共起ネットワークにおける最短経路を用いた英語文章のキーワード抽出

本田泰, 打越秀昭, 片川智弘

室蘭工業大学 情報工学科

Tv

本研究では, 英語文章から得られるネットワーク構造を用いてキーワード抽出を行った. 既存の研究では, 文書中の単語の共起関係により構成したネットワークにおいてある語を取り除くことによって平均経路長が増加するものをキーワードとした. 本研究では, キーワードが用いられている文で使われる語を高く評価することが可能になるように新たなキーワード抽出方法を導入し, 既存の研究より精度向上がみられた.

Keyword Extraction from English Documents using Minimum path length in Co-occurence Networks

Yasushi Honda, Hideaki Uchikoshi, Tomohiro Katagawa

Department of Computer Science and Systems Engineering, Muroran Institute of Technology

Abstract

We study a keyword extraction algorithm which utilizes co-occurence networks obtained from English documents. Matsuo et al. determined keywords by using the minimum path length in networks obtained by co-occurence in a document. In the present study, we concider a cluster of words in the word network and obtained high-accuracy in keyword extractions.

1 はじめに

近年, インターネットの普及に伴って新聞, 書籍, 技術論文などあらゆるものがデジタル化され保存されている. 我々が, これらの膨大な情報の中から必要な情報を得るために情報検索技術は, 必要不可欠である. 情報検索技術において, 文書に自動で索引付けを行う作業やキーワード抽出は, 重要な工程である.

最新の技術情報やニュースで使われる単語は, 日々新しく生まれている. キーワード抽出方法としてよく知られている TFIDF は, キーワード抽出を行う対象文書での出現頻度 TF(Term Frequency) と文書集合での対象文書の特定性 IDF(Inverse Document Frequency) を用いるため, 文書と共に新しい語も急激に増加する状況の今日, 頻度情報や文書集合と言った事前知識を用いずに新語もキーワードとして抽出が可能な方法が望まれている. 文章で用いられた単語の頻度情報や文書集合などの事前知識を用いずに, 対象文書のみを用いて単語をキーワードとして抽出するシステム, アルゴリズムが提案されている [1, 2].

World Wide Web, 俳優の共演関係, 高速道路網, 蜘蛛の巣など人工物から自然界に存在するものまでそれらが要素と要素間の接続で成り立っているとすると様々なものがネットワーク構造を持つ. 言語において単語は, 文中で無秩序に並ぶのではなく, 文法や表現によって相互に関係し, 影響し合う. このように言語も単語と単語間の関係から成り立っていると捉えるとネットワークであると言える.

本研究では, 文章構造をある種のネットワークとして捉えることで新たなキーワード抽出を行った松尾ら [2] が提案したキーワード抽出のための尺度 $C_B(v)$

の定義を改良し, 精度向上を行った.

2　文書中の語の共起ネットワーク

共起ネットワークはひとつの文書から修飾-被修飾語, 熟語, 定型化された表現などと言った相互関連性の高い 2 単語を一文中に出現した共起頻度をもとに抽出する. また, 得られた共起ネットワークは, スモールワールド構造を備えていることが示されている [2, 3]. 英語の文書から共起ネットワークを構築する手順について以下に示す.

1. 前処理:文書の特徴を表さない不要語を除去, 動詞の活用形を原形, 名詞の複数形を単数形に戻す.
2. ノードの生成:文書で規定回数 (f_0) [1] 以上出現する単語をノードとして抽出する.
3. リンクの生成:二つのノードが同一文中で多く用いられていればリンクを張る. 共起は, Jaccard 係数[2]を用いて, この上位 (k_0) の 2 ノード間にリンクを張る.

図 1 に上記の手順で得られる共起ネットワークの一例を示す.

