潜在隣接を用いた意味ネットワークの適正なクラスター化について

鄭　在玲‡　三宅真紀*　赤間啓之‡

†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

E-mail: ‡{catherina, akama}@dp.hum.titech.ac.jp, *mmiyake@lang.osaka-u.ac.jp

概要　グラフクラスタリング手法 MCL (Markov Clustering) を言語データに適用する際、単語頻度分布がクラスター・サイズの不均斉をもたらすことが知られている。高頻度の語彙をハブとした巨大なコアクラスターの内部分割のため、新しい Branching MCL(BMCL)の方法を提案する。すなわち、MCL が計算した「他の」代表ハブノードを媒介として、コアクラスター内の任意の点間の最短パス数を計算する。それが一定の閾値を越えた時、内部で辺が復元されるという、「潜在隣接」を導入し、潜在隣接行列を再度 MCL にかける。混沌とした意味ブロックをサブ系統概念に分岐させる手法として有効かどうか、石崎概念連想辞書に適用し検証する。

キーワード　グラフクラスタリング, MCL, BMCL, 潜在隣接

1. はじめに

Markov Clustering Algorithm(MCL)とは、グラフ上でマルコフ過程に従う random walk を反復させる時、遷移行列自体を漸次修正することで、次第に random walk するエージェント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ーに分割させるという手法である[1]。鄭らは、この MCL を、単語の連想関係を収集したコーパスである日本語連想概念辞書(慶應義塾大学、石崎俊研究室刊、以下「石崎連想辞書」と略)に施し、希少語を除く全 9373 語を 1408 個の類似・同系列の概念クラスターに分類した。さらに、鄭らは、これら MCL クラスターの生成過程である各ループでのクラスター間重複情報へ遡行することで、概念クラスターをさらに上位の概念クラスターへ統合する Recurrent MCL(RMCL)を施した[2][3]。

ただし、ここにひとつ問題が残されている。石崎連想辞書のように、単語が点ノードとして持つ次数がおおむね Zipf の法則に従う場合、MCL の結果として、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出力されるということである。そのようなコアクラスターは、意味のまとまりがつかず、しばしば異なった意味系列の単なる寄せ集めに見えてしまう。そこで三宅らは、次のような形でコアクラスターを分割する方法を考案した。すなわち、単語対データを取得するため、パラメーター閾値を次々に変えても、MCL の結果には共通するロバストな単語パターンが出力される。そこで分岐分類の検索法を用い、それらの中で元になる先祖パターンを自動抽出する、という方法である[4]。ただし、この方法は、Windowing を利用した探索的な方法ゆえ、反復的な計算量が大きく、またコアクラスターは必ずしもハードには分割されずいくつかの点ノードがオーバーラップするという欠点がある。

本研究では BMCL を根本的に改良し、MCL が計算した「他の」代表ノードとコアクラスターのメンバーとの隣接関係を利用し、コアクラスター「のみ」をふたたび MCL にかけるという方法を考案している。この方法をふたたび石崎連想辞書に基づくクラスター化された意味ネットワークに適用すると、最大次数語をハブとするコアクラスターはいくつかの意味系統に適正に分割されることが確認された。

2. コアクラスターと潜在隣接

一般に辞書の意味ネットワークにおける単語の次数の分布と、その意味ネットワークを MCL にかけて生成させたクラスターのサイズ(すなわち各クラスターに含まれる単語数)の分布の間には、密接な関係がある。グラフは石崎連想辞書のデータから両者を比較したものであり、双方の最底部において、右側に大きくずれた地点で、最大次数の単語とそれらを各ハブとするコアクラスターが見出せる結果となっている。(図 1 と図 2)

http://dl.dp.hum.titech.ac.jp/wiki/?miscellaneous#content_1_1 にコアクラスターに関する主なデータを収録しているので参照されたい。最大コアクラスターは、ハブ（最大次数ノード）が「家」でサイズ(ノード数)が 170 である。

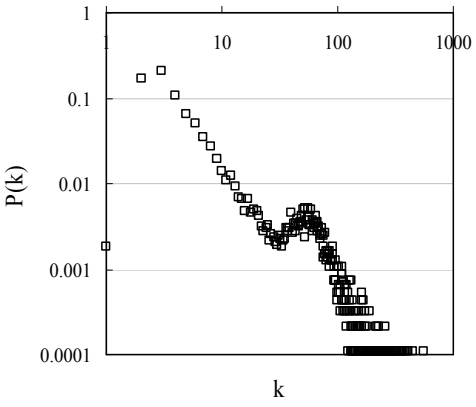

図 1 石崎連想辞書の次数分布

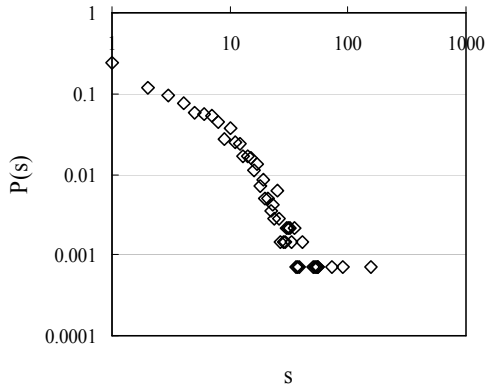

図 2 石崎連想辞書のクラスター・サイズ分布

コアクラスターに BMCL を適用するにあたり、まずコアクラスターの内部ノードと外部の代表ノードとの間の元データの結線状況から、内部ノードの間の帰属上の指向性差異を見てゆくことにする。

