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

三宅 真紀，Jaeyoung Jung，赤間 啓之
東京工業大学社会理工学研究科
{mmiyake, catherina, akama}@dp.hum.titech.ac.jp

本研究では、Van Dongen (2000)が提唱したグラフクラスタリング手法 MCL（Markov Clustering）に着目し、MCL が言語データに適用される際に生じる､クラスターサイズの不均斉の問題を解消するため､新たなアルゴリズム、BMCL（Branching Markov Clustering）を提案する。また、独自のウィンドウィング法により共起ペア頻度データを取得し、入力データとして単語グラフの隣接行列を計算の上、BMCL を適用する。具体的な言語資料としては、サン＝テグジュペリの「星の王子さま」を用い、立体的なストーリー・マップを表す意味ネットワークの自動構築を行い、手法の有効性を示す。

1　はじめに

言語資料から単語と単語の関係を表す意味ネットワークを形成する際には、隣接、共起、連関等の関係に基づく「単語のペア・インスタンス」の選定が重要な問題となる。ドキュメントベースの場合は、係り受けや同格といった統辞的特徴からペア・インスタンスを抽出することが一般的だが、これらの情報は十分に精度の高い形態素解析、統語解析の技術が求められる。このシンタクティカルな手法に代わるものとして、ウィンドウィング法の利用が知られており[1] [2]、意味ネットワークを形成するために、必要かつ十分な数の単語ペアを抽出することができる。しかし、出現する単語ノード数とペア結線数が大きくなると、グラフは複雑な模様を描き、意味関係を一目で把握することが困難になる。

そこで、類似する単語ノードをグラフ上でクラスタリングして概念エリアを形成するような技術が必要となるが、Van Dongen が提唱したグラフクラスタリング MCL（Markov Clustering）の適用は有効な手法であるといえる。MCL は、ランダムウォークに基づいたシンプルなアルゴリズムであり[3]、これによって、単語間ばかりでなく、ダウンサイジングした概念間での意味ネットワークの生成が可能になる。

しかし一方で、ドキュメント内の単語の頻度分布は、Zipf の法則に従い、Scale free-Small world なネットワークを形成するため [4]、MCL によるクラスタリングは、単語の頻度分布の偏りが原因となって、概念エリア=グラフクラスターのサイズに著しい不均斉といった問題が生じる。

以上のような先行研究を背景として、我々は、人間の直感的感性に合致した意味集合とそのネットワークを提示する統一的な方法論の研究を課題としてきた。まず、マクロ的な解決策として、MCL を発展させたアルゴリズム Recurrent Markov Clustering を考案し、MCL のクラスター外部間の再隣接化することに成功した [5]。本研究では、ミクロ方向のアルゴリズム、BMCL（Branching Markov Clustering）を提案する。そして、独自のウィンドウィング法によって BMCL を実装し、共起ペア頻度データを用いて小説「星の王子さま」のテキストに適用を試みる。

2　漸進ウィンドウ Incremental Advancing Window

ウィンドウィング法とは、ドキュメント上に一定幅のウィンドウ（フレーム）を走らせ、その内部に出現した共起単語をカウントする方法である[6]。この方法は、別の基準で選定された重要な意味形成語リストがあらかじめ用意されており、これらの語がウィンドウの中心に来た時、それと共起する単語インスタンスを抽出する場合が多い [7]。一方、意味ネットワークは、ネットワークが生成した後に、そ

の結果として重要語を決定するために用いられるものであり、従来の方法では不十分である。

そこで本研究では、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漸進ウィンドウ（Incremental Advancing Window）を新たに提案する。

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。先行するウィンドウ停止状態においてすでにカウントされた共起ペアは重複してカウントしないために、次の要領でカウントする。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カウントする。ウィンドウ内の単語をそれぞれノードとする「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。

つまり、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尾に達し、そこでカウントをすませたら終了となる。このようにして、ウィンドウサイズ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ことが可能になる。

3 Branching Markov Clustering

MCL を言語データに適用する際、クラスターサイズの不均斉が問題となる。特に、MCL が作り出す、飛び離れてサイズの大きいクラスターは、一般的すぎて命名が困難なものもが多い。その一方で、この大クラスターの特徴は、言語データにおける抽象的、包括的な意味の集結であり、メタデータの生成という可能性を持っている。

そのような MCL の大クラスターの存在意義を考慮しつつ、クラスターサイズの偏りを解消するために、我々は、MCL に分岐分類（branching method）の方法を導入し、「クラスタリング」と「分類」の組み合わせたアルゴリズム BMCL(Branching Markov Clustering)を提案する。この定義は、検索分類、分岐的方法を用いて、元になる祖先パターンの探索的な自動抽出を行い、その結果を MCL の大きなハードクラスターの内部に適用して、その再分割を行うものである。