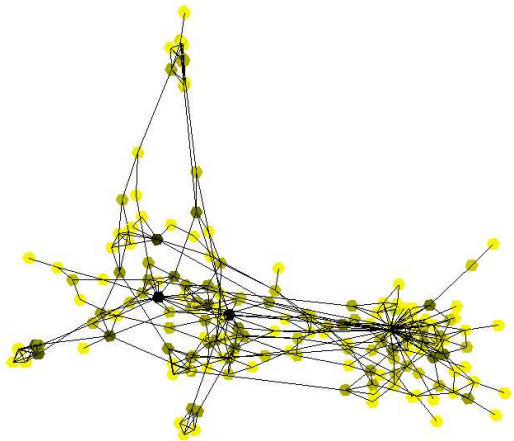

図 1: $f_0 = 5, k_0 = 4$ とした場合の実験 (5 節) で用いた文献の一論文から得られた共起ネットワーク, $C_B(v)$ をノードの色階調で表現した. 色が濃いノードほど $C_B(v)$ の値が高い.

[1]今回は, 出現頻度の閾値を f_0=5 とした. 同じ単語でも文書の長さによって出現頻度が異なることから, Zipf の法則 [4] などを用いて閾値を設定する必要がある.

[2]$T(w1, w2)$ = "単語 $w1, w2$ が共に出現する文の数", $S(w)$ = "単語 w が単独で出現する文の数" とすると単語 $w1, w2$ に対する Jaccard 係数は, $Jaccard(w1, w2) = \frac{T(w1,w2)}{S(w1)+S(w2)-T(w1,w2)}$ である. w1,w2 を用いる文がどの程度一致しているかを示す尺度となる. 今回は, $k_0 = 4$ とした.

2.1　媒介性 $C_B(v)$

一文中に出現する 2 単語間のリンクは, 文中での出現順を考慮せず, 向きはないとし, 共起ネットワークは, 無向グラフとする. 無向グラフ G の頂点集合を V とする. ノード $v(\in V)$ 以外のすべてのノードの組についての平均頂点間距離 L_v とノード v を削除したネットワーク G_v における平均頂点間距離 L_{G_v} との差として $C_B(v)$ を定義する. ただし, n は, G のノード数, $V' = \{x | x \in V, x \neq v\}$, $d(p, q, G)$ は, G における p, q 間の最短距離である. p, q 間に経路がない場合は, W_{sum} とする. ノード v の削除によって生まれるサブグラフ集合を G' とすると, $W_{sum} = \sum_{g \in G'} \mathrm{max_{p,q \in g}} d(p, q, g)$ とし, W_{sum} は, 有限の経路長とする.

$$
\begin{aligned}
C_B(v) &= L_{G_v} - L_v \\
&= \frac{1}{{}_{n-1}\mathrm{C}_2} \sum_{p,q \in V'} (d(p, q, G_v) - d(p, q, G))
\end{aligned}
$$

$C_B(v)$ はネットワークの全ノード間から算出される平均最短経路長に対して v がどの程度貢献しているかを計る指標である. 共起ネットワークにおいて, $C_B(v)$ 値が大きく, 媒介性の高いノードは, 互いに関連性の薄いある二つの概念を結びつけるといった重要な語とみなすことができる.

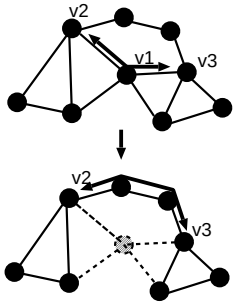

図 2: v1 の除去で v2,v3 ノード間の最短経路が延長する. v1 は, このネットワークにおいて他の 2 ノードを結びつける重要なノードと言える.

3　$C_B(v)$ を用いたキーワード抽出

文書を特徴付ける語や, 特定の概念を表す重要な語がその文書のキーワードと仮定すると, これらの語は, 他の互いに関係の薄い語を結び付ける語であると考えられる. 共起ネットワークにおいて, このような性質を計る指標として前節で定義した $C_B(v)$ を

用いる. 共起ネットワークを構成するすべてのノードにおいて $C_B(v)$ を算出し, $C_B(v)$ の値が高く, 共起ネットワークにおいて他の 2 ノードの媒介を行う語をキーワードと推定する.

