共起単語の選定を目的とするグラフクラスタリング評価の考察

三宅 真紀
mmiyake@lang.osaka-u.ac.jp
大阪大学 言語文化研究科

本研究では、階層化グラフクラスタリング RMCL （Recurrent Markov Clustering）の最適な意味ネットワークの作成を目的として、グラフクラスタリングの最適化とデータサイズを考慮した指標に基づき、クラスタリング結果について考察する。コーパスには新約聖書の福音書を使用し、ネットワーク指標に基づいて複数の意味ネットワークを作成する。また、データの特徴量から構造を観察し、単語・概念間における適切な意味ネットワークの構築が可能となるような共起単語ペアの選定を行う。

Thoughts on Evaluations of Graph Clustering for the Selection of Word Co-occurrence Data

Maki Miyake
mmiyake@lang.osaka-u.ac.jp
Graduate School of Language and Culture, Osaka University

In order to construct an optimal semantic network by employing the hierarchical graph clustering algorithm of Recurrent Markov Clustering (RMCL), this study discusses some graph clustering results in terms of optimal clustering and data sizes. The corpus used in this study is the Gospels of the New Testament for which two semantic networks are created based on network features. The network structures are investigated from the perspective of constructing appropriate semantic networks that capture the relationships between words and concepts.

1. まえがき

World Wide Web (WWW) や Social Network 等の実世界ネットワークのデータ解析に対しては、グラフ理論に基づいたネットワーク分析が直感的に把握しやすい方法論として一般的に用いられている。コーパス分析に関しても、意味ネットワークが複雑系ネットワークと共通した性質を持っていることが示され、スケールフリー・スモールワールドの構造が、自然言語の世界においても成り立っていることを、Roget のシソーラス[1]や WordNet[2]から作成された意味ネットワークを対象にして明らかにされた[3]。このように、単語や単語グループの関係性を理解するために、ネットワークを適用してデータの構造を体系的に把握することは有用である。

本研究では、Van Dongen が提唱した MCL（Markov Clustering）[4]から発展した階層化グラフクラスタリング RMCL（Recurrent Markov Clustering）[5]の最適化を目的として、クラスタリング精度、ならびに、ネットワークのデータサイズを考慮した２つの指標に基づき、クラスタリング結果について考察する。

コーパスには新約聖書の福音書を使用し、ネットワーク指標に基づいて複数の意味ネットワークを作成する。また、マクロ的な視点で、データの基本統計量からネットワーク特性について調べ、ネットワーク構造について観察する。さらに、単語・概念間における適切な意味ネットワークの構築が可能となるような共起単語ペアの選定を行う。その結果、RMCL を新約聖書の福音書に適用して、最適な意味ネットワークの構築を目標としている。

2. 意味ネットワークの作成

テキストは、古典ギリシャ語 Nestle-Aland26 版[6]を使用する。福音書の全出現単語（8361 単語）から、ウィンドウィング法によって共起単語ペアを取得し、隣接行列データを作成する。単語と単語の関係を表すような意味ネットワークを形成する言語データとして、隣接、共起、連関等の関係に基づいた「単語のペア・インスタンス」の選定が重要である。ペアの抽出には、係り受けや同格といった統辞的特徴を使用することが一般的だが、ウィンドウィング法[7]を利用した共起単語ペアデータによるグラフ化も効果的な方法である[8]。

ウィンドウ幅に関しては、単語の前後関係を考えたウィンドウ幅 1 における共起情報、また、節の前後情報を捉えた少し大きめの幅として、ウィンドウ幅 10 の 2 種類の幅を設定して、それぞれの共起単語ペアの頻度数をカウントする。

ここで、全ての共起データをグラフクラスタリングに適用した場合、機能語や高頻度出現単語の影響により、単一のクラスターに纏まる危険がある。実際に、クラスタリング係数 0.1 以下の単語をみると、και (and), δε (but), ο (the), εν (in)をはじめとした、接続詞、定冠詞といった単語が多く含まれている。幅 1 の MCL 結果を例に挙げると、2 クラスターに収束され、単語の意味関係を考察するには適切なクラスターとは言いがたい。そこで、意味生成にさほど関与しないノイズワードが持つ特徴に注目し、クラスタリング係数を閾値としてノイズワードを除去することが可能であると考えた[9]。ノード間の繋がりの度合いを表す指標であるクラスタリング係数[10]を閾値として、サブネットワークを作成する。各幅のデータから、係数を 0.1 ステップで 0.9 まで区切り、各係数値以上の単語抽出した結果、合計 18 個の隣接行列で表される、頻度数を重み付けした無向グラフを意味ネットワークとして作成する。

