共起語グラフの複雑性指標によるテキスト評価

小林　雄太†　中村　真吾††　橋本　周司†††
早稲田大学先進理工学部†　芝浦工業大学††　早稲田大学理工学術院†††

1. はじめに

近年，膨大な文書から有用な情報を抽出するテキストマイニングの研究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，語をノード，共起関係をエッジとした共起語グラフは，文書を可視化することでユーザの文書理解に大きく役立っている．従来研究では，共起語グラフから文書のキーワードを抽出する Keygraph[1] や，さらに共起語グラフの複雑ネットワークの指標を意図的に高めることで，可視化による文書理解を促進する手法[2]が提案されている．

山中ら[3]によると，文書理解の妨げとなる文書の難しさには大きく分けて 2 つの解釈があると考えられている．ひとつは文書の内容が理解できない事による難しさであり，もう一方は文書の構造が複雑であることによる難しさである．前者は文書に書かれている語句の難易度によって決まり，近藤ら[4]により文書中の語句の難易度レベルを用いて難易度を判定する手法が提案されている．

後者は係り受けの複雑さ等によって決まり，構造が複雑であると読み直しが必要になり理解が困難になる．更に後者は 3 つの要因からなると考えられ，文書を見た時の印象，係り受け構造と論理構成，記述内容に分類される．この中でも文書の記述内容は，ユーザの事前知識量に左右されるためその定量評価は難しい．

しかし，ある事柄について中心的に述べる専門書，ビジネス書のような文書では，背景を含め丁寧に単語間の関係を説明しているため，読み手の周辺知識が蓄えられることで記述内容全体の関連性を理解できる場合がある．一方で小説やブログといった様々な記述内容を含んだ文書は，各々の話題背景のつながりが薄いため全体を一つの一貫性でまとめることは難しい．Kritsada ら[5]は専門性が高く内容に一貫性のあるブログを面白いブログと定義し，トピック確率の特性を用いてこれを分類する手法を提案している．

本研究ではこの文書全体の記述内容の関係性を文書のまとめ方と考える．まとめ方は文書情報として語と語のつながりのパターンを表す共起と関連性があり，文書全体のまとめ方を測るには，文書全体の共起をネットワークとしてまとめた共起語グラフが有効であると考えられ，共起語グラフという文書の表層的な特徴からネットワーク特徴量を抽出することで，書籍程度の文章量の文書をまとめ方という新しい指標により評価する手法を提案する．文学および科学という異なる分野の文書分類に提案手法を適用し，文書の評価・分類に有効であることを実験的に確認した．

2. 共起と共起語グラフ

共起とは異なる語が同じ文中に現れることであり，語と語の間に何らかの意味的なつながりがあれば，共起が起こる確率は語によって偏る．特定の頻出語と選択的に多く共起するような偏りは，文書の著者が意味的なつながりを考慮して文書を書き進める上で生まれたものであり，分布が偏っている語は文書中において何らかの意味を担っている語であると考えられる．従って，文書の共起を調べることは文書の意味的な構造，即ち文書のまとめ方を把握することに繋がる．

語のネットワークマップとは，文中で用いられた語をノードとしたものである．特に，共起の関係を用いて結んだグラフを共起語グラフと呼ぶ．共起パターンに前後の順序の関係がある場合は有向グラフ，そうでない場合は無向グラフを用いる．本研究では共起について順序は考慮していないため共起語グラフは無向グラフとなる．ノードには名詞，動詞，形容詞等が用いられており，共起の測り方としては Jaccard 係数と Simpson 係数が一般的である．本研究では共起語グラフのノードとして頻出名詞上位 100 語を，共起尺度として Simpson 係数を採用している．

3. 複雑ネットワークとその指標

3.1　複雑ネットワーク

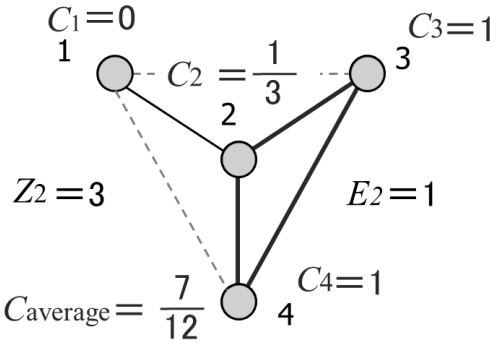
複雑ネットワークは Watts and Strogatz[6]のスモールワールド・ネットワークに関する研究以降，研究が盛んになった分野である．複雑ネットワークの研究では，その対象は格子のように解析しやすい規則的なネットワークではなく，規則的に表現することのできない現実の大規模なネットワークを扱っている．そこでは，個々の構成要素よりもネットワーク全体としての特徴が注目される．主に扱われる特徴量として，平均パス長，次数分布，クラスタ性等が挙げられる．平均パス長とは，文字通りグラフに含まれる全てのノード間

