ネットワークを用いたテキストマイニングによる類似ニュース記事の可視化

今井 貴之　　中村 啓太　　大豆生田 利章
群馬工業高等専門学校 生産システム工学専攻
ap15803@ipc.gunma-ct.ac.jp, {nakamura, mame}@ice.gunma-ct.ac.jp

1 はじめに

スマートフォンやタブレット端末の普及に伴い，電子化された文書が増加し，このような文書を扱う機会も増加傾向にある．このような電子化された文書の中で最も利用されているものがウェブサイト上のニュース記事である．例えば，多くの人が利用する SNS においても，ウェブサイト上のニュース記事の URL を付加することで情報の共有を行っている．このように，電子化された文書は扱いが容易であるが，ゆえに増大しやすい．膨大となった文書の中から求めている情報のある文書を見つけることは大変困難である．この問題に対する解決策として，電子化された文書を要約し，情報をまとめるという手法が多く取られている．この手法は効果的であるが，どの文書からどの文書までを要約することが最も効率的となるかの判断を行う必要があり，要約する文書を選別する基準が求められる．

そこで本研究では，ウェブサイト上のニュース記事を内容，話題に沿って分類する手法を提案する．ニュースサイトの記事に対して特徴語群を生成し，この特徴語群を用いて記事の類似度を算出する．算出された類似度を用いてネットワークを生成し，類似した内容かどうかの関係性を可視化する． これにより，情報を集約することができる．また，時系列データを用いた可視化を行うことで，あるニュース記事に関する時系列の推移を把握しやすくできる．

2 関連研究

ニュース記事を分類･要約し，情報推薦を目的とした研究は多く行われている．テキストマイニングによって類似している複数のニュース記事から重要な情報が書かれている部分を判断し，それらのニュース記事を要約するという研究 [1][2] では，多くの記事から重要な情報だけを抽出し，要約することができている．しかしながら， 要約された情報の中に必ずしもユーザの求める情報が含まれているとは限らない．

また，文書の話題を決定するトピックモデルの一つである LDA に基づいた Relational Topic Model(RTM) を開発し，これを用いて Document Networks を構成し，情報推薦を行う研究 [3] も存在する．この手法では文書間の関係性を可視化し，文書中の出現単語の推測を行うことができたが，LDA の特性を引き継いでいるため，文書の単語数が少ない場合などにうまく分類することができない．

さらに，ニュース記事を Name-Entity-Extraction（固有表現抽出）と TF-IDF 法を使った Bag-of-Words によるカテゴライズによってニュース記事を分類し，可視化する研究 [4] では，固有表現によってニュース記事を分類することができている．しかしながら，同一カテゴリ内における文書の内容についての類似性が少ないという欠点がある．

本研究ではこれらの研究とは異なり，テキストマイニングとネットワークを利用して可視化を行う．

3 提案手法

記事の本文から特徴語を抽出する手法として自然言語処理の TF-IDF 法と N-gram 法の組み合わせを用いる．ニュース記事に対し，形態素解析機 MeCab[5] を使用することで，文章を単語に分け，品詞ごとに分類する．本手法では「名詞」,「動詞」,「形容詞」の三つを扱う．これらの単語から特徴語を生成するため TF-IDF 法を使用する．

TF-IDF 法は，文章中における単語の重み付けの一種であり，以下の式 (1) から (3) で定義される．

$$tfidf = tf * idf \quad (1)$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}} \quad (2)$$

表 1: 重複表現の計算例

n-gram 数	単語	*tf* 値の倍率
1-gram	安倍 晋三 首相	0.95
2-gram	安倍晋三 晋三首相	0.95
3-gram	安倍晋三首相	1.00

$$idf_j = \log \frac{|D|}{\{d_j | n_{i,j} > 0\}} \quad (3)$$

ここで $n_{(i,j)}$ は単語 w_j の文章 d_i での出現頻度，$|D|$ は文章の総数である．本研究ではニュース記事を対象としており，一般語であっても重要な情報として考えられる．*idf* 値は一般語のような多くの文書で出現する語の数値を下げる役割をしている．

そこで，本研究では *tf* 値のみを採用する．しかしながら，*tf* 値のみでは未知語に対応できないため，特徴語の生成に不十分である．そこで本研究では N-gram 法を組み合わせる．

N-gram 法は，文章を単語単位ではなく文字単位で分解し，隣り合う n 個を一つの要素として扱う手法である．本研究では，単語の隣り合わせを用いた単語 N-gram 法を用いる．この手法によって，Twitter などのインターネットサービスで頻出する未知語に対応することができる．

