Vol.2019-MPS-123 No.43
Vol.2019-BIO-58 No.43
2019/6/19

オンラインニュースサイトにおける類似意見の抽出

山口 雄也[1,a)]　伏見 卓恭[1,b)]

概要：オンラインニュースサイトでは，記事に対する個人の意見を自由に投稿できる．ユーザの意見は多種多様であり，時折，議論が盛んに行われる．多様な意見から類似意見を抽出し，議論構造を可視化することで，他者の意見や考えの視点を網羅的かつ俯瞰的に知ることができる．本研究では，あるニュース記事に関するユーザのコメントに対して，返信機能を用いて議論しているコメント群を抽出，集約，可視化する手法を提案する．

Yuya Yamaguchi[1,a)]　Takayasu Fushimi[1,b)]

1. はじめに

Yahoo!ニュースなどのオンラインニュースサイトには，ニュース記事に対して様々なコメントが投稿されている．あるユーザのコメントに対して返信コメントを投稿する機能や，「そう思う/そう思わない」といった評価機能が備わっている．コメントによっては，多くの返信コメントが投稿され，議論が盛んに行われている．Web 上の投稿における議論に関する研究が多くなされており [1]，議論構造を分析することは重要な研究課題である．User Generated Contents であるニュースコメントにはよくあることであるが，ニュース内容とは関係のないコメントが書き込まれることも多々ある．それらの判別のために藤田らは，「建設的」という観点でコメントの順位を付けし議論を活発化させようと試みている [2].

本研究ではコメントに表出するユーザの意見を整理するため，議論ツリーと呼ぶ構造を新たに提案する．議論ツリーはルートコメントに対する返信コメントをノードとしたツリー構造であり，類似の観点でのコメントは同一のサブツリーとなり，異なる視点のコメントは別のサブツリーとなる．また，議論の過程で話が変わったり，トピックドリフトが発生した場合，サブツリーの枝分かれとして検出できると考える．

[1] 東京工科大学　コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology
[a)] c011627257@edu.teu.ac.jp
[b)] fushimity@stf.teu.ac.jp

2. 提案手法

提案手法の枠組みでは，コメント文の集合を $\mathcal{D} = \{d_1, \ldots, d_N\}$，単語集合を $\mathcal{W} = \{w_1, \ldots, w_M\}$ とし，各コメント文は M 次元の単語頻度ベクトル $\mathbf{b}_i = [b_{i,j}]_{j=1}^{M}$ で表現する．ここで，$b_{i,j}$ は，コメント文 d_i における単語 w_j の出現頻度を表す．以下，コメント文は一般化して文書と呼ぶ．任意の文書 d_i と d_j 間に類似度 $\rho(d_i, d_j) = \cos(\mathbf{b}_i, \mathbf{b}_j)$ が得られたとき，閾値 α を定め，$\rho(d_i, d_j) > \alpha$ となる文書間にリンクを付与することで，議論ツリーとよぶ木構造を構築する．

与えられた文書集合に含まれる文書 d をノードとみなし，類似度の高い文書間にリンクを付与することで，ツリー $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ を構築する．具体的には，文書 $d_i \in \mathcal{D}$ が投稿された時刻を $d_i.time$ としたとき，投稿された時刻が早い文書から順にツリーにノードとして追加する．つまり，時間発展するツリーを構築することになる．具体的には，文書 d_i が投稿された時刻より前に投稿された文書集合を $\mathcal{D}^{(d_i)} = \{d \in \mathcal{D} \ ; \ d.time < d_i.time\}$ と表す．文書 d_i について，各文書 $d \in \mathcal{D}^{(d_i)}$ との類似度 $\rho(d, d_i)$ を計算し，最も類似する文書ノード $\hat{d}$ から d_i にリンクを付与する．したがって，文書ノード d_i の親ノードは $P(d_i) = \hat{d} = \underset{d \in \mathcal{D}^{(d_i)}}{\arg\max}\, \rho(d, d_i)$ となり，ツリーのノード集合と $\mathcal{V} \leftarrow \mathcal{V} \cup \{d_i\}$ リンク集合 $\mathcal{E} \leftarrow \mathcal{E} \cup \{(\hat{d} \longrightarrow d_i)\}$ で定義される．すなわち，既に投稿されている（=既にツリーの一部になっている）ノード $d \in \mathcal{D}^{(d_i)}$ のうち，最も類似するノードの子ノードとして，d_i をツリーに追加する．

