6T-3

複雑ネットワークからのキーワード抽出

三澤 英樹 †　　大沢 英一 ‡
† 公立はこだて未来大学大学院 システム情報科学研究科
‡ 公立はこだて未来大学 システム情報科学部複雑系知能学科

1 はじめに

インターネットの普及，文書の電子化が進み，インターネットに存在する文書の数は膨大になっている．それにより，自分の望んだ文書を探すことが困難になっている．そこで，文書の選択を容易にする手段としてキーワード抽出が考えられる．キーワード抽出とは，文書における重要単語をキーワードとして抽出し，キーワードを見るだけでその文書で述べられている内容が自分の望んだ内容かどうか，推測することを補助する技術である．

キーワード抽出手法として tf-idf[1] が広く用いられている．しかし，この手法は idf と呼ばれる文書頻度を用いて行われるため文書頻度に使用するコーパスにより結果が左右されてしまうという欠点がある．そこで，松尾らは語の共起からネットワークを作成し，Small World 構造に基づいたキーワード抽出を行うアルゴリズムを提案した [2].

本研究では語の共起から得られたネットワークの構造的特徴を調査し，ネットワークの構造的特徴を変化させた時のキーワードの抽出結果の変化についての調査を行う．

2 関連研究

松尾らは語の共起からネットワークを作成することで，Small World 構造に基づいたキーワード抽出手法を提案した [2]．また，松尾らはひとつのノードの Small World 構造に対する貢献度を，平均経路長 L の定義を非連結に拡張した extend path length を利用することで計算した．extend path length の定義を次に示す．

定義 2.1 *ノード i，ノード j に対する extended path length $d'(i, j)$ を次のように定義する．*

$$d'(i,j) = \begin{cases} d(i,j) & \text{if } (i,j) \textit{areconnected}. \\ W_{sum} & \textit{otherwise}. \end{cases}$$

Keyword extraction from complex network
†Hideki MISAWA　‡Ei-ichi OSAWA
†Graduate School of Systems Information Science, Future University Hakodate
‡Department of Systems Information Science, Future University Hakodate

そこで，本研究では語の共起からネットワークを作成する際に共起指標の閾値を利用し，ネットワークの非連結を回避することで，より正確な計算を行い抽出精度の向上を図る．

3 語の共起を利用した文書ネットワーク

本研究では，英語の文書から文書ネットワークを次のように作成した．まず，文書の単語に対してステミングを行い,「...ing」，三単元の「s」などを取り除き語幹の形を得る処理を行う．また,「a」,「the」などのあらかじめ決められた不要語 (ストップワード) を削除する処理を行う．次に，N-gram によりフレーズを抽出する．ここでは $N = 4$ とした．その後，規定回数 (f_0) 以上出現する語，またはフレーズをノードとして取り出す．ここでは $f_0 > 3$ とした．最後に，2つのノードに対応する語の，同一文中での共起が共起指標の閾値以上であればリンクを張る．共起指標については次項で述べる．

4 共起指標

共起指標とは語 A と語 B の関係の強さを表す指標である．広く用いられている指標としては Jaccard 係数が挙げられる．また，本研究ではその他にも Simpson 係数を使用している．文章中の語 A，語 B の出現回数をそれぞれ $|A|, |B|$，語 A と語 B が同時に出現している文の数を $|A \cap B|$，語 A，語 B の少なくとも一方を含む文の数を $|A \cup B|$ としたとき，語 A と語 B の関係の強さは次のように表される．

Jaccard 係数 $\dfrac{|A \cap B|}{|A \cup B|}$

Simpson 係数 $\dfrac{|A \cap B|}{\min(|A|, |B|)}$

これにより，ネットワーク全体から見たそれぞれのノード間の関係の強さが得られる．本研究では，Jaccard 係数と Simpson 係数に着目し，この値に閾値を設定することで，閾値以上のノード間にのみリンクを張り，ネットワークの構造的特徴を変化させた．