本研究ではニュース記事の本文を形態素解析し，単語 N-gram 法によってワードリストを生成する．ここで単語 N-gram 法は 1 から 5-gram（5 単語連続）までを使用する．生成されたワードリストから *tf* 値を計算する．このとき単語 N-gram 法によって重複する表現があるため，重複表現がある場合は *tf* 値を下げている（表 1）．この *tf* 値の上位 30 語を特徴語として保持し，これによって他の記事に対する類似度を Jaccard 係数によって算出する．

Jaccard 係数は，ある二つの集合の共起の度合いを表すものであり，以下の式 (4) によって定義される．

$$sim(C_i, C_j) = \frac{|c_i \cap c_j|}{|c_i \cup c_j|} \quad (4)$$

ここで，C_n は n 個目の記事であり，c_n は n 個目の記事の特徴語群である．

本研究では，この係数によって決められたニュース記事間の類似度を用いて，情報の推薦を行うことができる．

4 比較実験

複数のニュースサイトで掲載された記事に対し，特徴語を生成する．また，その特徴語によって他の記事との類似度を算出する．この類似度を用いて，ネットワークを生成する． 提案手法を用いて生成したネットワークと既存の手法（TF-IDF 法）を用いて生成したネットワークを比較および検討する．

4.1 実験内容

本研究で扱うニュースサイトの記事は Yahoo!ニュース [6]，読売新聞 [7]，朝日新聞 [8]，日経新聞 [9]，産経新聞 [10] のサイトにおいてトップ記事として掲載されたものを使用する．新しい記事に対して解析を行い，提案手法によって特徴語を生成する．この特徴語から Jaccard 係数を用いて類似度を算出し，「Cytoscape」[11] によって記事をノード，類似している記事をリンクで接続したネットワークを生成，可視化を行う．このとき，ノードには属性として取得日時を付与する．本実験では 2015 年 5 月 15 日から 2015 年 5 月 22 日までの期間で収集した計 1454 件のニュース記事を扱う．提案手法を用いて作成したネットワークと TF-IDF 法のみを用いて作成したネットワークを比較し，提案手法の有効性を検証する．

4.2 実験結果・考察

図 1 にノード数 615，リンク数 1198 の提案手法によるネットワークの全体図を示す．また，図 2 にノード数 395，リンク数 339 の TF-IDF 法によるネットワーク全体図を示す．またこれらの図では，取得日ごとにノードの色，形を変更している．それぞれにおける主要な話題とそのサブネットワークについての関係を表 2 に示す．図 1，2 より，どちらの手法でもそれぞれの話題ごとにサブネットワークが生成されることがわかる．また，図 1 の提案手法の方が，図 2 の TF-IDF 法より大きいサブネットワークが生成されていることがわかる．また，図 2 のサブネットワークの多くが同じ日，もしくは近傍の日のニュース記事によって構成されている．これに対し，図 1 では同じサブネットワーク内であっても様々な日にちのニュース記事で構成されているものが多い．さらに，図 2 において，ノード数が少ないサブネットワークの数は図 1 のサブネットワークに比べて多くなっている．以上より，提案手法で構成されたネットワークの方がより多くのニュース記事を分類できていることがわかる．

