D-003

オンラインニュースサイトにおける議論構造の可視化
Visualization of Discussion Structure in Online News Site

山口 雄也 † 伏見 卓恭 †
Yuya Yamaguchi Takayasu Fushimi

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，図 1 のような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．

2. 関連研究

本稿では，ニュース記事へのコメント群に対して，投稿順序を保持し，類似する文書群をつなぐことによりツリー群を構築し，効果的に可視化する手法を提案する．時系列文書可視化の関連研究として，Ishikawa らの T-Scroll がある [3]．このシステムでは，時間的に離れた文書間の影響力が，指数的に小さくなる重みを導入した類似度を定義し，k-means 法によりインクリメンタルなクラスタリングを実現している．そして，隣接時刻間のクラスタ間に関連度を定義し，関連度の強いクラスタ間にリンクを付与することで，各時刻におけるクラスタリング結果を時間軸上にプロットしている．本稿のトピックフォレスト構築時にも，投稿間隔と文書間の類似度を考慮する点で関連する研究であるが，各文書を極座標平面に布置する点，クラスタとしてまとめず各文書を可視化する点，ツリー構造としての文書間の関係を表現する点で異なる．

キーワード抽出や特徴選択の技術を用いて重要な単語やトレンドワードを抽出し，その頻度をプロットすることで時系列全体を俯瞰する手法として，MemeTracker [4] や EventRiver [5]，CloudLines [6]，STREAMIT [7] などがある．Keim らの EventRiver [5] では，日々の出来事（イベント）について記されたニュースやブログの記事群に対して，短いタイムスパンのバケツに文書を分ける．バケツ内の文書をクラスタリングすることで，temporal-locality クラスタと呼ぶ時間的凝集性と意味的凝集性の高い文書群に分割する．そして，得られたクラスタ群をメタクラスタリングすることで，時間的に離れていても内容的に類似するような，長期間にわたるイベントに関する文書クラスタを一つのグループとしてまとめあげることができる．横幅をグループの生存時間，縦幅を各時点での影響度とした成長曲線を描くことで，イベントの起点や終点，盛り上がりなどが一目瞭然となり，関連イベントの発見も可能にする可視化結果を実現している．

