Vol.2014-NL-217 No.13
2014/7/4

話題追跡のための共起ネットワークによる語の重要度の検討

増田真也[†1]　石川孝[†2]

オンラインニュースなど Web 上で発信される情報の中には互いに時系列的関連を持ち，情報ストリームを形成するものが少なくない．話題追跡タスクでは，ある特定の話題に関する少数の記事が与えられた場合に，その話題に関する記事を情報ストリームから検出する．従来手法では，少数の記事から求めた語の重要度を補正するために，大量のコーパスを利用する手法が用いられている．一方で，テキスト分類の分野では，文書中の語の共起ネットワークから語の重要度を求める手法が提案されている．そこで本研究は，少数の記事だけを用いる話題追跡手法の提案を目的として，共起ネットワークによる語の重要度の検討を行う．

1. はじめに

オンラインニュースなど Web 上で発信される情報の中には互いに時系列的関連を持ち，情報ストリームを形成するものが少なくない[1]．これらの情報ストリームから，新規話題の発生を検出したり，興味ある話題の関連記事を収集したりすることができれば，Web 情報活用において非常に強力な武器になることが期待できる．

情報ストリームを扱う研究のひとつに，Topic Detection and Tracking（TDT）プロジェクトがある[2]．TDT は，オンラインニュースなどのデータストリームから話題構造を自動的に得るための技術確立を目的としている．TDT の代表的なタスクのひとつである話題追跡タスクでは，ある特定の話題に関する少数(1〜4)の記事が与えられた場合に，その話題に関する記事を情報ストリームから検出する．従来手法では，少数の記事から求めた語の重要度を補正するために，大量のコーパスを利用する手法が用いられている[3]．また，TDT の関連研究として，著者らは Twitter トレンドの話題追跡に関する研究を行ってきた[4]．

一方で，関連する文書分類の分野では，文書中の語の共起ネットワークから語の重要度を求める手法が提案されている[5][6][7]．松尾ら[5]は，論文から得た語の共起グラフが Small World 構造を備えていることを示し，Small World 構造に基づいたキーワード抽出法を提案した．また，小熊ら[6]は，[5]のキーワード抽出手法を文書クラスタリングに応用し，文書中の語の頻度に基づく手法よりも人手による分類に近い分類を再現するのに有効であることを示した．Blanco ら[7]は，テキストを共起グラフで表現しそこから得られるネットワーク特徴を情報検索へ適用した．

本論文は，少数の記事だけを用いる話題追跡手法の提案を目的として，共起ネットワークによる語の重要度の検討を行う．本論文のアプローチは，まず話題追跡タスクを情報検索タスクとして定式化する．次に語の共起ネットワークから求めた語の重要度を，情報検索において一般的なベクトル空間モデルへ適用する．そして，従来手法との比較によって，共起ネットワークによる語の重要度の有効性を評価する．

本論文の構成を次に示す．2 章では話題追跡タスクの定式化を示す．3 章では語の共起ネットワークの定義と作成手順について述べる．4 章で本論文の提案手法である共起ネットワークから語の重要度を求める手法について述べる．5 章では提案手法が話題追跡タスクにとって有効であるかを示すために，従来手法との比較実験について述べる．

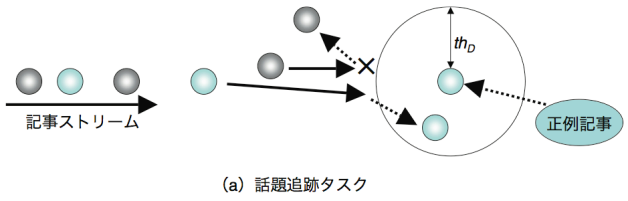
2. 話題追跡タスク

関連研究の手法を適用するため，話題追跡タスクを情報検索システムと同様のアプローチがとれるように，つぎの問題として定式化する．

Given: ある話題 T について述べられている記事 D が与えられる．

Seek: 記事ストリームから，類似度がしきい値 th_D を超える記事を関連記事として検出する．

この定式化での話題追跡タスクのイメージを図 1 に示す．



(a) 話題追跡タスク

図 1　話題追跡タスクのイメージ [1]より一部抜粋

3. 語の共起ネットワーク

本論文で扱う語の共起ネットワークは，ノードを記事中の語とし，リンクを同一文中の語の共起関係とする無向ネットワークと定義する．複数の文に同じ語が共起する場合はリンクが複数回張られ多重辺となるが，本論文では多重辺は除外する．語の共起ネットワークの作成手順は次の通りである．

