6T-8

共起ネットワークを用いたクラスタ性によるテキスト分類

小林　雄太[†]　村上　裕一[††]　中村　真吾[†††]　橋本　周司[†††]

早稲田大学先進理工学部[†]　早稲田大学大学院先進理工学研究科[††]　早稲田大学理工学術院[†††]

1. はじめに

近年，膨大な文書の中から有用な情報を抽出するテキストマイニングが盛んに行われている．文書情報のうち，異なる単語が同時に出現する共起のパターンは重要な情報であり，共起ネットワークによって文書を視覚化することでユーザの文章解析に大きく役立っている．共起ネットワークを用いた従来研究では，文書全体における語の共起関係がその文書の主張に至る道筋であるとして，主張を表す語を抽出する Keygraph や，さらにスモールワールド性を意図的に高めることで主張性を促進する手法が提案されている[1][2]．しかし，Keygraph は土台と主張という簡単な 2 層構造でしか文章を捉えていないため，複雑な論理構造の文章には適用できない．また，スモールワールド性も特定の文書のみ評価可能という問題点がある．

本稿では，テキストの論理構造を文章のまとめ方ととらえ，語の出現パターンを表す共起ネットワークを利用して，論理構造の特徴を抽出し，分類する手法を提案する．具体的には，文書から共起ネットワークを作成し，tf-idf 法[3]によって語に重みを与える．さらに，複雑ネットワークの閾値グラフに基づき閾値を変化させ，閾値とクラスタ性の関係であるクラスタ関数から，文書の論理構造を評価する．本手法は，従来では困難であった文章全体のまとめ方という観点で複雑な論理構造を持つ文書を評価できることから，新たなテキスト分類手法ということができる．実験では夏目漱石と森鴎外の小説を提案手法によって分類し，提案手法が実際の共起ネットワークに適用可能であること，及びその有効性を確認した．

2. 提案手法

2.1 閾値モデルによる共起ネットワーク

提案手法のネットワークは，共起と tf-idf 値，閾値により構成される．まず文書に対して文単位での共起を調べる．次に tf-idf 値が上位を占めるストップワードを除いた名詞から，tf-idf 値を重みとしたノードを持つ共起ネットワークを作成する．

Text classification with cluster property of co-occurrence network

Yuta Kobayashi[†]・Yuichi Murakami[††]・Shingo Nakamura[†††]・Shuji Hashimoto[†††]

Department of Applied Physics, Waseda University[† ††]

Faculty of science and engineering, Waseda University[†††]

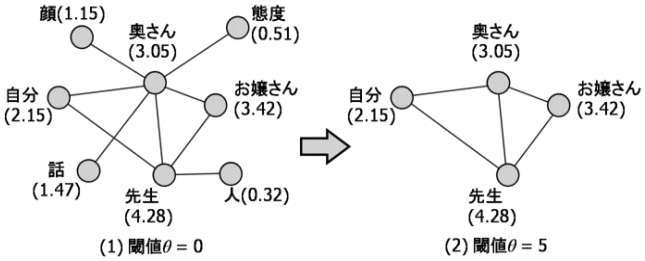

図 1 閾値によるネットワーク構造の変化

ここで複雑ネットワーク[4]の閾値モデルの理論を導入する．各語を結ぶリンクの強さを各々の語の重みの和とし，リンクの強さが閾値以上であればリンクを置き，閾値未満であればリンクを置かずにネットワークを構成する．したがって，閾値を変化させるとそれに付随してネットワーク構造が変化する．図 1 にその一例を示す．このとき，クラスタ性の指標である，ネットワークに三角形構造が含まれる割合を示す平均クラスタ係数[4]も構造と共に変化する．したがって，平均クラスタ係数は閾値の関数となる．本稿ではこれをクラスタ関数と呼び，この関数の特徴からテキストの分類を行う．

2.2 クラスタ関数

各ノードの重みが指数分布に従う理想的な閾値グラフの条件下では，クラスタ関数は複数のピークを持つ．閾値の値が 0 である場合には，全てのノード間にリンクが存在する．閾値を大きくすると，まず重みの小さいノード同士を結ぶ弱いリンクが切れ，三角形構造が失われるためクラスタ係数は減少する．次に，ある閾値 θ_1 まで大きくなると，重みの小さいノードと大きいノードを結ぶリンクが切れる．このとき，切れるリンクは三角形構造を持たないツリー構造部分のため，クラスタ係数は逆に大きくなる．さらに閾値が大きくなり θ_2 を超えると，大きい重み同士を結ぶリンクが切れ，再びクラスタ係数が減少する．クラスタ関数のピーク値 θ_1，θ_2 は，ネットワークに含まれるノイズ構造，ツリー構造，コア構造に分割する閾値となる．図 2 に，理想的な閾値モデルのクラスタ関数を示す．文書構造は，ノイズ構造が多いほどランダム性が，ツリー構造が多いほど展開性が，コア構造が多いほど中心概念性があると言える．

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.