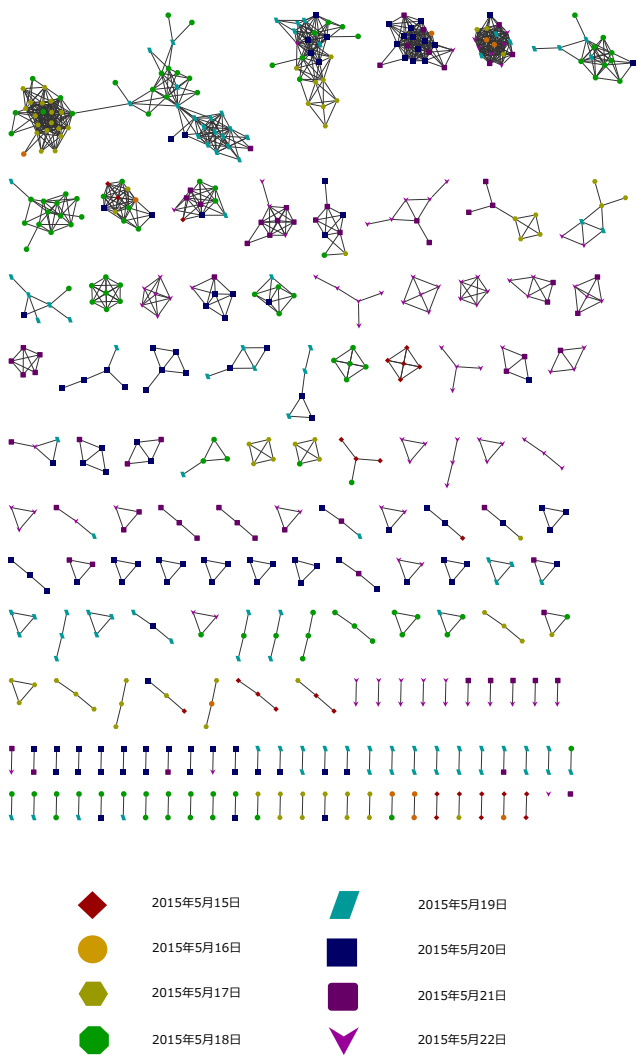

図 1: 提案手法によるネットワーク全体図 (ノード数 615，リンク数 1198)

次に，それぞれの図 1，2 において，一つのサブネットワークについて比較した結果を示す．図 3 に図 1 中の「大阪都構想」についてのサブネットワークを示す．図 3 のノード数は 57，リンク数は 260 である．また同様に，図 4 に図 2 中の「大阪都構想」についてのサブネットワークを示す．図 4 のノード数は 20，リンク数は 25 である．これらについて比較すると，図 3 では 1 つのサブネットワークであることに対し，図 4 では複数のサブネットワークに分散していることがわかる．また，全体のネットワークと同様に，図 4 では同じ日，もしくはその近傍の日のニュース記事ごとでネットワークが構成されていることがわかる．図 3 において左側の話題は「大阪都構想の是非を問う住民投票」についてであり，右側の話題は「橋下徹大阪市長の政界引退」についてである．これらはどちらも「大阪都構想」に関連している話題であり同一サブネットワーク内に集約していることがわかる．一方で，図 4 では

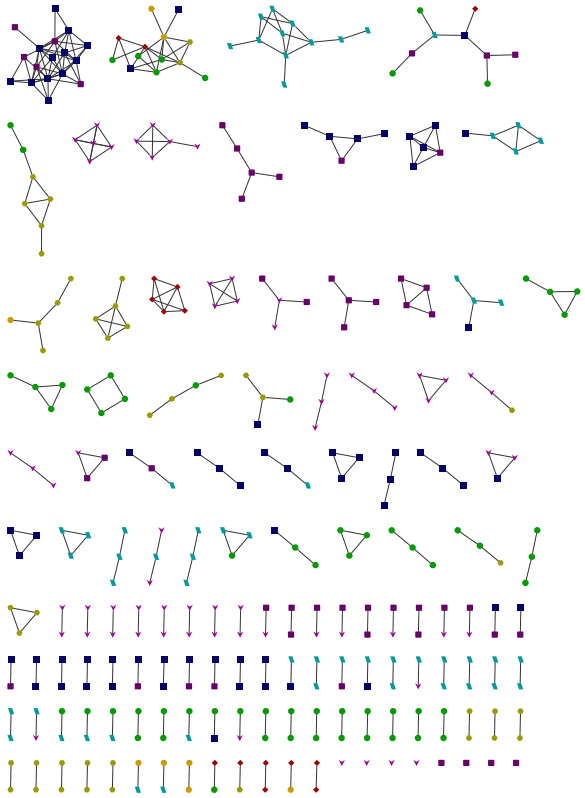

図 2: TF-IDF 法によるネットワーク全体図 (ノード数 395，リンク数 339)

表 2: 話題とサブネットワークの関係

話題	提案手法		TF-IDF 法	
	ノード数	リンク数	ノード数	リンク数
大阪都構想	57	260	20	25
吉田屋炎上	26	102	10	10
イルカ追い込み漁	20	115	17	52
核拡散防止条約	18	126	5	3
オスプレイ問題	14	39	3	3
大相撲	10	31	9	8
硫酸タリウム事件	12	50	13	23

これらの話題ごとにサブネットワークが生成されている．さらに，図 3 のネットワークにのみ出現しているノードがあり，それらのニュース記事はすべて「大阪都構想」に関連している記事であった．TF-IDF 法でこれらのニュース記事が出現しなかった理由として，TF-IDF 法は記事中においてわずかに異なる単語が出現すると，これらの単語が特徴語として抽出されるため，他の重要単語を抽出することができず，結果として他の多くの類似した記事と異なる記事であると判別してしまうことが挙げられる．