†1 日本工業大学大学院
The Graduate School of Engineering, Nippon Institute of Technology
†2 日本工業大学
Nippon Institute of Technology

ⓒ2014 Information Processing Society of Japan

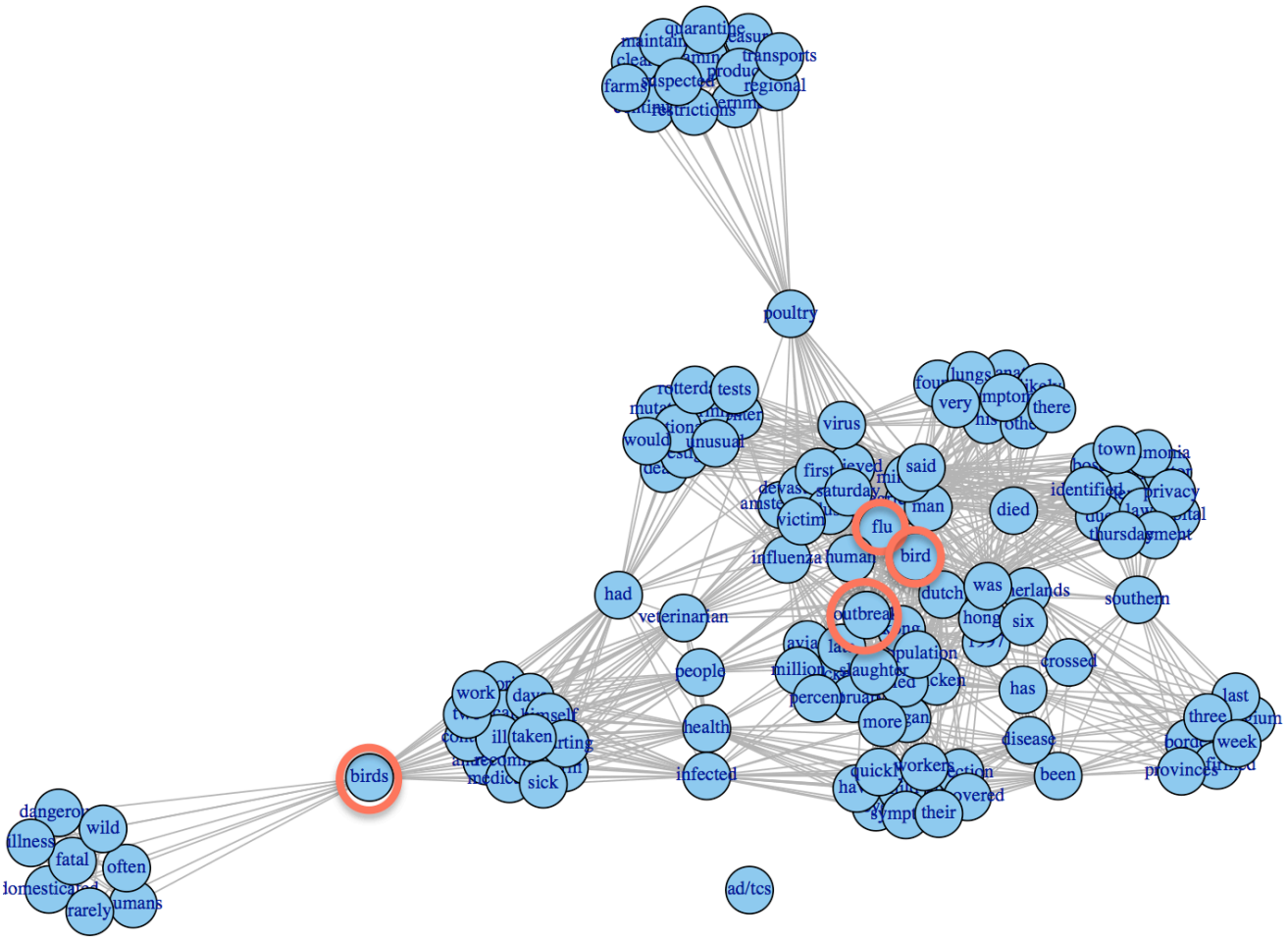


図 2　鳥インフルエンザに関する記事（図 3）から得た語の共起ネットワーク

```
1 AMSTERDAM , Netherlands ( AP ) _ A Dutch veterinarian is believed to be the first human
victim of an outbreak of bird flu devastating the poultry industry in the Netherlands ,
the Agriculture Ministry said Saturday . Bird flu , or avian influenza , has led to the
slaughter of more than 11 million chickens , around 10 percent of the Dutch chicken
population , since the outbreak began in late February . The illness is often fatal for
domesticated and wild birds , but is rarely dangerous for humans . The 57-year-old
doctor , who was not identified due to Dutch privacy laws , died Thursday of pneumonia in
a hospital in the southern town of Den Bosch , the Agriculture Ministry said in a
statement . ` ` It is very likely the man died of the bird flu because the virus was found
in his lungs and there is no other explanation for his symptoms , ' ' it said . The
veterinarian had not taken anti-viral medication recommended by health authorities for
people in contact with sick birds and became ill himself two days after starting work at
an infected farm . The Agriculture Ministry said it would investigate the unusual death
and that tests at the National Influenza Center in Rotterdam would determine if the virus
had mutated . More than 50 Dutch health workers have been infected with the disease since