の最短距離の平均値である．平均パス長が小さいほど，グラフ内のあるノードから他のあるノードへ短い距離で到達できる．言い換えると小さい世界になっているといえる．また，ノードから出ているエッジの数を次数といい，次数分布はグラフに含まれるノードの次数がどのように分布しているかを示すものである．クラスタ性とはある注目する一つのノードに対し，それに隣接したノードと生成されるネットワークにおいてどの程度つながりがあるかを示す指標である．平均パス長やクラスタ性に関しては，本研究では重要な指標であり，次節で詳しく説明する．

3.2 複雑性の指標

クラスタ係数 C とはネットワークがどの程度凝縮しているかを示す指標であり，ネットワークに含まれる三角形構造の割合である．クラスタ性はクラスタ係数によって示され，あるノードのクラスタ係数はそのノードと隣接するノードからなるネットワークの密度を表し 0 から 1 までの値を取る．全てのノードの平均である平均クラスタ係数を求めることによって，ネットワーク全体のクラスタ性を計算することができる．

例として図 1 のように 4 つのノードからなるネットワークのノード 2 のクラスタ係数を考える．ノード 2 を含む三角形は隣接するノード 1,3,4 のうち 2 つを含む 3 通りが考えられるが，実際に存在するのはノード 3,4 を含む 1 通りだけなので，クラスタ係数は 1/3 となる．



$$C_i = \frac{E_i}{Z_i} = \text{ノード}i\text{のクラスタ係数}$$

E_i ＝ノードiを含む三角形の数

$$Z_i = {}_kC_2 = \frac{k(k-1)}{2}$$

＝ノードiを含むように考えられる三角形の数

k ＝ノードiから出るエッジの数

図 1　ネットワークのクラスタ係数の計算

本研究では，クラスタ係数以外の複雑ネットワークの指標として，下記の指標を使用する．

- 平均パス長 ： $\boldsymbol{L}$
- スモールワールド性: $\boldsymbol{M}$（＝ $\boldsymbol{C/L}$）
- 同類選択性 ： $\boldsymbol{A}$
- 次数エントロピー ： $\boldsymbol{H}$

平均パス長は全ノード間の最短距離の平均値であり，スモールワールド性はクラスタ係数と平均パス長の比である．同類選択性は接続ノード間の次数の相関係数であり，-1 から 1 までの値を取る．次数エントロピーは次数のエントロピーであり，次数分布の均一性を表す．

4. 複雑性指標関数

本研究では，共起語グラフと複雑性指標の計算によってテキストを評価する．まず文書に対し文単位での名詞の共起を調べ，ストップワードを除いた tf 値上位 100 位までの名詞を抽出し，それらをノードとする共起語グラフを作成する．ここでノード同士を結ぶエッジの重みは代表的な共起指標である Simpson 係数を用いる．さらにこの共起語グラフのエッジに閾値 θ を設定し，重みが閾値未満のエッジを切断すると，θ の変化に付随してネットワークの構造が変化する．このとき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 C も構造と共に変化するので，クラスタ性の指標である平均クラスタ係数は閾値の関数とみなせる．閾値は各文書に対して閾値 0 から平均クラスタ係数 C が 0 になる閾値を百等分割して設定し，同文書では他の指標もこの百段階の閾値を利用する．同様にして複雑性指標である平均パス長 L，スモールワールド性 M，同類選択性 A，次数エントロピー H も閾値の関数とみなせる．本研究では，これらを複雑性指標関数と呼び，これらの関数を文書の特徴量ベクトルとしてテキストの評価・分類を行う．例えば 5 つ全ての複雑性指標を用いると，5 つの複雑性指標それぞれに対して 100 の閾値が存在するため，1 文書につき 500 次元の特徴量ベクトルが得られ，その中から有効な指標を選択し文書の特徴量とする．

実際に書籍の複雑性指標関数を作成し，その傾向を調べた．例として，図 2～6 に文学書と科学書の各指標における複雑性指標関数のいくつかを示す．このように実際の文書の複雑性指標関数は，閾値を共起語グラフの各部分がツリー構造や三角形構造へと変化するため，単調減少ではなく値が増減する．