定義より最短経路上に存在するノードが高い $C_B(v)$ を持ち, 文書のキーワードとして推定されることとなる. しかし, 共起ネットワークにおいて, 唯一の最短経路上にないノードの $C_B(v)$ の値は, 極端に低い値となる. 例えば, 図 3 のように複数の最短経路上にある場合, $v1,2$ ともに Group1,2 を結びつけ, ネットワークにおいて重要な位置に存在するノードと言えるが $v1, v2$ を個々に削除しても 2 ノード間の最短経路長に変化はなく, $C_B(v1) = C_B(v2) = 0$ となる. このような場合が多数存在するとキーワードとして適切な語を抽出できないこととなる. また, 図 4 のように最短経路上に存在していないノード $v1$ は, $C_B(v1) = 0$ となる. このように, 既存の研究による定義では, キーワードとするべき語が正しく得られない場合が存在すると考えられる. ここでは, 図 3 のような状況を想定したが, 最短経路が複数存在することが共起ネットワークにおいては, 多数見られないことから, 問題点とはならない.

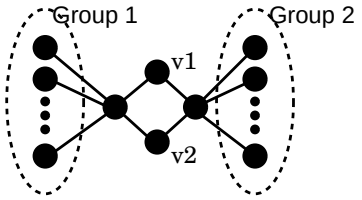

図 3: Group1,2 を結びつける複数の最短経路がある場合

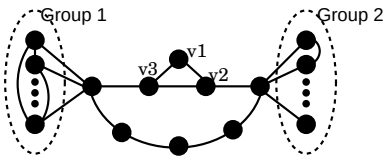

図 4: 最短経路上ではない位置にキーワードとなるべき単語がある場合

4 クラスターを用いた $C_B(v)$

図 4 のノード $v1$ がノード $v1, v2, v3$ によって三角形のクラスタ C_l を形成することで $C_B(v1) = 0$ となり, キーワードとして抽出できないことを回避するため $C_B(v)$ の定義を以下のように変更する. 無向グラフ G の頂点集合を V とする. 頂点集合 V' の要素をノード $v(\in V)$ 及び v と隣接するノードでかつ三角形のクラスタ C_l を形成するノードとする. $v'(\in V')$ 以外のすべてのノードの組についての平均頂点間距離 L'_v とノード v' を削除したネットワーク G''_v における平均頂点間距離 $L'_{G'_v}$ との差として $C'_B(v)$ を定義する. n は, G のノード数, $V'' = V - C_l$,n'は, V' の要素数 とする.

$$
\begin{aligned}
C'_B(v) &= L'_{G'_v} - L'_v \\
&= \frac{1}{{}_{n-n'}\mathrm{C}_2} \sum_{p,q \in V''} (d(p,q,G''_v) - d(p,q,G))
\end{aligned}
$$

このように定義を変更することで三角形のループ構造を形成するノードもキーワードとしての評価を高く見積もることができる. 従来の $C_B(v)$ で高い値を持つノードと隣接していながら低い値となっていたノードは, $C'_B(v)$ では, 高い値となる. このことにより, キーワードと推定していた単語と同一の文で用いられていた語をキーワードとして高く評価することになる.

5 評価方法

共起ネットワークからキーワードを抽出するための尺度である媒介性 $C_B(v)$,$C'_B(v)$ を英語の論文を用いたキーワード抽出実験で評価した. キーワード抽出法としての評価は, 論文の筆者が明記したキーワードと論文の本文のみから推定するキーワードを比較することで行った. 実験は, 論文データベース ScienceDirect[3] より"complex","network" の二つの単語で検索した最近の論文 40 篇を用いた. 精度は, 適合率及び再現率で比較を行う. 本手法は, 各論文ごとにキーワードの推定を行うが, 評価は, 40 篇での平均的な適合率, 再現率とし, 以下のように定義する.

$$
\begin{aligned}
\text{適合率} &= \frac{\sum_{d \in D} C(d)}{\sum_{d \in D} E(d)} \\
\text{再現率} &= \frac{\sum_{d \in D} C(d)}{\sum_{d \in D} K(d)}
\end{aligned}
$$