the outbreak began , but their only symptom was a mild eye infection , and all recovered
quickly . The disease has been confirmed in three provinces in the Netherlands and last
week crossed the southern border to Belgium . The government continued to clear farms
suspected of contamination and maintained quarantine measures and regional restrictions on
transports of poultry products . In 1997 , six people in Hong Kong died after an outbreak
of bird flu crossed into the human population . ( ad/tcs )
```

図 3　鳥インフルエンザに関する記事の例 [11]より” APE20030419.1630.0392”の記事

ⓒ2014 Information Processing Society of Japan

(1) 前処理：語の unigram により記事中から語を抽出する．ただし，表 1 に示す語は除外する．
(2) ノードの生成：前処理により抽出されたすべての語をノードとする．
(3) リンクの生成：2 つのノードに対応する語が同一文中で共起していればリンクを張る．ただし，多重辺は除外する．

作成した語の共起ネットワークの例として，図 3 の鳥インフルエンザに関する記事に対する共起ネットワークを図 2 に示す．図 2 から，いくつかの語の集まりがクラスタを作っている様子がわかる．リンクは同一文中での共起関係を表しているため，各クラスタは記事中の各文に対応する．また，クラスタをつないでいるノードは，複数の文に出現する語であることを示している．

前処理において表 1 の語を除外する理由は，これらの語が共起ネットワークから語の重要度を求める際にノイズとなることが予備実験で確認されたためである．この予備実験の詳細は 5 章後半で述べる．また，表 1 は，英語において対象間の関係を表す，定冠詞，接続詞，前置詞，関係詞[8]を参考に作成した．

表 1 除外語リスト

the, and, but, or, either, neither, not, only, nor, so, for, that, whether, if, while, before, after, since, until, till, as, soon, once, because, such, unless, though, although, even, long, far, in, case, at, on, from, to, during, of, by, with, about, along, across, trough, around, front, behind, opposite, into, out, onto, over, under, above, below, between, among, according, who, whose, whom, which, what, where, when, why, how, whoever, whichever, whatever, whenever, wherever, however, than