3. ネットワーク特性

ネットワークの特性を表す基本統計量として、次数分布とクラスタリング係数は、実世界の複雑系ネットワークが持つスケールフリー、スモールワールド性を明らかにするために欠かせない指標である。福音書ネットワークの構造を明らかにするために、2 種類のウィンドウ幅から作成した全ての共起単語ペアのグラフに対してネットワーク特性量を調べる。

3.1 次数分布

まず初めに、次数分布からスケールフリーネットワークの構造を調べる。Balabasi と Albert[11]によって導入した、次のような次数分布とべき乗則の関係を使用する。

$$P(k) \approx k^{-r}$$

図 1 に幅 1 （WS1）における全単語データ、の次数分布を示す。幅 1 （WS1）は、分布がべき乗側 $P(k) \approx k^{-r}$ （r=1.7）に従ってスケールフリー性を示し、また平均次数は 15.5（全ノードの 2%）と小さく、スパースな構造であることが分かる。また、エッジの結線率とノード数が WS1 ネットワークと等しいランダムグラフ作成し、その次数分布を同図にプロットしている。ランダムグラフの場合は、釣鐘型に分布をしており、スケールフリー性が保たれていないことが分かる。

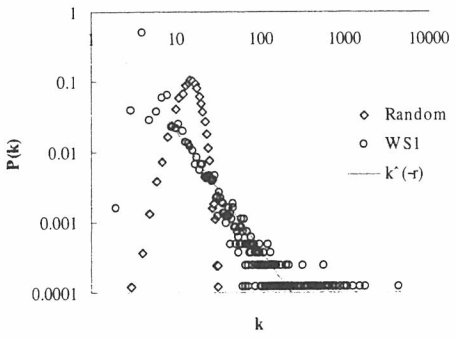

図 1. 次数分布（ウィンドウ幅 1）

一方で、幅 1 0 （WS10）の分布は図 2 に表されるような形状となった。分布の左側は、ランダムグラフにみられるような釣鐘分布をしており、右側はべき乗則に従う分布となっていることが分かる。ここで、右側の分布に合わせてべき乗のデータ（指数係数 r=1.3）をプロットした。以上から、釣鐘型とべき乗に従う 2 種類の合成分布から形成されているように見られる。べき乗則とは異なる分布をしている。平均次数は 106.4（全ノードの 13%）で、幅 1 と比較するとスパース性が損なわれている。

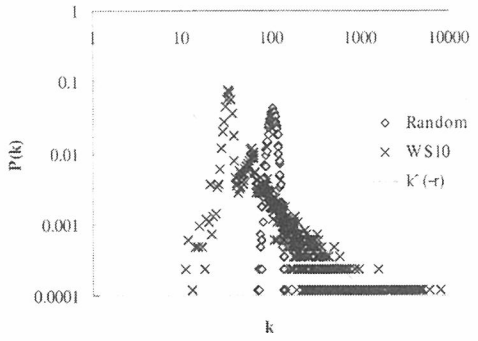

図 2. 次数分布（ウィンドウ幅 10）

3.2 クラスタリング係数

次に、クラスタリング係数の特徴量から、ネットワークのノード間の繋がりの度合いの性質を調べる。クラスタリング係数は、Watts と Strogatz [10]が「知り合いの知り合いが知り合いである確率」を定量的に表す指標として導入したものであり、ノード n に関して、隣接するノードを N(n) と表すとき、n のクラスタリング係数 C(n) は次のように定義される。

$$C(n)=\frac{\text{隣接ノード間のエッジ数}}{N(n)\times(N(n)-1)/2}$$

ここで、クラスタリング係数の値は、0～1 の範囲で表され、係数 0 のときは、図 3 のようなスターグラフを形成し、また係数 1 の場合は、図 4 のような完全グラフを形成することから、クラスター係数によってネットワークのクラスター性を知ることができる。スケールフリーな構造を持った、疎なネットワークにおいては、同レベルのランダムグラフと比較してクラスタリング係数が十分大きな値を持つとき、スモールワールド性が示される[3]。

