3K-03

異種ネットワークの結合とノード視点に基づくコミュニティ抽出*

片岡 拓弥† 須貝 康雄‡

千葉大学§

1 はじめに

現実社会における人間関係, 鉄道や道路における交通ネットワークなどの多様なネットワークが身近に存在している。また, 近年ではTwitterなどのWeb上におけるSNSなどに対する関心が高まっている。これらのネットワークにおいて, ノードは均一でなく, ノードごとに異なる視点, 関係性を持っている。ネットワーク全体においても全体が均等でなくリンクが密な部分や疎な部分が混在し, コミュニティ構造を有することが指摘されており, 様々なコミュニティ抽出法が研究されている [1]。

また, 現実世界にはグラフ表現されていない文章や表など様々な形式のデータが多くある。これらをグラフ表現することにより内容の理解を視覚的に手助けするとともに, 様々な異なるデータをグラフとして組み合わせて分析することが望ましいと考えられる。文章をグラフ表現する研究は行われているが, 文章が論理的に整理されているものが対象である場合が多い [2]。

そこで本稿では, word2vec[4] という文章中の単語をベクトル表現する手法を用いて各単語のベクトル間の距離を計算し, 重み付けをすることにより単語間にリンクを生成し, 文章データをグラフ表現する手法を提案する。提案手法により異種のデータをグラフとして組み合わせることが可能である。グラフとして組み合わせた異種ネットワークにおいて, CDB法 [3] を用いてコミュニティ抽出を行い得られたコミュニティに関する考察を行う。

2 共通コミュニティ抽出手法

文献 [5] では各ノードごとの視点 (以下シードノード) に立った時, それぞれ主観的なコミュニティを持っているとしており, 探索型コミュニティ抽出手法が有用であるとして, コミュニティ濃度に則って提案されたCDB法に着目し, 複数ノードの視点におけるコミュニティ抽出法を提案している。

CDB法は, コミュニティ探索の始点となるシードノードを選択し, コミュニティを構成する初期ノードとする。次にコミュニティ内のノードと連結している各ノードをコミュニティに加えたとした場合のコミュニティ濃度 Cd を計算し, そのうち最も Cd の値が大きいノードを実際にコミュニティに追加する。これを繰返し, 終了条件を満たすまで探索を続ける。

文献 [5] のアルゴリズムは3つ以上のシードノードに対しては複雑になり, 実用的ではない。そこで, 複数のシードノードを最初から1つのノードとして結合することにより, 3つ以上のノード視点からでも比較的容易に共通コミュニティを抽出できるように改良した。

3 コミュニティの評価

文献 [3] では, コミュニティであることの評価尺度をコミュニティ濃度 Cd として定義している。Cd 値は, 部分グラフ内の平均次数と, 外部に接続している部分グラフ内の境界ノードの平均次数との比である。部分グラフ内の平均次数が外部に接続している部分グラフ内の境界ノードの平均次数より大きい時にその部分グラフはコミュニティになる。コミュニティ濃度 Cd の値が大きいということは, コミュニティ内外の相対的な密度差が大きく, かつコミュニティ内部の密度が外部との繋がりに比べて大きいということを意味する。本稿では, Cd 値の極大値が見られる各時点で, それまでに抽出されたノード群をそれぞれコミュニティと見なす立場を取る。

4 文章データのグラフ化

word2vecとはニューラルネットワークを用いた単語の分散表現 (単語ベクトル) を作成する手法である。近年では, ニューラルネットワークによる学習モデルであるSkip-gramモデルがMikolovらによって新たに提案され [6], 従来の手法に比べ飛躍的に精度が向上した。

word2vecを用いると文章中の各単語に対応する単語ベクトルを得ることができるため, 単語間の距離, すなわちベクトル間の距離を計算できる。したがって本稿では, 単語ベクトルを正規化し, 全単語間のユークリッド距離を計算する。また, 文章をグラフ化する場合, 単語間の距離が遠いほど重みが小さいことが望ましい。大小関係の反転には逆数を用いることが多いが, ユークリッド距離の大きい領域では, 値の差が大きくても逆数の差は小さくなるため, 値の差を有効に表現できない。そこで, 本稿では文章中の全単語間のユークリッド距離の最大値からそれぞれの単語間のユークリッド距離を減じた値を用いる。

$$w_{ij} = d_{\max} - d_{ij} \tag{1}$$

また, 生成した重みを全て用いて文章をグラフ化する場合, 重み付き完全グラフ (ネットワーク) が生成される。しかし, そのまま用いると膨大な計算量となるため, リンクの除去を行う。具体的には, 生成されたリンクの重みがネットワーク全体に与える影響が少ないと考えれられるある値以下の場合, ネットワークから除去する。また, 同一単語ベクトル間の距離は計算の対象外であるため, 自己閉路がないネットワークが生成される。

5 異種ネットワークの結合

文章や表データから生成したネットワークや既存のネットワークを結合することを考える。本稿では

*An Integration of heterogeneous networks and Community extraction based on the nodes viewpoint

