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

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、Van Dongen (2000) が提唱したグラフクラスタリング MCL（Markov Clustering）を発展させたアルゴリズム RMCL （Recurrent Markov Clustering）を言語資料に適用して、共観福音書の意味ネットワークの半自動的構築を試みる。同時に、ネットワーク特性量であるクラスタリング係数を用いてノイズワードを除去するという新しいデータ処理の方法を提案、共観福音書に適用する。さらに、意味ネットワークの視覚化のため、web アプリケーションを開発して、単語・概念間の関係性を動的に表現する。さらに、RMCL クラスタリングによる共観福音書の細分化と従来の伝統的な共観表での分割とを比較し、本手法の有効性について考察する。

Implementing a Semantic Network of the Synoptic Gospels based on Graph Clustering

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

In this research, we propose the semi-automatic construction of a semantic network for the Synoptic Gospels employing a graph clustering algorithm called Recurrent Markov Clustering (RMCL), which is an improved form of Markov Clustering (Van Dongen, 2000). A new method of data processing for graph clustering is used where noise words are eliminated by a clustering coefficient which is a feature of the network. In order to visualize the semantic network, a web application is developed that dynamically represents the relationships between words and concepts. The effectiveness of the proposed technique is discussed by comparing aspects of RMCL clustering with the parallel synopsis tables traditionally used in biblical studies.

1. まえがき

グラフ理論を利用したネットワーク分析は、テキストマイニングやテキスト分析の領域においても、ベクトル空間言語モデルと同様に、意味知識や推論機構の体系を表現するのに直感的で有効な手段として着目されている。また、www をはじめとする大規模な実世界ネットワークの構造として知られている、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 のシソーラスや WordNet を用いて証明している[1]。

また、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[2]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[3]。

単語のグラフクラタリングに関しては、Van Dongen が提唱した MCL（Markov Clustering）[4]は、直感的な感性に合致した意味クラスタリング結果を提供するが、文書データの場合は、単語の頻度分布の偏りが原因となって、クラスターサイズの著しい不均斉が起きるという問題がある。これに対して、Jung[5]らは、ＭＣＬから発展したアルゴリズム RMCL （Recurrent Markov Clustering）を提案し、連想辞書に適用した。

本研究では、RMCL による共観福音書の意味ネットワークの半自動的構築を試みる。また、ネットワーク特性量であるクラスタリング係数に注目して、分析データの作成で必要なノイズワードの処理方法を提案する。さらに、単語・概念間の関係性を動的に表現する web アプリケーションを開発して、意味ネットワークの視覚化を行う。

2. RMCL グラフクラスタリング

本研究で適用する RMCL の基となる 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[4]。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[6]。

さらに Jung ら[5]は、単語の頻度分布の偏りから生じる、MCL クラスターサイズの著しい不均斉を解消するために、再帰的アルゴリズム RMCL を考案した。この手法は、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、ネットワークの階層化、ダウンサイジングを実現する。その結果として、図 1 に示すような、単語・概念間における適正な意味ネットワークの構築を可能にしている。

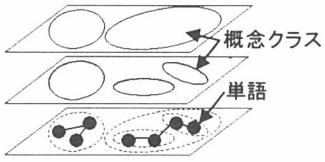

図 1：意味ネットワークの階層化

3. クラスタリング係数

Watts と Strogatz [7]によって、「知り合いの知り合いが知り合いである確率」を表す指標として、導入されたクラスタリング係数は、ノード間の繋がりの度合いを表す重要な指標である。ここで、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 2 のようなスターグラフを形成し、また係数 1 の場合は、図 3 のような完全グラフを形成する。

図 2：C=0　　図 3: C=1

最近では、Dorow ら[8]によって、単語の Curvature（クラスタリング係数と同意語）に基づき、隣接する単語の意味的関連性を測定することを提案した。さらに、Curvature を閾値として、多義語などの曖昧性の強い単語の除去を行い、Curvature クラスタリングを British National Corpus(BNC) に適用している。

