の {10(6),ia10,st09}と、因子4は RMCL-9 の {5[6, 17]}(すなわち{ia6, st02})と明確にマッチし、因子5は RMCL8 の {1[1, 18]}(すなわち{ia1, st3})および RMCL9 の {1[1, 3, 23]} (すなわち{ia1, ia3})と、因子8は MCL の {21(17), st10, st12}と弱くマッチする。他の4個の因子は照応関係をつけることはできなかった(むろん、変数間の相関ベースによる類似度とパス長・パス分布をベースとした類似度では、変数ノード間で食い違う局所の扱いが異なることに注意する必要がある)。ただし、BpGC によって抽出された変数間関係のうち因子にも明確に現れたものは、BpGC、因子分析双方において元のテキストに照合して分析可能であるばかりか、意味抽出上、重要性を帯びていた。詳しくは本論の目的ではなく、実際の「解釈学」の領域に踏み込むので割愛し稿を改めるが、この5個の照応可能な因子のうち、因子2と因子5は、キーワードインスタンス間のページ数が極端に大きく、「同一思考の間を置いた変奏」「同一テーマのフラッシュバック的回帰」を明らかにする重要なパターンであり、他の照応関係の付けられなかった因子よりも、潜在的意味の抽出という点で有益であった。

さらに、抽出された因子をグラフ情報の観点から分類した(表II)。ここで、matchQ は MCL・RMCL の結果とのマッチングが可能か否か、matchcond はマッチングの仕方、overlaprate はそれぞれの因子において最大因子負荷量をもつ2個のキーワードと距離1で隣接する共起語のうちの重複率(すなわち両キーワード間の最短パス上にある中継する共起語の割合)、netflow は両キーワード間の最大フロー数、directflow はそのうちの最短パス（直接経路)、indirectflow はそのうちの間接経路であり netflow-directflow の値になる。

表が示す通り、原理的に言って MCL クラスタは、netflow 数がある程度限られた上で、直接結合の確率が高く、間接結合（迂回パスを通る）の確率が低い場合生じやすい（因子8の箇所も MCL クラスタに従い、st11, st12 の組ではなく、st10, st12 の組で選択するとこの傾向が現われる。Cf. overlaprate=0.2381, netflow=15, directflow=10, indirectflow=5)。その観点(低い netflow 値、directflow>indirectflow)は上の表でも明確に現われており、本来回転によって寄与

率などの差を問えなくなった因子間で、グラフクラスタリングを介し性質の再分類、情報の質差の提示が可能であることを示している。

以上、多変量解析の場合のように、変数（キーワード）、サンプル（共起語）を別々に扱い、変数、サンプル間で2部グラフを作り MCL->RMCL を施すと、部分的に回転付き因子分析から得られるのとほぼ同様な重要な情報が、グラフクラスタリングによっても得られることがわかった。ただし、BpGC が検知できた類似性を因子分析が検出できるとはかぎらないことに注意する必要がある。

たとえば、MCL の第 8, 9 ステージとの間で計算される RMCL8、RMC9 からは、安定して (ia02, ia05) というパターンが現れるが、"ia02"、"ia05" がともに因子負荷の大きい因子パターンはいかにしても生成することができない。だが、MCL クラスタ 2 すなわち "ia02" とその共起語、MCL クラスタ 5 すなわち "ia05" とその共起語の文脈上の関連性は、クラスタ内の共起単語の同系列要素性を見ると解釈学上十分に認められる。しかしベクトル空間モデルでこの意味関係を捕捉できないのは、overlaprate が 0.13 と小さい値で、なおかつ netflow 数が 29 と大きい局所的な結線状況が影響していると考えられる。このような情報の源脈を因子分析では追ってゆくことが困難である。なお、MCL クラスタ 2 と MCL クラスタ 5 の 2 つはそれぞれ最大、2 番目に大きい MCL クラスタであり、ドキュメントの内容に触れると、ともに「個人が身心の関係を通じて発話行為を行う」という大きなテーマを表しているが、このテーマは st の共起関係（解釈学の内容上、記号論的な背景を持つ）には一切見られない（本稿ではこのテーマには深入りしない）。

ちなみに同一のデータに対して WARD 法による階層的クラスタリング分析を行ったところ、因子分析と同じような傾向が現われた。すなわち MCL, RMCL の双方によって、近接する単語ペアのいくつかがカバーされている。しかし、RMCL では 2 度現われる MCL クラスタ 2 と MCL クラスタ 5 が、デンドログラム上で遠い位置（最長距離）に配されており、このことからも多変量解析データが Network flow に関する細部の特徴をとらえていないことがわかる。なお以上ソシュールデータの詳細は、http://dl.dp.hum.titech.ac.jp/wiki/?MCL#content_1_3 の BpGC compared with Factor Analysis (from Saussure's Document) を参照されたい。

5. まとめ BpGC の利点と欠点

BpGC は、変数とサンプルの関係を二分グラフに表現した上で MCL,RMCL を施すクラスタリング手法である。BpGC の利点をまとめると以下のようになるだろう。もしグラフクラスタリングによっても、多変量解析が抽出できる情報のうち重要な価値のあるものを抽出できるとすれば、もともとグラフクラスタリングは、多変量解析から落ちる重要情報、すなわち、対象ノード間の経路情報、パラダイムシークエンス情報など、ベクトル空間モデルではローカルでエピソード的 (Burgess,1999) と軽視される直接的・間接的な連関情報を保持しているのが有利な条件として働く。