以下に示すリストであるが、内部リストの第 1 要素、鞄(以下、高級、室内…)は、「家クラスター」の要素をあらわす。内部リストの第 2 要素はリストになっているが、それぞれの要素は第 1 要素と隣接する MCL 代表ノードである。

{{鞄, {箱, 麻, 封筒, 漫画, 亭主, 教科書, 鈴, 手帳, 容器, 会社員,カメラ, ノート, ふくろ, 傘, ファッション, 弁当, シャープペンシル, 学生, 支度, ペン, かご, 荷物}}

以下は上記 URL の 4)に全データを収録している。

これでわかるのは、第 1 要素は、力関係で「家クラスター」に入れられたが、リファレンス・グループは他の複数の MCL クラスターだということである。これら外部代表ノードからなるクラスター(第 2 要素)は、上位概念がコアクラスターによって与えられれば、即座にそれらをつなげている「心(こころ)」が理解可能となるものである。

3. BMCL のアルゴリズム

新しい BMCL では、元の隣接行列から、コアクラ

スターの内部ノードと外部の諸代表ノードとの間の隣接部分のみを抽出し、その部分行列とその転置行列の積により、外部ハブを介したコアクラスター内の(元データにおける)結線数を計算する。さらにそれをマルコフクラスター内に潜在隣接を復活させる指標として利用し、潜在隣接行列をふたたび MCL にかけると、コアクラスターがサブ・マルコフクラスターとして再分割されることになる。

まず、元の隣接行列から、コアクラスターの内部ノードと外部の諸代表ノードとの間の隣接部分のみを行列 Acn*rn として抽出し、

NSPcn*cn=Acn*rn×Transpose(Acn*rn)-(1)

を計算する。ここで A は隣接行列(の部分行列)、NSP は Number of shortest paths(最短パス数)の略、cn,rn という添え字はそれぞれ corecluster nodes(コアクラスターノード), representative nodes(代表ノード)を意味する。NSPcn*cn は、外部の諸代表ノードを媒介としたコアクラスター内部ノード間の長さ 2 の最短パス数を表している。そこに適宜、閾値を設定し、

Acn*cn=MakeAdjacency(Threshold(NSPcn*cn))

により潜在隣接行列 Acn*cn を生成、それをもとに MCL を使いコアクラスター内部の再分割を行う。ここで隣接を意味する 1 の値を設定する閾値条件を、NSPcn*cn(i,j)>=p && MAX(NSPcn*cn(i,_),q) という形で表現する。上式の p は、結線行列の要素の最小値、q は各行から残す最大値の順位に対応する。

このように、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。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、直接に結合する辺として現実に生成されると考える。これを潜在隣接と呼び、潜在隣接行列の値が 1 であると表したりする。

ここで注意すべき点は、コアクラスターの代表ノードのデータは外しておくということである。コアクラスターの代表ノードは、コアクラスターのすべてのメンバーと、他の代表ノードを介しては距離 2 で結合しており、この計算では、コアクラスターのメンバーはことごとく家ノードとの間で潜在隣接する。よってその場合は MCL の計算はサブグラフを生まない。さらに対角成分はすべて 0 と置き、コアクラスター内の潜在的隣接行列として MCL にかけるわけである。

4. 最大コアクラスターに対する BMCL

このように、最大コアクラスターからハブノード（家）は省いたが、参照する代表ノード群からは外していない。かつ代表ハブノードを介した最短パス数の最大値を 3 個取って潜在的に隣接するとした。なおここでは、NSPcn*cn(i,j)>=p 条件は、特に設けなかった。結果は以下の通りである。

{{鞄，袋，軽い}，{高級，低い，高い，立つ}，{持つ，なくす，振る}，{室内，扉，金具，家具屋，四角い，引き出し，外す，物置，拭く，磨く，しまう，整理する，勉強部屋，こたつ，蝶番}，{実家，ゆったり，大切，気持ちいい，両親，母親，爺さん，のんびり，核家族，散歩する，飲む，あたたかい，落ち着く，温める}，{建てる，コンクリート，鉄筋コンクリート，煉瓦，セメント}，{帰る，座る，静か，暖かい，広い，畳，椅子，心地いい，入る，街，郊外，住まい，寛ぐ，狭い，町中}，{やわらかい，ツルツル，フカフカ，フワフワ，丸める，ザラザラ，スベスベ}，{閉める，閉まる，閉じる，あける，開く}，{丈夫，なくなる，使う，便利，引く}，{あつい，消す，分厚い}，{安い，渋い，貰う，落とす，並べる，売る，破れる，洗う，子供部屋，紐，ビニール，盗む，飾る，ベランダ，かっこいい，かける，買う，高価，おしゃれ，屋敷，結婚式，干す}，{暗い，寂しい，深い}，{古い，新しい}，{釘，変える，壊す，造る，直す}，{隠す，出す}，{壁，城，一軒家，ドア，ハウス，平屋}，{団地，暮らす，住む}，{片づける，汚れる，野外，臭い，運ぶ，汚い，汚す，