4. 意味ネットワークの作成方法

4.1 共起ペア情報の取得

テキストは、Aland の共観表[9]との比較分析も考慮して、Nestle-Aland26 版を使用する[10]。そして、RMCL の計算に必要な単語の隣接行列データは、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。出現単語の持つ概念だけではなく、統語的な要素の抽出も考察するために、出現単語の形態素処理は施さず、設定ウィンドウ幅は、幅は比較的小さな幅 2（各インスタンスの前後 2 単語を抽出）を用い、また 2 回以上出現するペアに限定する。以上の条件を満たす共起単語ペアは、7277 単語であった。

4.2 ネットワーク特性量

ここで、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ー係数を計算して、共起単語ペアのグラフの特性を調べる。

スケールフリーネットワークの構造を知る方法として、Balabasi と Albert[11]が導入した、次のような次数分布とべき乗則の関係を使用する。

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

単語をノードとする無向グラフにおける次数分布 P(k)と、それにフィッテングさせたべき乗則分布（指数は 1.4）をプロットしたものを図 4 に示す。ここで、次数平均は、38.4 であった。この結果から、共起単語ペアのグラフは、スケールフリーネットワークを形成していることが明らかになった。

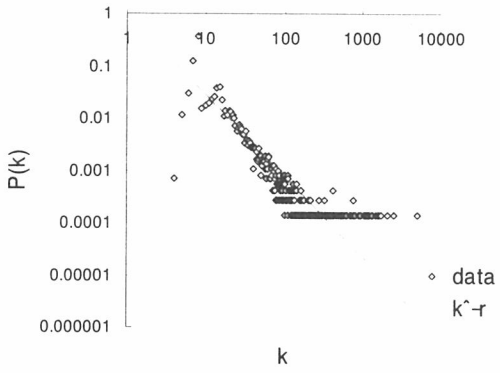

図 4：次数分布

図5に、次数に対するクラスタリング係数をプロットしたものを示す。また、全単語における、クラスタリング係数の平均は、0.68であった。

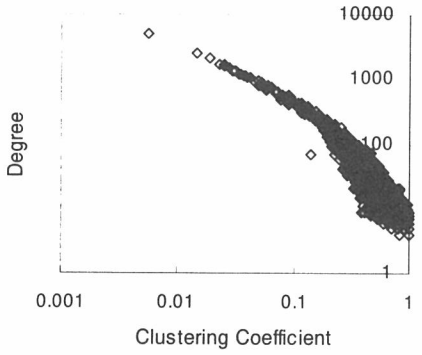

図 5：次数－クラスタリング係数

4.3 共起単語ペアの選定

全データを MCL にかけた場合、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ため、クラスタリング係数を基準にして、共起単語ペアを選定する。

本研究では、意味生成にさほど関与しないノイズワードが持つ高次数かつ隣接単語間の繋がりの低さという特徴に注目し、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。0.1以下の単語をみると、και, δε, ο, εν をはじめとした、接続詞、定冠詞といった単語である。そこで、係数を0.1ごとに区切り、各係数値以上の単語からペアデータを作成した。

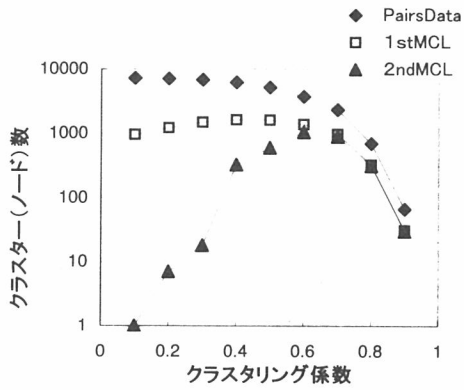

図 6：クラスター数の推移

図2に、各クラスタリング係数における、ペアデータのノード（単語）数、第1回 MCL、第2回 MCL の クラスター数の推移を示す。平均値以上のクラスタリング結果からは、ノード数が 5 以下の小クラスターの集まりで、隣接関係が密であり、熟語を表すクラスターが多い。係数 0.3 以下のクラスタリングは、クラスターサイズの分散も大きく、第2回 MCL の収束クラスター数の少なさから、多くのノイズワードが含まれていると考えられる。以上から、係数 0.4～0.6 の間が閾値として採択するのが適している。今回は、第 1、2 回 MCL 圧縮の度合いを基準にして、係数 0.5 以上の共起単語ペアデータ(5195 語)を意味ネットワーク生成の対象とする。