ただし, C(d) は, 文書 d($\in D$) での推定キーワードのうち実際にキーワードである単語数, E(d) は, 文書 d($\in D$) での推定キーワード数, K(d) は, 文書 d($\in D$) での実際のキーワード数, D は, 40 篇の文書集合と

[3] http::://www.sciencedirect.com

する．適合率は，キーワードとして誤って抽出した語 (ノイズ) の比率を表し，再現率は，キーワードとして抽出できなかった語 (漏れ) の比率を表す．

6　結果と考察

実験では，事前知識を用いない手法と対峙して，文書集合を事前知識としたキーワード抽出法 TFIDF との比較も行った．結果を図 5,6 に示す．推定キーワード数とは，各論文からキーワードとして抽出する語の数であり，キーワードと高く推定される語から順に追加した．実験の結果，TFIDF には及ばないが従来の定義 $C_B(v)$ より今回変更した $C'_B(v)$ を用いたキーワード抽出法の精度が向上していることが確認できる．精度が向上した要因の一つとして，実験で用いた論文において同一文中で複数のキーワードが用いられている傾向があったと考えられる．

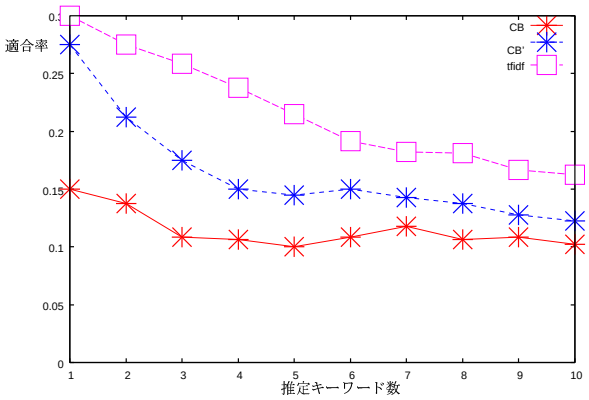

図 5: 論文 40 篇での平均的な適合率

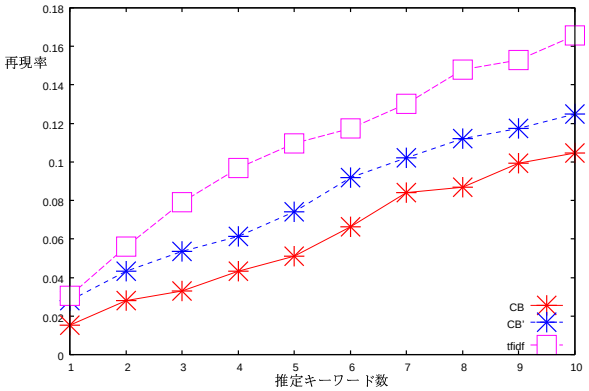

図 6: 論文 40 篇での平均的な再現率

7　まとめ

本研究では，松尾ら [2] によって提案された文章構造を利用したキーワード抽出の精度向上を目的とした．共起ネットワークにおけるクラスターを用いて新たなキーワード抽出法を開発し，実験によってその有効性を確認した．また，従来の方法では，共起ネットワークのノード数を n とすると，ノード数 $n-1$ のネットワークの最短経路を全探索しなければならないが，本研究における方法を用いると，ノード数 $n-n'(n'$ は，削除対象のノード数) のネットワークとなるため計算量も削減される．しかし，従来の TFIDF の計算量が $O(n)$ であるのに対し，グラフの最短経路を探索する $C_B(v)$ は，$O(n^3)$ である．キーワードの抽出を行う文書の要約など前処理による計算量の削減が今後の課題である．

参考文献

[1] 梅村恭司” 未踏テキスト情報中のキーワードの抽出システム” 平成 12 年度 IPA 未踏ソフトウェア創造事業

[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 Vol43.No.6 (2002)

[3] Y.Matsuo,Y.Ohsawa,M.Ishizuka：A Document as a Small World, Proceedings the 5th World, Multi-Conference on Systemics, Cybenetics and Infomatics(SCI2001),Vol.8,pp.410-414(2001).

[4] 徳永健伸” 言語と計算-5 情報検索と言語処理” 東京大学出版会 (1999)