以上より，ニュース記事を分類する手法において，TF-IDF 法より提案手法の方が有効である．

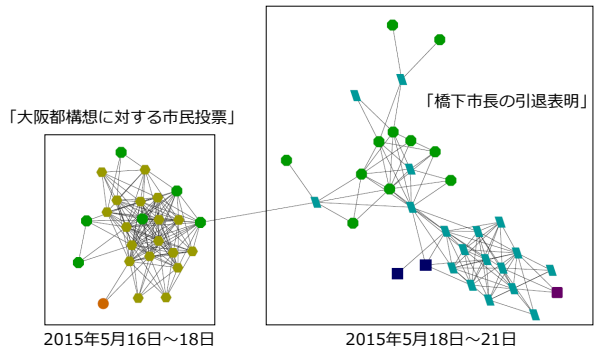

図 3: 提案手法における「大阪都構想」についてのサブネットワーク図 (ノード数 57，リンク数 260)

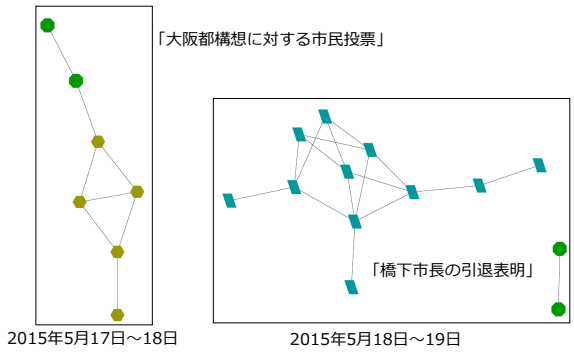

図 4: TF-IDF 法における「大阪都構想」についてのサブネットワーク図 (ノード数 20，リンク数 25)

5　おわりに

本研究ではニュース記事から TF-IDF 法と N-gram 法の組み合わせによって特徴語を生成し，Jaccard 係数によって類似度の算出を行った．この類似度からネットワークを作成し，類似ニュース記事の可視化および分類を行った．さらに，既存の手法である TF-IDF 法との比較を行った．結果として，提案手法は TF-IDF 法よりも多くの類似した記事を分類することができた．

今後の課題として，三つのことが挙げられる．

一つ目は可視化の手法の検討についてである．本研究ではリンク数のみを用いてネットワークを構成し，ノード数の大きいものから表示している．ネットワークの構成をさらに理解しやすくするため，ネットワークごとの話題を解析し，その話題別にネットワークを表示させることを目標とする．

二つ目は対象期間についてである．本実験では一週間文のニュース記事を対象としたが，これを一か月間などに増加させた場合に，ネットワークがどのように変化するかの検証を行う予定である．

三つ目は比較対象についてである．本実験では TF-IDF 法を比較対象としていた．このほかに文書分類の手法であるトピックモデルの一つの LDA との比較実験も行う予定である．

参考文献

[1] Simone Teufel, Marc Moens, *Summarizing Scientific Articles:Experiments with Relevance andRhetorical Status*, Association for Computational Linguistics, 409–445, 2002.

[2] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam, *Centroid-based summarization of multiple documents*, Information Processing & Management, vol.40, num.6, pp919–938, 2004.

[3] Jonathan Chang, David M. Blei, *Relational Topic Models for Document Networks*, Proc. of Conf. on AI and Statistics 2009 (AISTATS'09), 2009

[4] Marko Grobelnik and Dunja Mladenic, *Visualization of News Articles*, Informatica (Slovenia), vol.28, num.4, pp375–380, 2004.

[5] MeCab: Yet Another Part-of-Speech and Morphological Analyzer: `http://taku910.github.io/mecab/`, (2015/1/9 アクセス).

[6] Yahoo! Japan ニュース: `http://news.yahoo.co.jp/`, (2015/1/9 アクセス).

[7] 読売新聞 (YOMIURI ONLINE): `http://www.yomiuri.co.jp/`, (2015/1/9 アクセス).

[8] 朝日新聞デジタル: `http://www.asahi.com/`, (2015/1/9 アクセス).

[9] 日本経済新聞: `http://www.nikkei.com/`, (2015/1/9 アクセス).

[10] 産経ニュース: `http://www.sankei.com/`, (2015/1/9 アクセス).

[11] Cytoscape: `http://www.cytoscape.org/`, (2015/1/9 アクセス)