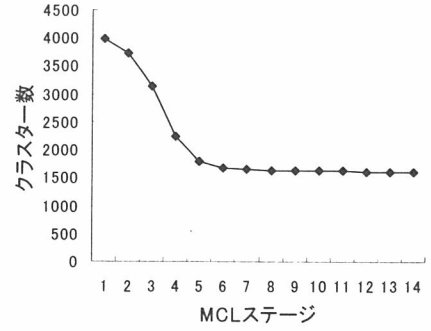

図 7：第一回 MCL プロセス

4.4 RMCL の適用

まず、第 1 回 MCL を計算し、収束クラスター数を求めた。MCL の計算は、第 13 ループ目でほぼ収束し、その収束クラスター数は 1613 であった。図 7 に、第一回 MCL のクラスター数の推移を示し、次に、第 2 ループ目 MCL クラスターから、収束クラスターを再隣接化し、再度 MCL を計算した。第 2 回 MCL は 第 6 ループ目で収束し、収束クラスター数は、602 であった。

5. 共観福音書意味ネットワークの実装

5.1 システム構成

RMCL により生成された意味ネットワークの視覚化に関しては、WebMathematica による web アプリケーション
http://nerva.dp.hum.titech.ac.jp/semnet/RmclNet/synopsis.jsp
を開発し、RMCL&MCL クラスター単位でグラフを描画しながら、意味ネットワークのインタラクティブな表示を可能にしている。図 8 に、簡単なシステム構成図を示す。まず、ユーザがアクセスする画面は JSP と WebMathematica の MCP スクリプトによって、書かれている。WebMathematica を通して、クラスタリング結果のデータとの通信を行い、また Mathematica のカーネルを起動させて、グラフ描画の計算を実行している。

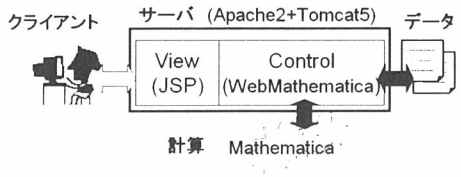

図 8：システム構成図

5.2 アプリケーション部分

Web アプリケーション部分は、予め全ての RMCL クラスターがリストで表示されおり、プルダウン機能で各 RMCL クラスターを選択すると、それに含まれる MCL クラスターと、その MCL クラスターに属するノードが順次選択されて、結果として図 9 に示すような、選択ノードに隣接するグラフ図が表示される。デフォルトのグラフ図は、選択ノードからパスの長さが 1 のノードを抽出して表すが、オプション指定を行うことで、パスの長さの設定が 1 から 5 まで出来るようになっており、図 10 に示すような、より広範囲なネットワーク図を描くことも可能である。

また、各 RMCL、MCL クラスターのラベルは、クラスター内で次数の一番多いノードを代表ノードとして選択している。さらに、グラフ表示のノードに連結している他ノードをクリックすると、他の RMCL 又は MCL クラスター内へリンクすることも可能である。そして、選択ノードは、赤色で示し、選択したノードと同一の MCL クラスターに属しているノードは、ピンク色で表す。代表ノードに関しては、青色のノードで表されており、このように、一目で MCL クラスターと代表ノードが識別できるようになっている。

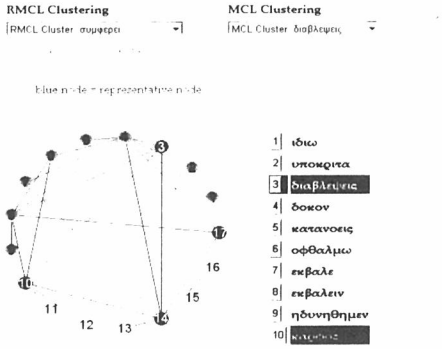

図 9：共観福音書ネットワーク

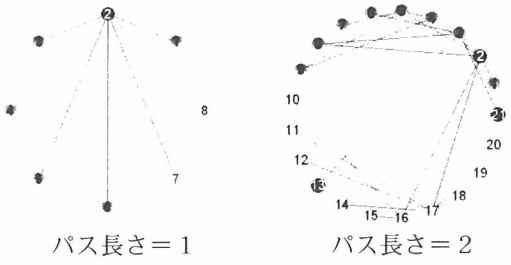

図 10 :パス長によるグラフの変化

6. 考察

RMCL クラスタリング結果から、各クラスターが、テキストの節や共観箇所に対応していることが分る。

一例として、以下のような RMCL クラスター（3MCL クラスター、20 単語）を取り上げる。MCL クラスターを点線ボックスで表し、マタイ・ルカ共通単語は太字、マタイ・マルコ共通単語は下線、マルコ・ルカ共通単語は斜体で識別する。