4. 語の重要度

4.1 ベクトル空間モデル

情報検索システムでは，文書を文書中の語の重みを値とするベクトルで表現するベクトル空間モデルが一般的である[9]．いま，記事ストリーム中に記事 $D_1, D_2, \ldots, D_n$ があるとし，これらの記事集合全体を通して全部で m 個の語 $w_1, w_2, \ldots, w_m$ があるとする．このとき，記事 D_j は次のような文書ベクトル d_j^T で表現される．

$$d_j^T = \begin{bmatrix} d_{1j} & d_{2j} & \cdots & d_{mj} \end{bmatrix} \quad (1)$$

d_j^T はベクトル d_j の転置を表し，d_{ij} は語 w_i の記事 D_j における重みである．

記事中の語が，記事の内容を表すうえでどれだけ重要度を持っているかを測ることができれば，より高い精度で情報検索タスクを実行できると考えられる．この考えを元に行われるのが語の重み付けであり，一般的に TFIDF 手法が用いられる[9]．TFIDF 手法は，直感的には記事中に多数出現する語の重みは上げ，記事集合全体で多数の記事に出現する語の重みは下げる手法であると捉えることができる．

$$d_{ij} = tf_{ij} idf_i \quad (2)$$

$$tf_{ij} = f_{ij} \quad (3)$$

$$idf_i = \log \frac{n}{n_i} \quad (4)$$

ここで f_{ij} は記事 D_j における語 w_i の頻度である．また，n は記事ストリーム中の記事の総数であり，n_i は語 w_i を含む記事の数である．

記事 D_i と D_j の文書ベクトルを d_i, d_j とすると，一般的にそれらの類似度は，d_i と d_j のなす角度である次式の $\cos(d_i, d_j)$ で定義される．この類似度は，話題追跡タスクにおいて目標とする文書との近さを表す尺度となる．

$$\cos(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\| \|d_j\|} \quad (5)$$

4.2 ネットワーク特徴

人間関係などの社会的なネットワークでは，重要な位置を占めると大きな影響力を得られると考えられる．この視点から，文書から得た語のネットワークにおいて重要な位置を占める語は，もとの文書においても重要性が高いと考えるのが自然である[5].

本論文では，記事から得た語の共起ネットワークの特徴から語の重要語を抽出する手法を提案する．ネットワーク特徴には，あるノードの付近の情報だけから計算できる局所的なものと，ネットワーク全体の情報を用いて計算する大域的なものがある[10]．ネットワークをより深く理解するためには局所的・大域的の両者をあわせる必要がある．本論文では，共起ネットワークを用いた話題追跡手法への第一歩として，大域的なネットワーク特徴のひとつであるノードの中心性に着目する．

4.3 ノードの中心性

本論文では，ノードの中心性を測る指標のひとつである媒介中心性（Betweenness）に着目する．媒介中心性は一般的な中心性の指標のひとつであり，直感的には，ノード v_i の媒介中心性 b_i は v_i がネットワーク上の流れを橋渡ししたり制御したりする度合いである．媒介中心性 b_i は次の式(6)で定義される[10].

ⓒ2014 Information Processing Society of Japan

$$b_i \equiv \frac{\sum_{i_s=1:i_s\neq i}^{N}\sum_{i_t=1:i_t\neq i}^{i_s-1}\frac{g_i^{(i_s,i_t)}}{N_{i_s,i_t}}}{(N-1)(N-2)/2} \qquad (6)$$