もともと、R&B MCL、BpGC 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、大規模なデータに向いている点、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。

ただし、BpGC にはいくつか克服すべき問題点が残されている。たとえば、実データの場合、クラスタリングがデータセット中の過結合や最高次数語に鋭敏に反応し、それを除去しないと全体で 1 個のクラスタしか生成しないことがある。つまり、全テーマが最も一般的な汎用語の存在によって強固に一体化してしまうわけである。また BpGC は、パス長がすべて偶数になり、最短パス以外は必要以上にパス長がもつ重みが大きくなるという欠点もある。さらに飛び石型の RMCL は MCL のすべてのクラスタステージに対応して計算結果を出力するため、最適なデータをユーザー・読者が自ら判断する必要に迫られるということも挙げられる。しかし、まだこの方面の研究は端緒についたばかりであり、グラフクラスタリングが今後、ドキュメント解析でますます有効性を発揮するであろうことは確かなように思われる。

文　献

[1] Burgess, C., Lund, K., "The Dynamics of Meaning in Memory, Cognitive Dynamics: Conceptual Change in Humans and Machines", Dietrich & Markman (Eds.), 1999

[2] 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

[3] Enright, J., Van Dongen, S. and Ouzounis, C. A., "An efficient algorithm for large-scale detection of protein families", Nucleic Acids Res. Apr 1;30(7):1575-84, 2002

[4] Gfeller, D., Chappelier, J.-C. , De Los Rios P. "Synonym Dictionary Improvement through Markov Clustering and Clustering Stability", International Symposium on Applied Stochastic Models and Data Analysis, 106-113, 2005

[5] Jung, J., Miyake, M., Akama, A., "Recurrent Markov

Cluster (RMCL) Algorithm for the Refinement of the Semantic Network", LREC2006, pp.1428~1432,2006

[6] 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

[7] Miyake, M., Akama, H., Sato, M., Nakagawa, M., "Approaching to the Synoptic Problem by Factor Analysis", Proceedings of the Institute of Statistical Mathematics 48(2), pp.327-337, 2002

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

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

[10] 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.

[11] 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).

表 I：因子パターン行列

パターン行列a

	因子								
	1	2	3	4	5	6	7	8	9
ia01	.094	.101	.065	.088	.371	-.120	-.184	.159	-.065
ia02	-.065	-.095	.073	.044	.354	-.238	.113	.051	.030
ia03	-.055	-.018	-.165	.025	.728	.171	.007	-.130	.041
ia04	-.115	.022	.161	.022	.094	.117	-.209	-.001	.217
ia05	.032	-.107	-.068	.178	-.005	-.052	-.041	-.014	.645
ia06	-.006	.099	.021	.417	.110	-.075	-.025	.176	.123
ia07	.059	-.010	.642	-.019	-.129	-.119	-.061	.001	-.049
ia08	.000	-.192	.493	-.028	.097	.032	-.088	.153	.167
ia09	-.072	.111	.518	.030	.005	.074	.187	-.128	-.138
ia10	.015	.510	.128	.033	.123	.090	-.084	.032	-.020
ia11	.135	.115	.005	-.155	.030	-.093	.056	.083	.322
ia12	-.017	.120	.184	.173	-.227	.089	-.115	-.003	-.009
ia13st04	.810	.142	-.021	-.103	-.024	.004	-.001	-.071	.177
ia14	.305	-.043	.054	-.138	.146	-.013	.256	.062	-.089
ia15st07	.324	-.079	.076	.152	.006	.431	-.124	-.066	-.086
st01	-.055	.298	.129	-.094	-.004	.100	.256	-.078	.284
st02	-.070	.040	-.014	.627	-.021	.072	.177	.039	.035
st03	.031	.366	-.083	.241	.167	-.062	-.069	-.008	-.053
st05	.640	-.114	.009	.052	-.074	.088	.081	.034	-.036
st06	.116	.048	.038	.322	.091	-.137	.495	-.084	-.093
st08	.015	-.040	-.043	.128	-.103	.517	.134	.001	.154
st09	-.022	.824	-.114	.075	-.161	-.011	-.086	.084	-.026
st10	-.011	.123	-.056	-.172	.030	.596	.046	.187	-.182
st11	.045	-.164	-.070	.056	-.034	.116	.669	.204	.005
st12	-.019	.096	.013	.121	-.019	.106	.180	.796	.026

因子抽出法: 主因子法
回転法: Kaiser の正規化を伴うﾌﾟﾛﾏｯｸｽ法
a. 13 回の反復で回転が収束しました。

表 II：因子と MCL、RMCL クラスタの対応表

Factor	matchQ	matchcond	Overlaprate	netflow	directflow	Indirectflow
1	yes	MCL(strong)	0.346	13	9	4
2	yes	MCL(strong)	0.306	19	11	8
3	no	×	0.25	26	11	15
4	yes	RMCL(strong)	0.25	17	8	9
5	yes	RMCL(weak)	0.233	23	10	13
6	no	×	0.227	19	11	9
7	no	×	0.185	13	5	8
8	yes	MCL(weak)	0.225	14	9	5
9	no	×	0.13	21	6	15