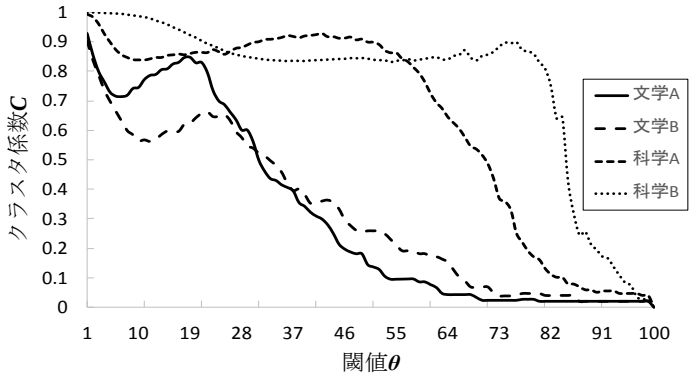


図 2　クラスタ係数 C の複雑性指標関数

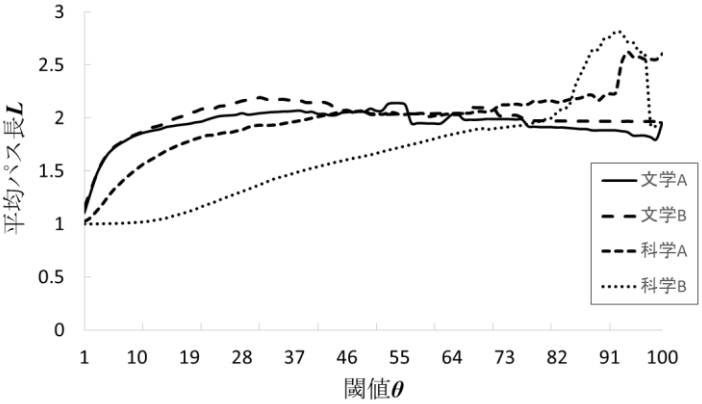


図 3　平均パス長 $\boldsymbol{L}$ の複雑性指標関数

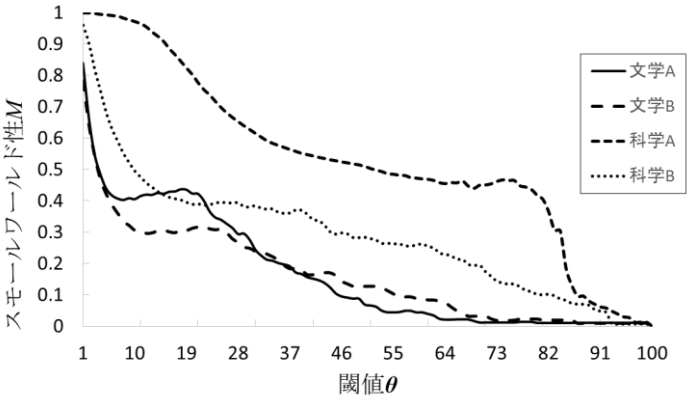


図 4　スモールワールド性 $\boldsymbol{M}$ の複雑性指標関数

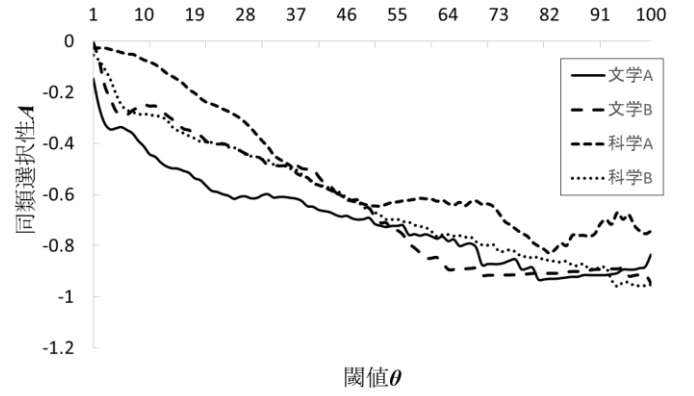


図 5　同類選択性 $\boldsymbol{A}$ の複雑性指標関数

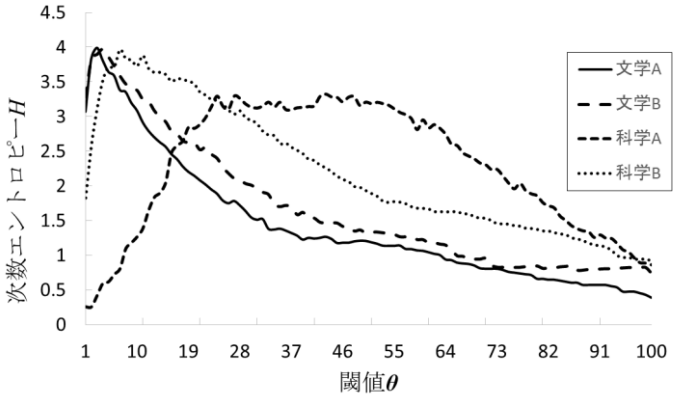


図 6　次数エントロピーHの複雑性指標関数

5.　評価実験

5.1　分野別文書分類実験

提案手法の有効性を確認するため，文書分類実験を行った．データセットには青空文庫から得た文学書と科学書を各 20 文書の合計 40 文書を用いた．文学書は様々な話題について述べており，科学書は特定の事柄について述べているものが多いため，前者より後者がよりまとまりが高いと評価されると考えられる．2 種類の文書分類では，分類器に SVM を用いて 10 分割交差検定を 5 回行い，その平均を結果とした．表 1 に選択した複雑性指標と分類正解率を示す．図 2 のクラスタ係数 C では，71%の精度であったが，図 5 の同類選択性を指標にした場合には約 90%の精度が得られ，有用な指標であることを確認できた．