ここで，$g_i^{(i_s,i_t)}$ は始点 v_{i_s} から終点 v_{i_t} へ行く最短経路の中で v_i を通るものの数，N_{i_s,i_t} は v_{i_s} から v_{i_t} へ行く最短経路の総数である．式(6)の分子は，ノード v_i が他の２点を結ぶ最短経路上にあると，v_i に得点が入ることを意味する．総和記号における $i_s, i_t \neq i$ は，v_{i_s} または v_{i_t} が v_i ならば v_i が最短経路上にあるのは当たり前なので，そのような場合を除外することを表す．分子の二重和は，i_s と i_t の組を i 以外の $N-1$ ノードから選ぶことを表す．このような i_s と i_t の選び方は $(N-1)(N-2)/2$ 通りある．式（6）の分母は規格化定数である．

4.4 データ例

3 章の図 2 の共起ネットワークから求めたノードの媒介中心性が上位の語を表 2 に示す．図 2 の共起ネットワークは鳥インフルエンザに関する記事から得たものであるが，媒介中心性の高い”birds”，”outbreak”，”bird”，”flu”などの語が，鳥インフルエンザに関する記事であることを表す重要な語であることが分かる．本論文の提案手法は，この媒介中心性の値を語の重みとして文書ベクトルを構築する．

表 2　図２における媒介中心性が上位の語

Term	Betweenness
poultry	0.175
birds	0.112
veterinarian	0.103
dutch	0.089
said	0.064
had	0.057
outbreak	0.056
was	0.053
people	0.048
health	0.046
infected	0.046
agriculture	0.042
ministry	0.042
bird	0.041
flu	0.041

5. 比較実験

5.1 目的

提案手法が話題追跡タスクに対して有効であるかを評価するために，従来手法である TFIDF 手法との比較実験を行う．

5.2 方法

比較実験で用いる記事は，TDT5 コーパス[11][12]を使用した．TDT5 コーパスは，TDT プロジェクトにおけるベンチマークデータであり記事の話題について人手によるアノテーションが施されている．TDT5 コーパスには English, Mandarin, Arabic と 3 言語からなる Web ニュース記事が含まれるが，本論文では English のみを扱う．その理由は，Mandarin, Arabic の記事は English に機械翻訳された状態であるため，機械翻訳の問題を避けるためである．TDT5 コーパスに含まれる English の記事の総数と，話題の総数を表 3 に示す．

表 3　TDT5 コーパスに含まれる English の記事数と話題数

Doc Count	278,109
Topic Count	126

話題追跡タスクでは，与えられた記事と同じ話題の記事（正例という）をストリームから検出することを目的とする．このためには同時に，与えられた記事と異なる話題の記事（負例という）を除外する必要がある．情報検索システムの文脈では，この検出と除外の関係は情報検索と同様に再現率と適合率のトレードオフがある．この観点から，実験では提案手法がより精度よく同じ話題の記事とそうでない記事を判別できるかを評価する必要がある．これを踏まえて比較実験は次の手順で行う．

1. TDT5 コーパスに含まれる話題のうち, ひとつの話題を選択する．
2. 選択した話題に属している記事をひとつ選択し，種とする．
3. 種に直近の同じ話題の記事を目標とする．
4. 記事ストリームのうち，種と目標の間の異なる話題の記事を負例集合とする．
5. 提案手法および TFIDF 手法でそれぞれの語の重要度を求め，各記事を文書ベクトル d で表現する．
6. 種と目標の cos 類似度，および種と負例集合の cos 類似度を算出し，類似度の分布を比較する．

5.3 結果

実験条件を表 4 に，選択した話題のアノテーションの一部を表 5 に示す．また，提案手法および TFIDF 手法によるそれぞれの結果を次に示し，比較を表 6 に示す．

ⓒ2014 Information Processing Society of Japan

(1)　提案手法

提案手法の cos 類似度の分布を図 4 に示す．種と目標の cos 類似度は 0.63 であった．負例集合の cos 類似度の平均 $\bar{x}$ は 0.068 であり，標準偏差 σ は 0.059 であった．

(2)　TFIDF 手法

TFIDF 手法の cos 類似度の分布を図 5 に示す．種と目標の cos 類似度は 0.72 であった．負例集合の cos 類似度の平均 $\bar{x}$ は 0.020 であり，標準偏差 σ は 0.018 であった．