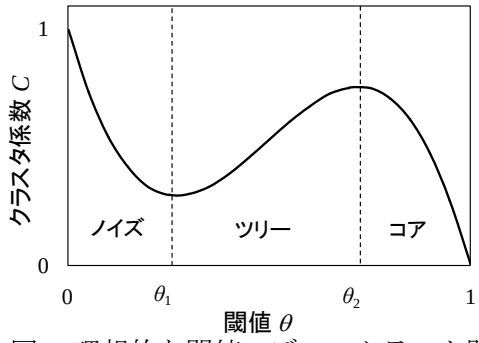

図 2 理想的な閾値モデルのクラスタ関数

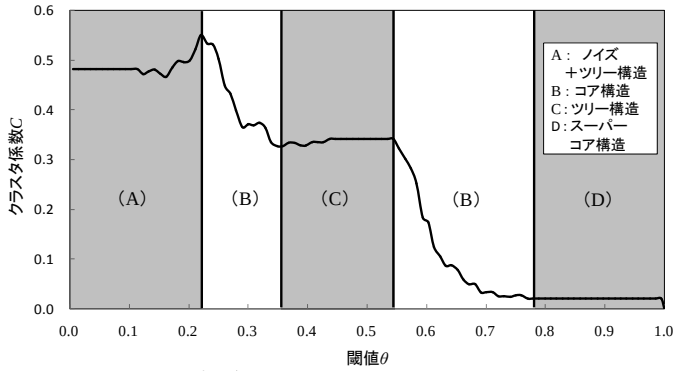

図 3 森鴎外「鶏」のクラスタ関数

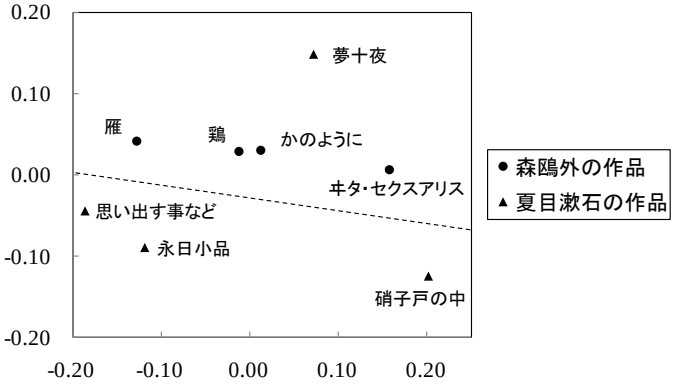

図 4 クラスタ関数の構造幅による主成分分析

3. 実験

3.1 文書におけるクラスタ関数

実際にいくつかの小説のクラスタ関数を作成し，その傾向を調べた. 一例として, 図 3 に小説「鶏」のクラスタ関数を示す．実際の文書の場合，閾値を増大させると，始めは最も小さいリンク強度を持つノイズ構造と一番外側のツリー構造とが混じりながら，リンクが切れる．これは理想的な閾値グラフとは異なり，共起した語のみが繋がっているため，出現頻度の低い重みの小さい語は他の語と繋がりにくいためである．そのため，弱いリンクのツリー構造をしており，ノイズ構造とツリー構造の組み合わさった構造をしていると推測できる．その後は，コア構造とツリー構造が交互に現れる．最後には，非常に重みの大きいノードのみで構成された強固な構造だけが残る．この部分は，非常に重要で中心的な概念であるスーパーコア構造を持つ.

このように実際の文書のクラスタ関数は，ツリー構造とノイズ構造が交互に現れる階層構造をしており，最終的にはスーパーコアが残る構造をしている．次のテキスト分類実験では，これらの特徴を利用し，文書の評価と分類を行う.

3.2 テキスト分類

提案手法を用いて夏目漱石と森鴎外の小説各 4 作品，計 8 作品について，tf-idf 値が上位 100 位の語群からクラスタ関数を求めた. ただし, 上位 100 位の語の重要性が同じになるよう，各小説の文書の長さを同程度になるよう編集してから作成した． 各作品について得られたクラスタ関数のピーク閾値から 5 つの構造領域に分割し，構造の幅の大きさをその文書の特徴量とした.

図 4 に得られた特徴量に対し主成分分析を行い, 第 1 と第 2 主成分をプロットした結果を示す．漱石と鴎外の作品を線形分離できていることが確認できる．ただし, 夏目漱石の「夢十夜」だけは分離することができなかった． これは, 漱石が夢の内容を淡々と綴った試作的なものであり，漱石の著作の中でも異色な作品である．そのため, 同じ著者の作品であっても，文調が異なるため漱石の作品として分離できなかったと考えられる.

4. まとめ

共起ネットワークからクラスタ関数を求め，その構造を特徴量とすることで, テキスト分類をおこなった．実験では, 実際の小説を用いて分類を行い, その有効性を確認した．今後は，さらに高精度なクラスタ関数の計算法を検討するとともに，より多くの文書の評価と分類を試みたいと考えている.

謝辞

本研究の一部は，グローバル COE プログラム「グローバル ロボット アカデミア」，科学技術振興機構 CREST 研究「人を引き込む身体的メディア場の生成・制御技術」の研究助成を受けて行われた.

参考文献

[1]大澤幸生, ネルスＥ．ベンソン, 谷内田正彦, "KeyGraph：単語共起グラフの分割・統合によるキーワード抽出", 電子情報通信学会論文,J82-D21,No.2,pp.391-400, 1999.

[2]Xavier Llorà et al.,"Discovering Chance Scenarios using Small-World KeyGraph and Evolutionary Computation ",The First International Workshop on Chance Discovery, pp.51-61,ECAI 2004.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", International Journal on Digital Libraries, vol. 24,pp. 513-523, 1988.

[4] Watts, D. and Strogatz, S., "Collective dynamics of small-world networks", Nature, Vol. 393, pp. 440-442 ,1998.

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.