Vol.2019-MPS-123 No.43
Vol.2019-BIO-58 No.43
2019/6/19

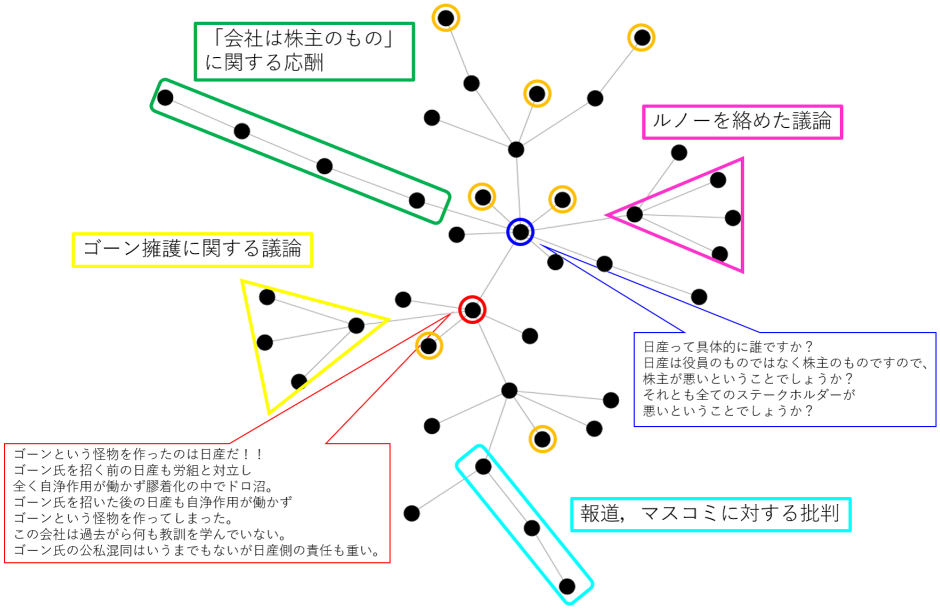

図 1 4 月 11 日：「ゴーンという「怪物」を生んだのは誰か　日産“権力闘争史”から斬る」に関する議論ツリー

次に，類似度閾値パラメータ α を導入する．すなわち，文書 d_i について，最大類似度 $\rho(\hat{d}, d_i) = \max_{d \in \mathcal{D}^{(d_i)}} \rho(d, d_i)$ が閾値 α を超える $(\rho(\hat{d}, d_i) > \alpha)$ 場合のみ $\hat{d}$ から d_i にリンクを張り，そうでない場合にはリンクを付与せず，d_i は新たなツリーの根（root）となる．このことを便宜上，$P(d_i) = d_i$ と記す．適切な閾値 α を設定することで，異なるトピックの文書群は異なるツリーを形成する．以降，この一連の手順により得られるツリー群を議論ツリーと呼ぶ．$\alpha < 1$ の値が大きいほどツリーの数は増え，$\alpha = 0$ で単一のツリーとなる．

3. 評価実験

評価実験では，Yahoo!ニュースの記事に対して投稿されたコメントを収集したものを用いる．本稿では，4 月 11 日の記事「ゴーンという「怪物」を生んだのは誰か　日産“権力闘争史”から斬る」に対するコメント群を用いる．その中でも，多数の返信コメントがついたコメント（ルートコメント）および返信コメントから議論ツリーを構築した．類似度閾値パラメータは $\alpha = 0.15$ とした．図 1 に，議論ツリーを可視化したものを示す．図 1 において，赤ノードがルートコメントである．橙ノードは，「脱兎ダットサン」，「タンスにゴーン」などのふざけた投稿であったり，特殊な表現の投稿などである．これらは，類似度の意見が存在しないため，子ノードが存在しなくリーフノードとなったと考えられる．青ノードは，「ゴーンという怪物を作ったのは日産だ」というルートコメントに対する反対意見として「日産というのは株主のもので，株主が悪いのか？」という趣旨のコメントをしている．青ノードコメントで初めて「株主」という単語が現れ，以降の「株主」に関するコメントがサブツリーのノードとなっている．顕著な例として，緑で囲ったノード群であり，それらは，「会社は株主のもの」という主張の青ノードに対する返信やそれに対する応酬コメントが連なっている．その他にも，桃ノード群は「ルノー」を絡めたコメント，水色ノード群は「報道，マスコミに対する批判」のコメント，黄色ノード群は「ゴーン擁護」の意見を持つコメントである．このように，ルートコメントに対して多種多様な観点で返信されており，それらをサブツリー構造として表現できている．

4. おわりに

本研究では，Web 上に投稿されたユーザの意見を整理，集約し，コメント間の議論構造を分析するために，議論ツリーという構造を提案した．実データを用いた評価実験では，サブツリーに類似のコメントが集まっており，どのような観点の意見が存在するのかを俯瞰できることを確認した．今後の課題として，コメント間の類似度に投稿時間の時間差を導入することで，より関連の深いコメントをつなげることを目指す．さらに，議論ツリーからコメントの観点となるキーワードを自動で抽出する手法を模索していく．

謝辞　本研究は，JSPS 科研費（No.16H02904）の助成を受けたものである．

参考文献

[1] Habernal, I. and Gurevych, I.: Argumentation Mining in User-Generated Web Discourse, *Computational Linguistics*, Vol. 43, No. 1, pp. 125–179 (2017).
[2] 藤田綜一郎，小林隼人，奥村　学：建設的ニュースコメントの順位付けのためのデータセット構築，技術報告 14 (2018).