5.4 考察

提案手法が話題追跡タスクに対して有効であるかを考察すると，図 4 および表 6 から，種と目標の cos 類似度が負例集合の cos 類似度の平均よりも約 9.5 σ の差で分かれていることがわかる．一方で，TFIDF 手法は約 38 σ の差で分かれている．この結果から TFIDF 手法の方が目標と負例記事をより精度よく判別できるといえる．しかし，提案手法も目標と負例記事を判別するには十分な性能であると考えることができる．したがって，提案手法は話題追跡タスクに対して有効であるといえ，さらに検討する価値があると考えられる．

実験した 2 つの手法で抽出された重要語を具体的に比較するため，提案手法および TDIDF 手法で求めた語の重要度を表 7 に示す．表 7 の語のうち，表 5 の記事のアノテーション中に存在する語は太字で強調した．表 5 と表 7 から媒介中心性が高い語は，記事のアノテーション中に出現する語であることが分かる．

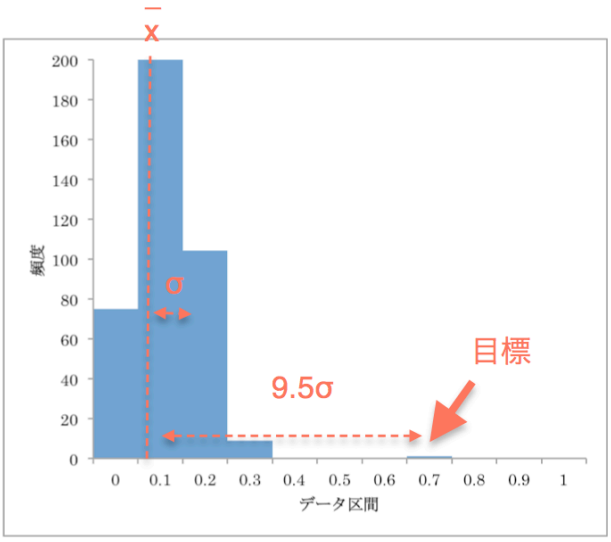


図 4　提案手法の cos 類似度の分布

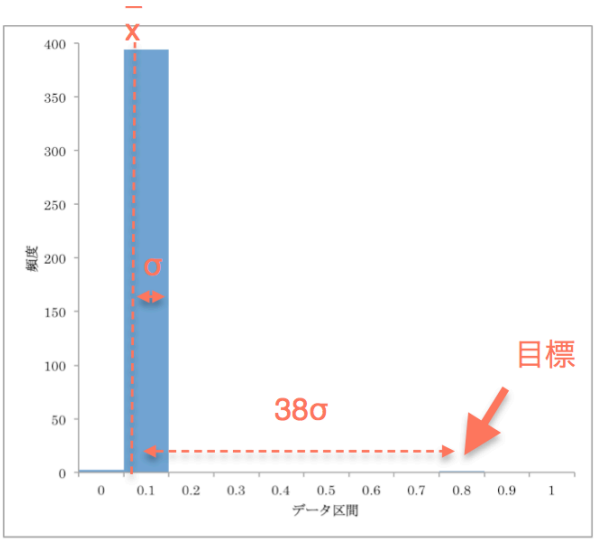


図 5　TFIDF 手法の cos 類似度の分布

表 4　実験条件

選択した話題	55066. Bird Flu outbreak in Belgium
種とした記事 DATE_TIME	APE20030419.1630.0392, 2003-04-19 19:00:00
目標とした記事 DATE_TIME	APE20030420.0630.0101, 2003-04-20 07:13:00
種と目標の間の負例集合の数	398

表 5　「55066. Bird Flu outbreak in Belgium」のアノテーション（[12]より一部抜粋）

Seminal Event	
WHO	Belgian
THAT	The government issued tough restrictions on the transport of chickens and eggs.
WHEN	04/16/2003
WHRE	Brussels, Belgium
Topic Explication	
The Belgian government issued tough restrictions on the transport of chickens and eggs Wednesday after confirming its first outbreak of bird flu. The disease is often fatal for birds but is almost never contracted by humans. The significance of this particular article is that there have been a few deaths associated with this recent bird flu epidemic. The entire poultry sector will be immobilized and no contact between farms and other related businesses will be made. The government agency has begun culling some 250000 birds including, chickens, lay hens, ducks and geese, on farms located in a three kilometer zone around the Dutch border, in Meeuwen Gruitrode, where the first case was confirmed.	