BMCL の実装にあたっては、幾つかの方法が考えられるが、本研究では漸進ウィンドウによる共起ペア頻度データを用いて、アルゴリズムを実現する。この手法は、ストーリー・マップを形成する上で有効なものである。

具体的には、ウィンドウ幅を変化させながら、その中で共通して出現する単語パターンを分岐分類の方法を用いて系統化する。パターンの分類は、ペア頻度の閾値を固定して、ウィンドウ幅を変えた結果を総当り的に比較する。このとき、重複して出現する単語パターンは、その単語列の長さによって分類が可能である。後に見るとおり、ペア頻度閾値は、経験的に 4,5 が適当である。

ここで、単語列の長さが異なるものの中で、内包や分岐のような関連性を想定でき、それらの類似性から系統分類の適用を許すものが多数見られる。そして、それらを系統樹かベン図で表すことができ（図１）、root として最短のもの（ベン図では、他のパターンを集合として含まない集合）を祖先パターンと定める。

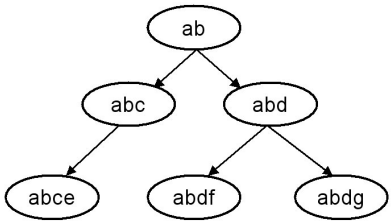

図 1： 系統樹

BMCL の場合、祖先パターンはその最短長を前もって決めて計算することが可能である。この場合、

ウィンドウ幅の拡張の間接的な結果として 1 単語ずつ増えるパターンの長さによって、系図的なつながり（分類学における世代）を明らかにする。そして、この長さのインクリメントによって世代進行を表す。この祖先パターンは、数代にわたって相同な形質を維持するので、祖先パターンに帰属する単語の連関は安定して強いことがいえる。

各 MCL クラスターの内部に祖先パターンを可能な限り発見することで、各 MCL クラスター内部の単語ノード間で連関の強さに差をつけることが可能になる。すなわち、祖先パターン中で共起するかどうかに従い、それらの単語を再隣接化、再グラフ化することができる。これは、ほぼ星グラフ化した MCL クラスター内部の再隣接化によって、そこに成分(コンポーネント)を再導入することを可能にする。

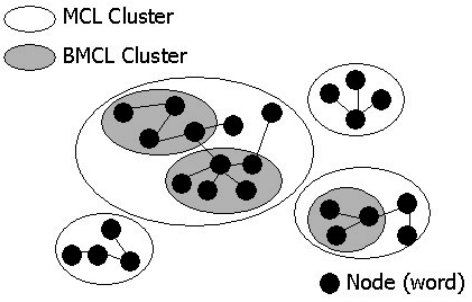

図 2 ：MCL&BMCL クラスタリング

4 BMCL の「星の王子さま」への適用

この節では、言語資料としてサン＝テグジュペリの「星の王子さま（Le Petit Prince）」を使用し、共起ペア頻度データから MCL にかけ、ペア頻度別の結果について考察する。そして、その結果を入力データとして単語グラフの隣接行列を計算し、BMCL を適用した結果を示す。さらに、BMCL の結果から、立体的なストーリー・マップを表す意味ネットワークの自動構築を試みる。

4.1 共起ペア頻度データ

まず、テキスト中の出現単語に対して、形態素解析をし、全単語を原形に戻した。そして、２で記述した漸進ウィンドウ法を用いて、共起ペア頻度データを取得した。また、ノイズワード・機能語を除いた 1312 語を使用単語として用いた。

ここで、ペア頻度閾値別に、ウィンドウ幅を拡張させたときの MCL クラスター数の推移を図３に示す。さらに、各ペア頻度閾値において、MCL クラスターの 1 番目と 2 番目にサイズの大きいクラスターを比較し、ウィンドウ幅を拡張させたときのその差を図４に表す。

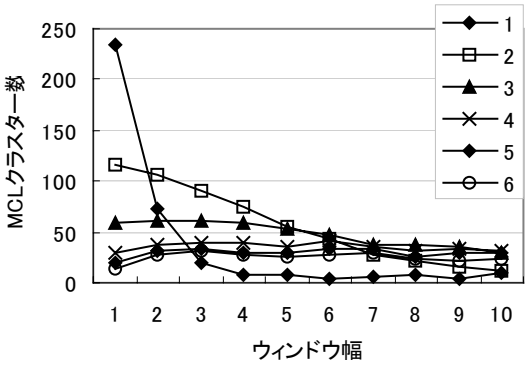

図 3：ウィンドウ幅とペア数閾値

図３から、単語ペア頻度閾値が 1,2,3 においては、ほぼ単調に減少し、閾値が小さいほど、クラスター数の落ち込みが激しいことがわかる。また、単語ペア頻度閾値が 4,5 を越えると、ウィンドウ幅ごとの MCL クラスター数が(幅=1 を例外として)、安定する傾向がみられた。