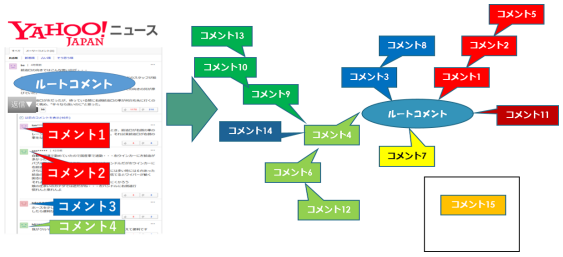

図 1: 議論ツリー構築

3. 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

†東京工科大学コンピュータサイエンス学部

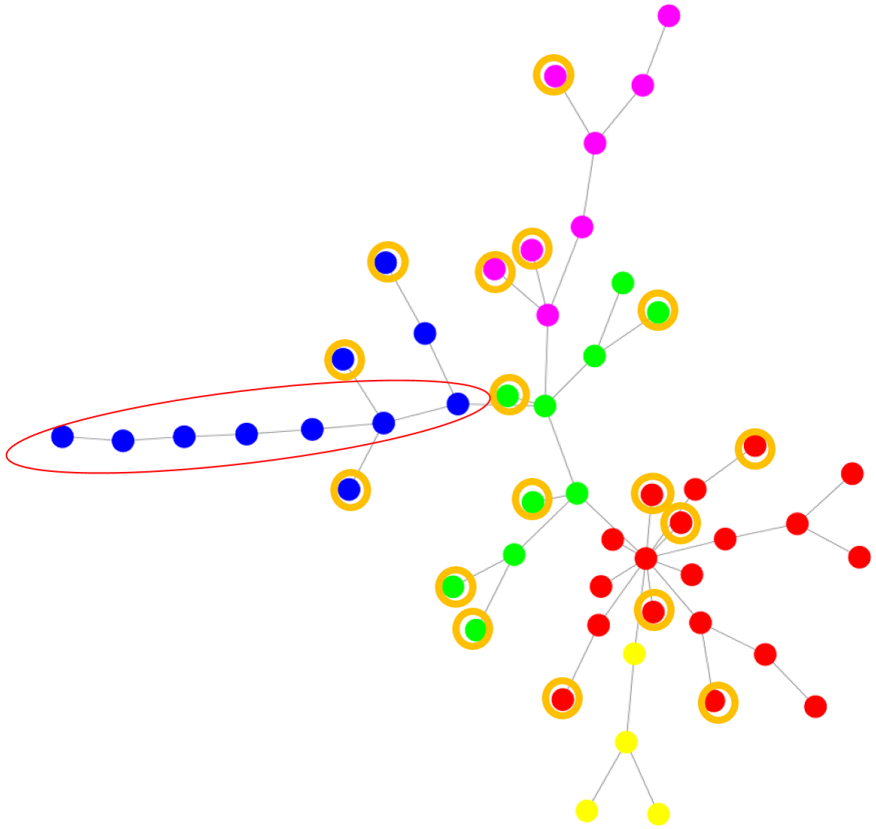

図 2: 4 月 11 日 :「ゴーンという「怪物」を生んだのは誰か　日産“権力闘争史”から斬る」に関する議論ツリー

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 $P(d_i) = \hat{d} = \arg\max_{d \in \mathcal{D}^{(d_i)}} \rho(d, d_i)$ 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．

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．

4. 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」と 4 月 22 日の記事「「母の死刑執行の映像が頭に浮かぶ」林真須美死 刑囚の長男が語る，和歌山毒物カレー事件」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 2 に，議論ツリーを可視化したものを示す．図 2 において，赤ノード群にルートコメントが含まれており，赤，黄緑，黄色，ピンクで色付けされているノード群は同一観点に関するコメントを検出したものである．ルートコメントは，「ゴーンをという怪物を作ったのは日産だ」というゴーンに絡めた批判的な意見であり，その他の赤ノードもゴーンに関しての内容でまとまっている．黄色のコメントは「ゴー