ⓒ2014 Information Processing Society of Japan

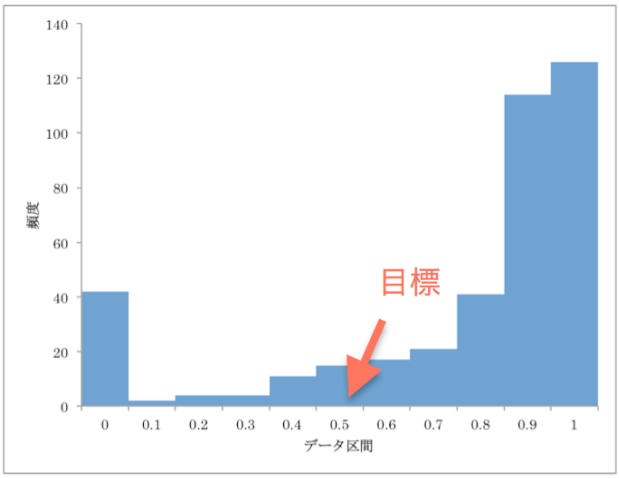


図 6　提案手法（全語使用）の cos 類似度の分布

表 6　比較実験のまとめ

	Betweenness	TFIDF
$\cos\left(d_{seed}, d_{target}\right)$	0.63	0.72
$\bar{x}$	0.068	0.020
σ	0.059	0.018
$\frac{\cos\left(d_{seed}, d_{target}\right)-\bar{x}}{\sigma}$	9.5	38

表 7 各手法における語の重要度が上位の語

Term	Betweenness	Term	TFIDF
poultry	0.175	**bird**	0.120
birds	0.112	**dutch**	0.120
veterinarian	0.103	**flu**	0.112
dutch	0.089	agriculture	0.080
said	0.064	netherlands	0.074
had	0.057	influenza	0.065
outbreak	0.056	**poultry**	0.065
was	0.053	veterinarian	0.065
people	0.048	**birds**	0.060
health	0.046	crossed	0.060
infected	0.046	**outbreak**	0.056
agriculture	0.042	infected	0.041
ministry	0.042	population	0.041
bird	0.041	ad/tcs	0.037
flu	0.041	amsterdam	0.037

表 8　提案手法（全語使用）における語の重要度が上位の語

Term	Betweenness
the	0.327
and	0.145
dutch	0.035
was	0.022
said	0.022
outbreak	0.021
bird	0.017
flu	0.017
for	0.017
not	0.015
died	0.013
agriculture	0.013
ministry	0.013
veterinarian	0.010
had	0.010

予備実験における全語使用の意味について述べる．まず，cos 類似度の分布を図 6 に示す．種と目標の cos 類似度は 0.56 であり，図 6 から種と目標の cos 類似度が負例集合の分布に埋もれてしまい十分な性能を発揮しないことがわかる．次に，この原因を探るために，この手法で求めた語の重要度を表 8 に示す．表 8 から，”the”や”and”など，どのような話題の文書にも普遍的に出現する語が重要語として上位になっていることがわかる．これは，定冠詞や関係詞などの文法上，語や文をつなぐ働きをする語は，どのような話題の文書においても媒介中心性が高くなる傾向にあるからだと考えられる．本論文では，こうした文法上語や文をつなぐ働きをする語を除外語として表 1 にまとめ，共起ネットワークのノードから除外する前処理を行うことで性能の改善を試みた．表 1 の語以外にも除外すべき語があるかは，今後の検討で探っていきたい．