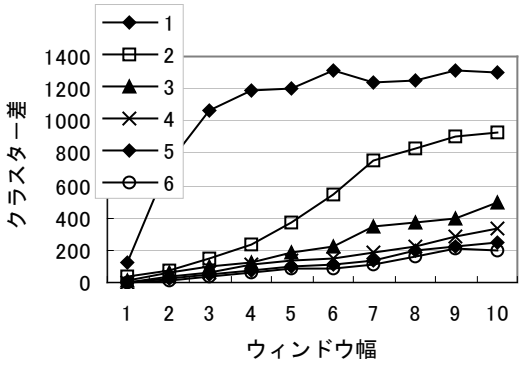

図 4：クラスターサイズ差の推移

図４から、単語ペア頻度閾値が小さい場合は、ウィンドウ幅を拡張させていくと、ひとつのクラスターが異様に大きい数の単語を吸収し、クラスター間で極端な偏りが確認された。また、ウィンドウ幅が大きさに比例して、出現する単語の種類が増えるこ

とは明解であるが、その増分は、一番大きなサイズのクラスターが取っている。

4.2 BMCL

次に、4.1 で取得した共起ペア頻度データをそれぞれ MCL にかける。今回は、ペア頻度閾値を 5 に設定し、ウィンドウ幅が 1 から 10 までの、共起ペア頻度データを使用する。そして、MCL の各ハードクラスターからこの先祖パターンを抽出し、MCL クラスターを BMCL によって内部分割する。ウィンドウ幅10におけるMCLクラスターの一つを例にとり、BMCL によって内部分割した結果を表す。

{ {« addition », « **air** », « champignon », « **sérieux** »,}
{« **air** », « monsieur »,}
{« compte », « manquer »,}
{« croire », « **sérieux** »,}
{« fleur », « rencontrer »,}
{« abri », « anéantir », « arroser », « béte », « boulon », « caravane », « déranger », « deviner », « fabriquer », « globe », « griffe », « irriter », « marteau », « naïf », « pâle », « pétale », « prince », « rendre », « reprendre », « semblant », « sot », « tâche », « tousser », « traverser »}
}

ここで、クラスターを {} で囲み、各単語ノードを « node »で表す。また、重複するノードに関しては、太字で表した。BMCL で内部分割されたクラスターをみると、{« addition », « air », « champignon », « sérieux »}, {« air », « monsieur »}, {« croire », « sérieux »}の３つのクラスターは、連結成分を作り、MCL クラスター内に４つの成分が生成されたことがわかる。

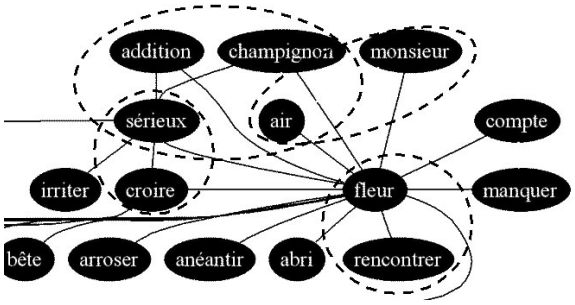

図 5 : BMCL クラスタリング例

5 おわりに

本研究では、グラフクラスタリングとパターン分類法を組み合わせたアルゴリズムを提案し、独自のウィンドウィング法を用いて実装した。そして、サン＝テグジュペリの「星の王子さま」を用いて、立体的なストーリー・マップを表すセマンティックネットワークの自動構築を行い、提案した手法の有効性を示した。今後は、他の文学テキストや概念辞書などへの適用を試みる予定である。

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。

【参考文献】

[1] Burgess et al., Explorations in context space: words sentences and discourse, Discourse Process. 25, pp.211-257, 1998.

[2] Lemaire B. and Denhière G., Incremental Construction of an associated Network from a Corpus, Proceedings 26th Annual Meeting of the Cognitive Science Society, pp. 825-830, 2004.

[3] Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, 2000.

[4] Steyvers, M., Tenenbaum, J., The Large Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, Cognitive Science, 29 (1) pp.41-78, 2005.

[5] Jung J., Miyake M., and Akama H. Markov Cluster Shortest Path Founded upon the Alibi-breaking Algorithm, CICLing-2006, LNCS 3878, Springer Verlag Berlin Heidelberg,pp55-58. 2006

[6] Schutze H. and Pederson, J.O., A cooccurrence-based thesaurus and two applications to information retrieval, Information Processing & management, vol.33, No.3, pp.307-318, 1997.

[7] Takayama Y. et al., Information Retrieval Based on Domain-Specific Word Associations. In Proceedings of PACLING '99, Waterloo, Ontario, Canada, 1999.