掃除する}，{滑る，喜ぶ，歌う，転ぶ，遊ぶ，休む，いる}，{割れやすい，壊れやすい，壊れる}，{拾う，捨てる，抱く}，{ねじ，塗料}，{素敵，起きる，怖い，うるさい，殴る，倒す}，{茶の間，詰める，入れる，蓋}，{出来る，探す，押入れ，見つける，投げる，しっかり，小さい}，{借りる，貸す}，{玩具，玩具屋}，{置く}，{屋内，屋外}}

このように新しい BMCL の適用によって、第 1 コアクラスターは、意味のある 30 個のサブクラスターに分割されたことが明確に見て取れる。

一方、(1)式で右辺の 2 項を交換し、

NSPrn*rn=Transpose(Acn*rn)×Acn*rn---(2)

を計算すると、逆にコアクラスターの内部ノードを媒介とした、外部の諸代表ノード間の長さ 2 の最短パス数が計算可能になる。「家」クラスターで NSPrn*rn を計算すると、1419 個ある代表ノードのうち、全体の 93%にあたる 1316 個が(「家」ノードは外す)、170 個の「家」クラスター内部ノードと隣接関係を持っていた。さらに、BMCL と同様に、

Arn*rn=MakeAdjacency(Threshold(NSPrn*rn))

を計算し、これを MCL にかけると、「家」と関連する概念クラスターが、コアである「家」クラスターの外部に生成した。これらは、「家」という語のもつコノテーションとして、道具、建物、収納・仕切、中身、狭さ、衣料、安置、家族、身体、光、火というテーマ系を成す。これらテーマ系のサイズを MAX(NSPrn*rn(i,_),q)の q という閾値ごとに表したものが図 3 になる。

たとえば、閾値 10 の場合を例に挙げると、道具:{鏡，封筒，眼鏡，電気製品，容器，器，ノート，宝庫，ふくろ，入れ物，ペン立て，虫かご，かご，かばん，鍵，勉強机};建物:{巣，教室，建物，アパート，トイレ，工場，邸宅，えんとつ，玄関，全身，飲み屋，マンション，建築物，住居，住宅，自宅，床，部屋};収蔵仕切:{箱，家具，檻，靴箱，戸，戸だな，食卓，食器棚，本棚，棚，カーテン，たんす，収納家具，小屋};中身の詰まったもの:{おもちゃ，機械，ぬいぐるみ，人形};狭空間:{穴，隙間};安置する場:{ソファベッド，ラブソファ}となる。なお、他は 1 クラスター1 ノードである。

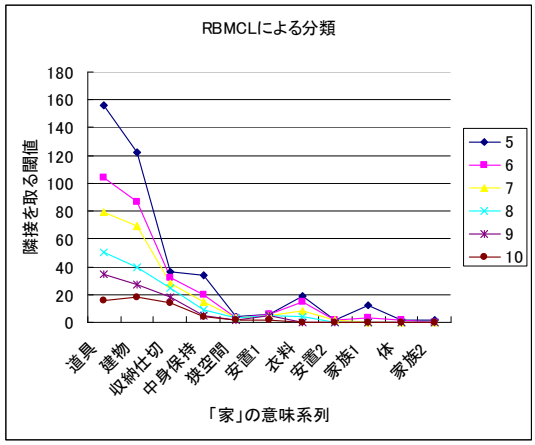

図 3 第 1 コアクラスターの RBMCL による意味抽出

5. まとめ

本研究では、単語の次数分布の偏りによる MCL クラスターのサイズ不均斉の問題を解消し、類義語の集合としての概念が理解可能な規模に落ち着くための方法として、BMCL というアルゴリズムを提案、石崎連想辞書に適用してその有効性を検証した。この方法は辞書資源から自動的にシソーラスを生成する上で利用可能性が期待できる。

今後は連想概念辞書にとどまらず、一般の国語辞典や WordNet、通常のドキュメントの語彙データに関してもこの方法を適用し、分類精度がいかに向上するか検証したいと考えている。

文　献

[1] Van Dongen, S. (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht,http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm

[2] Jung, J., Miyake, M., Akama, H. (2006) Recurrent Markov Cluster (RMCL) Algorithm for the Refinement of the Semantic Network, LREC2006, 1428-1432.

[3] Jung, J., Miyake, M., Akama, H. (2006) Markov Cluster Shortest Path Founded upon the Alibi-breaking Algorithm, CICLing-2006, LNCS 3878, Springer Verlag Berlin Heidelberg, 55-58

[4] 三宅、Jung、赤間(2006) 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、NLP2006、644-647.