TFIDF 手法と比較した提案手法の利点について考察する．提案手法は，ひとつの記事から語の共起ネットワークを生成し，そのネットワーク特徴から語の重要度を得る．これは話題追跡タスクにおいて，TFIDF 手法のように大量のコーパスを必要とせずに，与えられた少数の記事のみから話題追跡が可能であることを意味する．またこれは話題追跡タスク以外にも大量のコーパスを集めることが難しいようなタスクに適用できる可能性があることを意味する．さらに提案手法では，ノードの中心性を表す指標である媒介中心性のみに着目したが，他のネットワーク特徴を組み合わせることで更なる性能の向上が見込めると考えられる．

ⓒ2014 Information Processing Society of Japan

6. 今後の課題

話題追跡のための共起ネットワークからの語の重要度における今後の課題を次に示す.

- 中心性の比較：媒介中心性のみでなく，次数中心性や近接中心性などの他の中心性指標との比較を行う.
- 従来技術との融合：性能を向上するためにネットワーク特徴のみでなく，従来の TFIDF 手法などとの融合を行う.
- 評価実験の実施：いくつかの話題に対して実験を行い，cos 類似度の分布を調べる評価実験を行う.
- 有効性の理論化：評価実験結果を元に，なぜ共起ネットワークによる語の重要度が話題追跡タスクに有効であるかを理論化する.

7. おわりに

本論文は，話題追跡タスクに対し，少数の記事だけを用いる話題追跡手法の提案を目的として，共起ネットワークによる語の重要度の検討を行った．本論文は，まず話題追跡タスクを情報検索システムのアプローチがとれるように定式化を行った．次に語の共起ネットワークの特徴から求めた語の重要度をベクトル空間モデルへ適用した．TFIDF 手法との比較実験の結果，提案手法が話題追跡タスクに対して有効であることを示した.

今後の計画は，無作為にいくつかの話題を選択し cos 類似度の分布を調べる評価実験を行う予定である．また，種と目標の cos 類似度と負例集合の cos 類似度の平均の差の分布を調べ，提案手法がどのような話題に対しても適用できる手法であるかを調べる予定である.

参考文献

1) 高間康史. "6.Web 情報ストリーム (< 特集> Web ダイナミクス: 膨大で動的な Web 情報の知的処理に向けて)." 情報処理 44.7 (2003): 720-725.
2) J. Allan, editor. “Topic Detection and Tracking: Event-based Information Organization”, Kluwer (2002).
3) Lo, Yuen-Yee, Jean-Luc Gauvain. "The LIMSI topic tracking system for TDT2001." *Proc. TDT* (2001).
4) 増田真也, 石川孝. “SVM 手法を用いた Twitter トレンドのトピック追跡.” 情報科学技術フォーラム講演論文集 (2013).
5) 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 43.6 (2002): 1825-1833.
6) 小熊淳一, 内海彰. “語の共起情報を用いた文書クラスタリング.” 人工知能学会第 19 回全国大会論文集, 2E1-01 7 (2005).
7) Blanco, Roi, Christina Lioma. "Graph-based term weighting for information retrieval." *Information retrieval* 15.1 (2012): 54-92.
8) 塊 タカユキ, 石黒 昭博. “総合英語 Forest, 5th edition”, 桐原書店 (2006).
9) 北 研二, 津田 和彦, 獅々堀 正幹. “情報検索アルゴリズム” 共立出版 (2002).
10) 増田 直紀, 今野 紀雄. “複雑ネットワーク―基礎から応用まで” 近代科学社 (2010).
11) TDT5 Multilingual Text.
URL://catalog.ldc.upenn.edu/LDC2006T18
12) TDT5 Topics and Annotations.
URL://catalog.ldc.upenn.edu/LDC2006T19

ⓒ2014 Information Processing Society of Japan