MCL1
δοκον/δοκος(log beam of wood), οφθαλμω(eye), *ιδιω(one's own)*,αδελφου(brother)
εκβαλε/εκβαλειν(take out),
διαβλεψεις(see clearly),
ηδυνηθημεν (can), **κατανοεις(notice)**, ερεις(say) ,
υποκριτα(hypocrite),σκανδαλιζη(cause to sin)

MCL2
οδοντα/οδοντος(tooth), οφθαλμον/**οφθαλμου** (eye)

MCL3
αδελφε(brother), εκβαλω(take out),καρφος(speck)

これらの単語から、「目の中の梁（マタイ 7 章 3 節：なぜあなたは、自分の兄弟の目にある屑を見て、自分の目に梁には気づかないのか。…）[佐藤研訳]」の箇所（ペリコーペ）とマタイ 5 章 38 節の「目には目を歯には歯を」の言葉が容易に想起される。さらに、Aland の共観表と比較すると、共観表 no.81. "On judging (Mt 7:3-5, Lk 6:41-42)"に対応していることが分る。表 1 に、該当する共観箇所の一部を抜粋する。MCL クラスターは「目の中の梁」という譬えの単位であるに対して、Aland の共観表は、譬え話の意味内容から「さばくな」という訓戒として、範囲を広げて前後の節を付加していることが分る。

表 1：共観表 no. 81（cf. no. 68）一部抜粋

マタイ 7:3	ルカ 6:41
τι δε βλεπεις το **καρφος** το εν τω **οφθαλμω** του αδελφου σου την δε εν τω σω **οφθαλμω δοκον** ου **κατανοεις** η πως ερεις τω αδελφω σου αφες **εκβαλω** το **καρφος** εκ του **οφθαλμου** σου και ιδου η **δοκος** εν τω **οφθαλμω** σου	τι δε βλεπεις το **καρφος** το εν τω **οφθαλμω** του αδελφου σου την δε **δοκον** την εν τω *ιδιω* **οφθαλμω** ου **κατανοεις** πως δυνασαι λεγειν τω αδελφω σου *αδελφε* αφες **εκβαλω** το **καρφος** το εν τω **οφθαλμω** σου αυτος την εν τω **οφθαλμω** σου **δοκον** ου βλεπων

7. まとめと今後の課題

本研究は、共観福音書を対象にして、グラフクラスタリングによる意味ネットワークの半自動構築を行った。得られた結果は、共観表との対応も認められ、複数テキストから共通部分抽出法としての有効性を示した。今後、共観福音書意味ネットワークが表す構造と従来の共観表との詳細な比較を行い、また、ネットワークのグラフ図の改良をして、出力結果が一目で理解しやすい表示方法を研究する。そして、クラスタリング結果を共観福音書問題の計量分析に本格的に適用する予定である。

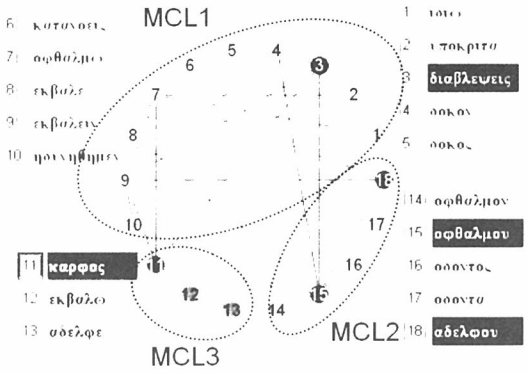

図 11：MCL クラスター例

参考文献

[1] 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

[2] 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.

[3] 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.

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

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

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

[9] Aland, K. (ed.), Synopsis Quattuor Evangeliorum, 15th edition, Stuttgart, 1996.

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

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