5.2　オリジナルとランダム化文書の分類実験

提案手法がまとめ方によって文書を分類できているか確認するため，オリジナルの文書から意図的に生成したでたらめな文書との分類実験を行った．ランダム化文書の生成には，オリジナル文書に対し，異なる文間での名詞の入換を繰り返し行うことによって，共起のパターンを意図的にランダム化した．また，入換回数を 100 回，1000 回，10000 回とした．データセットと分類方法については分野別文書分類と同様のもので行った．表 2 に選択した複雑性指標と各入換回数での分類正解率を示す．名詞入れ替え回数が少ない程，分類が難しいタスクであることを確認できた．入れ替え回数が少なくオリジナルに近いほど，文書としてはまとまりがあると考えられるので，提案手法はまとめ方に沿った分類手法として妥当であると考えられる．

5.3 人手ラベル付き文書のクラスタリング

人手で文書がまとまっているか，まとまっていないかを被験者 1 名が判断したラベル付き文書を用い分類実験を行った．データセットは分野別文書分類と同様のもので行った．分類器については，主成分分析による特徴量の次元圧縮を行い，全ての指標で累積寄与率が 95%以上となる第十主成分までを使用し，SVM によって 10 分割交差検定を行った. 表 3 に選択した複雑性指標と分類正解率を示す．全ての指標を用いた場合に 90%という高い精度で，人手ラベル付き文書に対してまとめ方を評価することができた．

表 1　選択した複雑性指標と分類正解率

選択指標	正解率（%）
クラスタ係数 *C*	71
平均パス長 *L*	69
スモールワールド性 *M*	73
同類選択性 *A*	**90**
次数エントロピー*H*	81
5 種全て *C*～*H*	74

表 2　選択指標と各入換回数での分類正解率

選択指標	100 回入換	1000 回入換	10000 回入換
C	69	72	75
L	65	66	68
M	68	69	70
A	63	66	70
H	65	71	73
C～*H*	75	78	80

表 3　選択指標と分類正解率

選択指標	正解率（%）
クラスタ係数 *C*	58
平均パス長 *L*	50
スモールワールド性 *M*	53
同類選択性 *A*	78
次数エントロピー*H*	85
5 種全て *C*～*H*	**90**

6. 考察

分野別文書分類実験では，科学書の中に様々な科学分野の小さな話題を集めた書籍があり，そういった分野内で特殊な書籍は正確に分類されない傾向があった．分野毎にどの程度話題が定まっているかなど，文学，科学以外の分野の書籍のまとめ方についても調べる必要があると考えられる．

本研究では文書の共起をランダム化させる際，どの文も等確率で名詞の入換が行われるようにしたが，距離が近い程交換されやすく，距離が離れる程交換されにくいといったように，入換規則を確率的に設定することも今後検討していきたい．

人手ラベル付き文書のクラスタリングについては，分野別分類で正確に分類することができなかったまとめ方が分野内で特殊な書籍も，まとめ方という意図にそって正確にクラスタリングされていた．しかし，ラベル付の主観依存が大きいと考えられるため，より詳細な判定項目を設定する必要があると考えられる．

7. まとめ

文書の共起語グラフの複雑性指標を求め，閾値との関係を文書特徴量ベクトルとすることで，テキスト評価・分類を試みた．実験では書籍の分野別分類実験とまとめ方によるクラスタリング実験を行い高精度な分類を達成できた．今後は複雑性指標関数の特徴量の改善を行うとともに，他分野・他言語文書での有効性を確認したい．

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] 山中翔太，山崎高弘，常盤欣一朗，長谷川哲子, "構文解析を用いた日本語論文の読みやすさ判定法"，情報科学技術フォーラム講演論文集 8(2), pp.275-276, 2009.
[4] 近藤洋介，松吉俊，佐藤理史，"教科書コーパスを用いた日本語テキストの難易度推定", 言語処理学会第14回年次大会論文集，pp.1113‐1116,2008.
[5] Kritsada Sriphaew, Hiroya Takamura, Manabu Okumura ,"Cool Blog Identification using Topic-based Models", Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence-2008, pp.402-406, 2008.
[6] Watts, D. and Strogatz, S., "Collective dynamics of small-world networks", Nature, Vol. 393, pp. 440-442, 1998.