5　実験

5.1　ネットワークの構造的特徴の変化の調査

実験の手順として，まず，人工知能分野の論文から文書ネットワークを作成し，共起指標の閾値を変化させることで，ネットワークの構造的特徴の変化に関する調査を行った．

共起指標の閾値を変化させた場合の拡張された extend path length の変化を図 1 に示す．図の横軸は共起指標の閾値，縦軸が extend path length となっている．

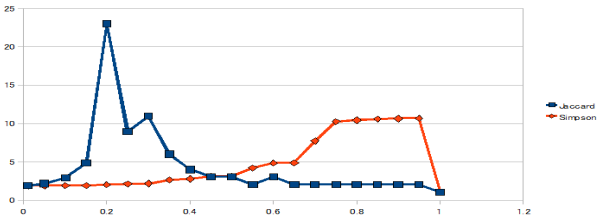

図 1: 拡張された平均経路長の変化

共起指標の閾値を高くしていく際に，extend path length が急激に増加するのは，共起指標の閾値を高くすることでリンクが急激に減少してしまい，ネットワークのほとんどのノードが繋がっておらず，ネットワークの崩壊が起きたためである．

次に共起指標の閾値を変化させた場合のクラスタ係数の変化を図 2 に示す．図の横軸は共起指標の閾値，縦軸がクラスタ係数となっている．

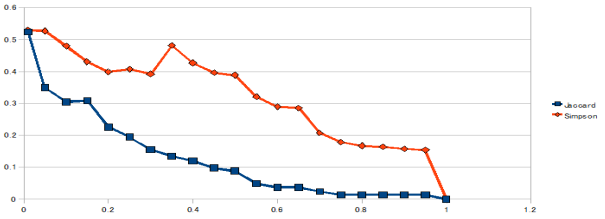

図 2: クラスタ係数の変化

共起指標の閾値を変化させた場合，クラスタ係数が一時的に上昇しているのは，共起指標の閾値により共起の強いリンクでネットワークが構成されることで，密なネットワークが生成されたためであると考えられる．

また，これらの傾向は他の論文で実験した場合でも見られたため，文書ネットワークの共起指標を変化させたときの特徴であると考えられる．

表 1: precision の比較

	先行研究	提案 A	提案 B	提案 C
precision	0.27	0.27	0.33	0.5

そこで，本研究ではこのクラスタ係数と平均経路長に着目し，クラスタ係数が極大を取る値，平均経路長が急激に増加する直前のネットワークでキーワード抽出を行い，抽出精度の比較を行う．

5.2　抽出精度の比較実験

実験の評価は，松尾らの先行研究，提案手法のそれぞれよって抽出されたキーワードの上位 15 個を出力し，各種法から得られたキーワードをひとつにまとめシャッフルする．その後，論文中に示されたキーワードを利用し，キーワードであると考えられる語を数え，各手法による出力語中でキーワードとされた語の割合を precision として評価した．

この際に Jaccard 係数を利用したクラスタ係数が極大を取る値を閾値としたネットワークを提案 A，Simpson 係数を利用した際にクラスタ係数が極大を取る値を閾値としたネットワークを提案 B，Simpson を利用した際に平均経路長が急激に増加する直前の値を閾値としたネットワークを提案 C とした．

先行研究と提案手法での本実験の結果を表 1 に示す．

この結果から，先行研究よりも全てのネットワークでの precision が高くなっていることがわかる．また，提案 C がさらに precision が高くなっていることがわかる．

6　まとめ

本研究ではネットワークの構造的変化によるキーワード抽出結果の変化を調査した．結果として，共起指標に閾値を利用することで抽出精度が向上した．また，今回の実験によりネットワークの構造的特徴はキーワード抽出の抽出結果を変化させるのに重要な役割を持っていることがわかった．

今後の課題として，tf-idf による抽出結果の比較を行い，評価を行う必要があると考えられる．

参考文献

[1] G. Salton and C.S. Yang. On the specification of term values in automatic indexing. *Journal of documentation*, Vol. 29, No. 4, pp. 351–372, 1993.

[2] 松尾豊, 大澤幸生, 石塚満. Small world 構造に基づく文書からのキーワード抽出. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1825–1833, 2002-06-15.