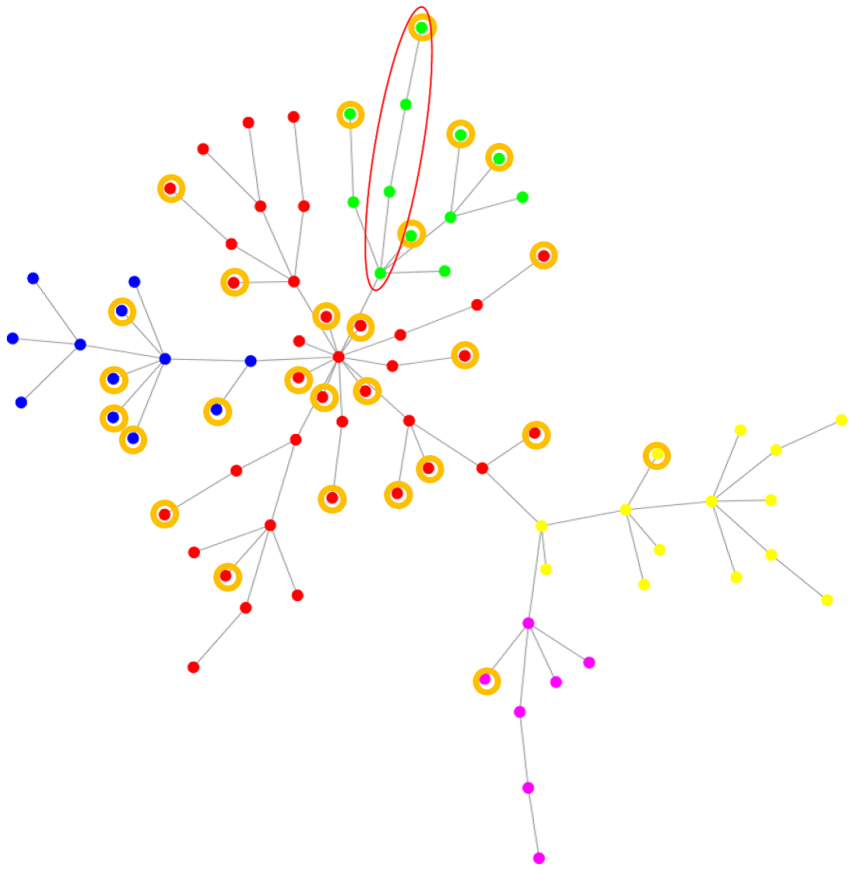

図 3: 4 月 22 日：「「母の死刑執行の映像が頭に浮かぶ」林真須美死刑囚の長男が語る，和歌山毒物カレー事件」に関する議論ツリー

ンだけが悪いのではなく，日産も悪いのではないか」というコメントを筆頭に話題がゴーンから日産に関しての話に移っている．黄緑色のコメントは，青やピンクに続く大きなサブツリーのルートになっている．このことから黄緑色コメントは新たな議論の起点となっていると考えられる．黄緑色のルートコメントは「確かにゴーンは非難される行動をとったと思うが，日産を回復させて稼げる会社にした業績は認めないといけない.」という内容になっており，そのため黄緑色はゴーン関しての批判ではなく，ゴーンを擁護する内容となっている．青のサブツリーのルートノードが「株主総会で叫んでいる人みたいだな.」という発言であり，緑までのゴーンを擁護する内容から株主へと話が変わっている．青のサブツリーでは特に議論が盛んであり，赤い丸で囲まれた部分では「企業は株主のものなのか？」ということについての議論を続いている．ピンクのサブツリーのルートノードは,「ゴーンもブラジルからの移民で元々のエリートじゃないから金に対してアグレッシブ.」という内容であり，事件に関してのゴーンの話ではなく，ゴーンという人物がマスコミやフランスでのゴーンの話になっており黄緑色のゴーン擁護や赤色のゴーン批判とはまた違った視点からコメントである．オレンジで囲んだノードは野次や煽りのコメントであり，後続するコメントの中にこのコメントに類似するコメントがなかったため，リーフノードとなっている．

図 3 に，議論ツリーを可視化したものを示す．図 3 を見ると，赤色ノード群にはルートコメント「確たる証拠がなく，動機もなく，印象だけって感じの事件.」が含まれており，事件に対しての感想や疑問を持っている意見が多い．青のサブツリーのルートノードは「保険金詐欺については長男の言う通り両親だけではないのでは？」というコメントであり，赤ノードのように事件の感想から林容疑者が行った保険金詐欺へと話題が移っている．黄緑色サブツリーでのルートノードでは「この事件は物的証拠も無いし，動悸も不透明．死刑執行だけはするべきではない.」という主張がなされており，今回の犯罪は不自然な点が多く，林容疑者を死刑にするべきではないという擁護の意見であった．赤い丸がついている部分ではルートコメントから「不自然な点が多い」,「犯人は他にいる」,「林容疑者の家族はカレーを食べていないため，彼女以外が犯人とは考えずらい」と

続いている．黄色のサブツリーでは，当時のメディアで報道されていた事件の行動が話されており，特にカレーに入れられていたヒ素についての議論が行われている．ピンクツリーのルートノードは「なんか真犯人説をいうやついっぱいいるけど，さすがに 99%クロなのは間違いない.」となっており，林容疑者が犯人だという話に賛同をしている．

このように，提案手法により構築した議論ツリーでは，

- 類似意見を子ノードとして連結することで意見を集約でき，
- サブツリーへの枝分かれにより話題の転換を検出でき，
- 議論観点の抽出
- リーフノードとしてヤジ，あおりコメントを判別でき，
- ツリーの深さにより，議論の白熱・紛糾レベルを測定できる

ことがわかった．

5. 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).

[3] Ishikawa, Y. and Hasegawa, M.: *T-Scroll: Visualizing Trends in a Time-Series of Documents for Interactive User Exploration*, pp. 235–246, Springer Berlin Heidelberg (2007).

[4] Leskovec, J., Backstrom, L. and Kleinberg, J. M.: Meme-tracking and the dynamics of the news cycle, *KDD*, pp. 497–506 (2009).

[5] Keim, D. A., Luo, D., Yang, J., Ribarsky, W. and Krstajic, M.: EventRiver: Visually Exploring Text Collections with Temporal References, *IEEE Transactions on Visualization & Computer Graphics*, Vol. 18, pp. 93–105 (2010).

[6] Krstajic, M., Bertini, E. and Keim, D. A.: CloudLines: Compact Display of Event Episodes in Multiple Time-Series., *IEEE Trans. Vis. Comput. Graph.*, Vol. 17, No. 12, pp. 2432–2439 (2011).

[7] Alsakran, J., Chen, Y., Luo, D., Zhao, Y., Yang, J., Dou, W. and Liu, S.: Real-Time Visualization of Streaming Text with a Force-Based Dynamic System, *IEEE Comput. Graph. Appl.*, Vol. 32, No. 1, pp. 34–45 (2012).