図 3：C=0　　図 4: C=1

この指標は、言語データの単語間の類似度を表すのに関しても重要な指標として用いられている。その一例として、Dorow ら[12]は、クラスタリング係数をグラフの上の Curvature（クラスタリング係数と同意語）とみなし、単語をノードとした意味ネットワークの Curvature を基準にして、隣接する単語の意味的関連性を測定することを提案している。また、クラスタリング係数を閾値として計算する Curvature クラスタリングを British National Corpus(BNC) に適用して、多義語などの曖昧性の強い単語の除去を行い、クラスタリングとしての適応性を示している。

さらに、Ravasz と Barabasi は、クラスタリング係数の次数ノードへの依存関係を示すことで、複雑系ネットワークのスケールフリー性と階層構造を明らかにした[13]。これは近年、Dorogovtsev ら[14]が発見した、スケールフリーネットワークを決定付ける手段として、次数 k を持つノードのクラスタリング係数の平均値は、次数 k の-1 乗に従う法則に基づいている。Ravasz と Barabasi は、指数定数の－1 から発展しての階層指数係数 β を導入し、は、$C(k)\approx k^{-\beta}$ で表されるように、クラスタリング係数分布も次数分布と同様にして、次数に従うことを示した。

図 5 は、次数ごとのクラスタリング係数の平均値をプロットした分布である。WS1,WS10 の両分布とも同様にべき乗則に従っていることから、階層構造を持ったネットワークであることが分かる。従って、本研究で適用する階層的グラフクラスタリングである Recurrent MCL を適用する妥当性が確認できた。また、このとき、２つのネットワークの平均クラスタリング係数は、WS1 の場合は 0.62、WS10 は 0.75 であり、非常にクラスター性の高い性質を持っていることが分かる。

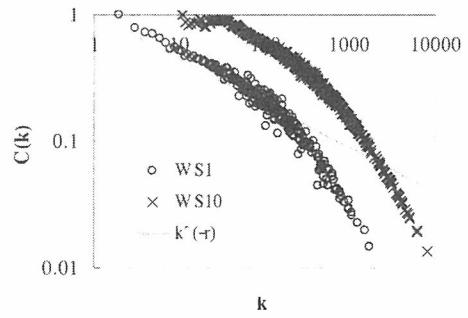

図 5. クラスタリング係数分布

4. グラフクラスタリング

4.1 RMCLの適用

Jung らによる再帰的アルゴリズム RMCL は、Van Dongen による MCL（Markov Clustering）の結果の単語の頻度分布の偏りが原因だと考えられるクラスターサイズの著しい不均斉を解消することを目的として考案された手法である。

MCL グラフクラスタリングは、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、グラフの隣接行列から得られた遷移行列に対し、マルコフ過程の操作を反復して行うことで、全体をクリスプな部分グラフに分割する。MCL は、パラメータ操作の容易さと収束の速さから、大規模データからのパターン抽出に適しており、類義語辞書等の言語データへの応用研究も行われている[15]。

RMCL は、MCL のクラスタリング過程を利用して、収束ハードクラスター間を再隣接化し、再度 MCL を計算することで、ネットワークのダウンサイジングを行いながら、単語・概念間における適正なネットワークの階層化を実現する。

4.2 MCL 結果

図 6 に幅 1 ,10 におけるクラスタリング閾値ごとの単語数と MCL 計算結果のクラスター数を示す。閾値 0.1 を除く、幅 10(WS10)の傾向に注目すると、データの数に関わらず一定した MCL クラスター数に収束していることから、幅をある程度広げていくと、クラスタリング係数に依存されないクラスターが抽出されることが分かる。

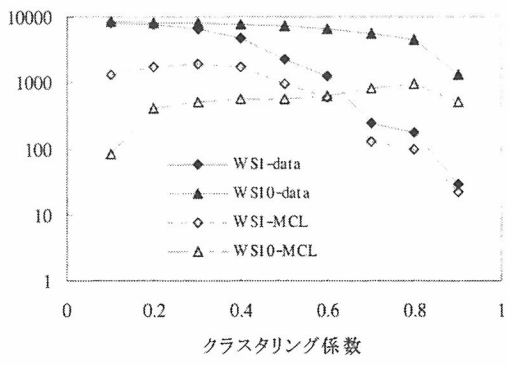

図 6. MCL 結果

4.3 MCL クラスタリング評価

MCL のクラスタリング結果に関して、Modularity（Q 値）と F 尺度による評価を行う。Modularity は、クラスタリングの精度を評価するさいに用いられる指標である[16]。また、F 尺度は、精度（precision）と再現率（recall）のトレードオフの関係性を用いて最適な条件の選択が可能となる。赤間ら(2007)[17]は、MCL のクラスタリング結果の精度にこの 2 つの指標を導入し、共起単語ペアの頻度数を閾値とする意味ネットワークからクラスタリング結果が最適であるネットワークを選定した。さらに、この 2 つの指標を基準に、MCL 計算のパラメータである Γ 係数に注目し、最適なパラメータ値を求めた研究も報告されている[18]。