†Kataoka Takuya

‡Sugai Yasuo

§Chiba University

結合するネットワーク間に共通の単語が存在することを前提とする。そして, 結合するネットワーク間の共通単語は共通ノードとして扱う。異種のネットワークの結合において, リンクの除去を行うため, 非連結になることが考えられる。その場合, コミュニティ抽出はシードノードが属する連結成分を対象とする。

6 計算機実験

6.1 実験設定

文献 [1], [6] を対象データとしてネットワーク化したグラフと, 英和辞書 [7] をネットワーク化したグラフの 3 つの異種ネットワークを結合する。本研究ではノード視点のコミュニティ抽出を前提としているため, 文章を読む前に得られる文章のタイトルを視点としたコミュニティ抽出を行う。したがって, 文献 [1], [6] のタイトル中の名詞を複数シードノードとしてコミュニティ抽出を行う。

6.2 異種ネットワーク結合結果

3 つのネットワークを結合したネットワークの最大連結成分を図 1 に示す。

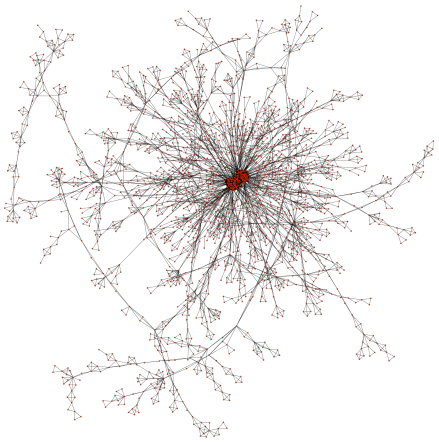

図 1: 異種ネットワークの結合結果

6.3 コミュニティ抽出結果

図 2 に Cd 値の推移の一部を示す。横軸はコミュニティを成長させる各時点におけるコミュニティ内ノード数を示している。また, 1 つ目の極大値の時点でのコミュニティ内ノード 34 ノード中の一部を表 1 に示す。

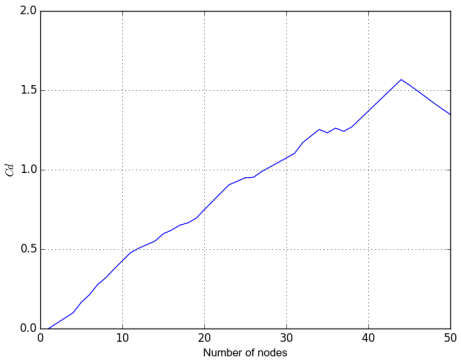

図 2: Cd 値の推移

表 1: コミュニティ内ノード

距離	表現	長さ
performance	成績	record
score	wording	interlude
register	代表	記録
点数	性能	空間
distance	acting	wording
rendering	肖像	言い回し

6.4 考察

abstract 内の名詞とコミュニティ内ノードを比較すると共通してる単語が見られるため, コミュニティ抽出により論文の内容を示せており, abstract に記載されていないコミュニティ内ノードは, abstract では示せていない論文の内容を把握するために効果的であると考えられる。また, ネットワーク化した論文に辞書ネットワークを結合し, コミュニティ抽出することにより未知の言語で書かれた文章の内容理解の手助けとなることが考えられる。

7 まとめ

本稿ではグラフ表現されていない文章データをネットワーク化し, 既存のネットワークや異なる文章から作成したネットワークを結合してコミュニティ抽出を行った。文章から文章のタイトル視点でコミュニティ抽出することにより文章の内容理解に効果的であることが分かった。今後, 様々な研究分野の論文や既存のネットワークをグラフ化することにより結合し, ノード視点でコミュニティ抽出することにより, 萌芽分野の発見などにも応用できると考えられる。

また, 同じ単語でありながら異なる意味で用いられている場合の処理や, 論文中の式や記号などのノイズの除去が今後の課題である。

参考文献

[1] M. E. J. Newman, Park J: “ Why social networks are different from other types of networks ”, Phyical Review E, Vol.68, No.3, 036122(2003).

[2] 野坂卓矢, 原口誠: “ 文章の隣接グラフ化とグラフマッチングに基づく判例文の類似度計算 ”, 第 25 回人工知能学会全国大会論文集, 3H2-OS3-8(2011).

[3] 高橋篤，荒井幸代: “任意ノードの視点からのコミュニティ抽出手法”, 第 23 回人工知能学会全国大会論文集, 2I2-2(2009).

[4] word2vec: “ https://code.google.com/p/word2vec ”, 2018 年 1 月 10 日アクセス.

[5] 多比羅大, 須貝康雄: “ ノード視点におけるネットワークからのコミュニティ抽出 ”, 電気学会全国大会講演論文集, 3-032, pp.42-43(2015).

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: “Efficient estimation of word representations in vector space ”, ICLR, arXiv:1301.3781(2013).

[7] “ http://tokoton-eitango.com/ ”, 2018 年 1 月 10 日アクセス.