先行研究に倣って、本研究においては、MCL によって得られたクラスタリングの適切さを図る指標として、Q 値を使用する。また、Q 値を精度とし、再現率を全出現単語に対する各閾値のノード数の割合として計算した F 尺度の 2 つの指標を考慮する評価方法を用いて、最適な共起単語ペアを選定する。

図 7 に、各閾値で得られた MCL クラスタリングの Q と F 値を計算したものを表す。WS1 においては、Q の最大値は 0.6、F 値は 0.4 とピークが異なる。再現率であるデータ数を考慮し、F 値のピークのときを選択する。WS10 においては、閾値に比例して Q の値は大きくなりピークがないので、F 値のピーク 0.7 を選択する。選出した WS1 と WS10 のノード数はそれぞれ 4710 と 5541 であり、5000 程度のほぼ同じ大きさのネットワークである。

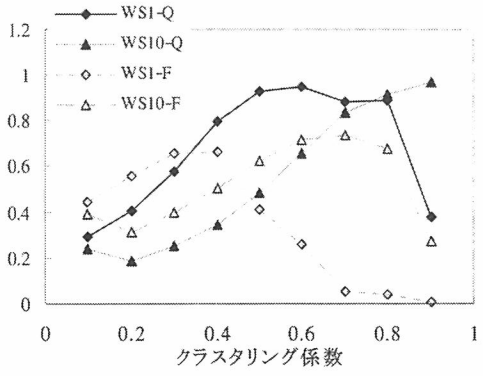

図 7. 各クラスタリング結果の Q/F 値

図８に WS1,WS10 の MCL 結果から得られたクラスターのコンポーネントサイズの分布を表す。単語ノードがどの単語ともグルーピングされない、コンポーネントサイズ１のデータを考慮にいれないと、WS1 の分布は、べき乗分布のような形状をしている。一方で、WS10 に関しては、緩やかな右下がりの分布を描いている。

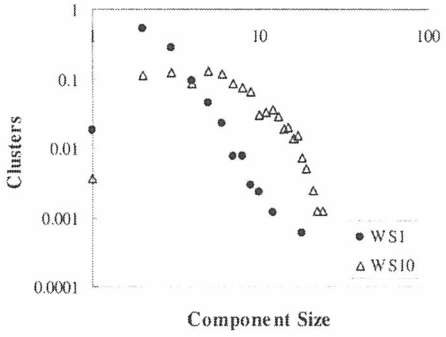

図 8 . MCL コンポーネントサイズ分布

4.4 RMCL 結果

RMCL の入力データある仮想隣接行列は、MCL プロセス過程においてできる各ステージクラスターと収束 MCL クラスターの関係を基に作成される。従って、収束までのステップ数だけ RMCL のバリエーションができるが、どのステージの結果を採択するにあたっては、クラスターの分散など様々な指標を閾値として決定することが可能であると考えるが、本研究では、各ステップに対して得られた RMCL クラスター結果の Q 値を求め、精度 Q が最大になるものを選んだ。

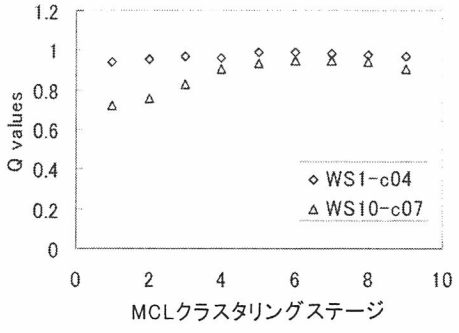

図 9. MCL ステージ別による RMCL の Q 値

図９に、WS １、WS10 における、収束前の MCL クラスタリングステージごとの RMCL クラスターの Q 値をプロットしたものを示す。WS1 の場合は、ステージ５のときに Q は最大値をとり、WS10 のピークは、ステージ７のときであった。RMCL のクラスター結果に関しては、WS1 は、収束クラスター数は 1567 であり、W10 は 725 クラスターであった。

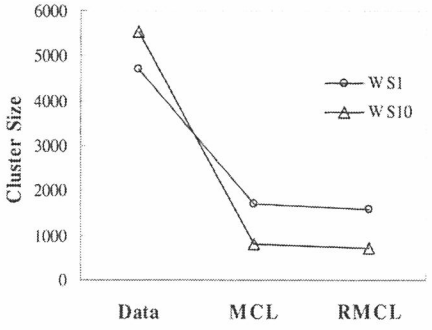

図 10. クラスター数の推移

図 10 に、各幅におけるネットワークが MCL/RMCL によってクラスターサイズの推移を示す。共起単語ネットワークデータから、順次グラフクラスタリング MCL と RMCL を施すことによって、階層的に小さなネットワークを実現している。

表 1：コンポーネント数の平均（SD）値

	MCL 平均(SD)	RMCL 平均(SD)
WS1	2.81(1.34)	1.08(0.30)
WS10	6.87(4.10)	1.06(0.27)

表 1 に、MCL.RMCL で得られたクラスターのコンポーネント数の平均と標準偏差を計算したものを示す。MCL のコンポーネント数に関しては、WS1 と比較して WS10 のほうが平均値も高く、ばらつきも大きいことが分かる。一方、RMCL のコンポーネント数においては、両者ともほぼ同様の結果を示しており、非常に小さなクラスターが出来ていることが確認できる。

5. まとめ

本研究は、テキストデータとして福音書を対象にして、２種類のウィンドウ幅で得られた単語の共起情報を基にして、意味ネットワークの特性を調べた。ウィンドウ幅の小さいネットワークに関しては、スケールフリー性が確認された。また、両ネットワークとも、クラスター性の高さが確認された。さらに、クラスタリング係数を閾値として複数のネットワークを作成し、グラフクラスタリング RMCL を行った。グラフクラスタリング結果の評価に関しては、クラスタリングの精度とデータの再現率を考慮した２つの指標を基準にして、分析に最適なデータの選出を行う手法を示した。

謝辞

本研究は、科学研究費補助金若手研究（B）19700238 の支援を受けて行ったものである。東京工業大学・佐伯元司研究室の三浦新氏が作成した Windowing 方法によるテキスト共起情報取得のプログラムを使用することにより分析データを円滑に作成することが可能となった。また、RMCL の適用・評価に関して、東京工業大学・赤間啓之准教授ならびに鄭在玲さんから有益な助言を頂いたことに心より感謝いたします。

参考文献

[1] Fellbaum, C. (ed.) , *WordNet: An Electronic Lexical Database*, Cambridge, MA, MIT Press, 1998.

[2] Roget, P. M.: Roget's Thesaurus of English Words and Phrases, *http://www.gutenberg.org/etext/10681*, 1991.

[3] 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.

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

[5] Jung J., Miyake M., and Akama H. Recurrent Markov Cluster Algorithm for the Refinement of the Semantic Network, *LREC2006*, pp.1428-1432, 2006.

[6] Nestle-Aland, *Novum Testmentum Graece 26th edition*, German Bible Society Stuttgart, 1979.

[7] 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.

[8] 三宅真紀、Jaeyoung Jung、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.

[9] 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、*人文科学とコンピュータシンポジウム論文集 じんもんこん-2006*、pp.161-165、2006.

[10] Watts, D. and Strogatz, S., Collective dynamics of 'small-world' networks, *Nature*, 393:440-442, 1998.

[11] Barabasi, A.L. & Albert, R., Emergence of scalling in ramdom networks, *Science*, 286, pp.509-512, 1999.

[12] Dorow, B. et al., Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination, *MEANING, 2005*.

[13] Ravasz, E., Barabasi, A. L., Hierarchical Organization in Complex Networks, *Physical Rev. E*, 67, 026112, 2003.

[14] Dorogostev, S.N., Golstev, A.V., and Mendes, J.F.F, Pseudofractal Scale-free Web, *Physical Rev. E*, 65, 066122, 2002.

[15] Gfeller, D., et al., Synonym Dictionary Improvement through Markov Clustering and Clustering Stability, *ASMDA*, 106-113, 2005.

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

[17] 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*、No.49, pp.49~56, 2007.

[18] Miyake, M. and Joyce, T., Mapping out a Semantic Network of Japanese Word Associations through a Combination of Recurrent Markov Clustering and Modularity, *the 3rd Language & Technology Conference (LTC'07)*, pp.114-118, 2007.