1R-03

凝縮性に基づく有用単語検出によるトレンドワードの注釈付き可視化

藤野 まり菜 †, 佐藤 杏奈 †, 伏見 卓恭 †, 斉藤 和巳 †, 大久保 誠也 †, 池田 哲夫 †
† 静岡県立大学経営情報学部

1 はじめに

近年 Twitter のような SNS をはじめとするソーシャルメディアが様々な分野において利用されており，その研究が盛んに行われている [1]．また，膨大な量のデータの特徴や関係を理解するための手段として可視化があり，これまでに様々な可視化法が提案されている [2]．

本研究では，Twitter 上で話題になっているトレンドワードから有用単語を検出し，それらの関係を注釈付きで可視化する手法を提案する．有用単語の検出には，凝縮性（cohesiveness）と呼ぶ指標 [3] を用いることにより，長期的にも頻繁に出現し，話題の一貫性度の高い単語（トレンドワード）を求める．得られた単語群を注釈付き可視化 [2] することにより，話題の共通性により単語群をグループに分割するとともに，各グループに対し，そこでの特徴的な単語をアノテーションワード（注釈語）として抽出し，可視化する．また，2014 年 1 月から 6 月までのトレンドワードを対象にした計算機実験により，提案手法の特性や有効性の評価を行う．

2 提案手法

提案法は次の 3 ステップで構成される．

1. 凝縮性の高い単語集合抽出
2. 抽出単語の類似度に基づく最小全域ツリー構築
3. ツリーカットによるアノテーションワード抽出

以下では，各ステップの詳細について述べる．

まず，凝縮性の高い単語集合抽出について述べる．トレンドワードとは，今現在 Twitter 上を流れているツイートの中から，短時間で何度も話題に上っているキーワードをリアルタイムに抽出して表示する Twitter の機能である．トレンドワードは 5 分毎に更新され，10 個表示される．各トレンドワードの単語頻度ベクトルを構成する共起単語は，トレンドワード取得時点でのトレンドワードを含む最新 100 件のツイートを対象に日本語形態素解析して，トレンドワードと共起する名詞を抽出した．トレンドワードを収集した時刻集合を τ とし，時刻 $t \in \tau$ のトレンドワード集合を V_t とする．時刻 t のトレンドワード $v \in V_t$ に対し，最新 100 件のツイートより得られる単語頻度ベクトルを $\mathbf{x}_{t,v}$ とし，総共起単語種の集合を η とする．単語頻度ベクトルのペア間の平均類似度は次式となる．

$$\mu = \frac{1}{100|\tau|^2} \sum_{s\in\tau} \sum_{t\in\tau} \sum_{u\in V_s} \sum_{v\in V_t} \rho(\mathbf{x}_{s,u}, \mathbf{x}_{t,v}) \tag{1}$$

ここで $\rho(\mathbf{x}_{s,u}, \mathbf{x}_{t,v})$ は $\mathbf{x}_{s,u}$ と $\mathbf{x}_{t,v}$ のコサイン類似度を表す．また，総ペア数は $|V_t| = 10$ より $100|\tau|^2$ となる．任意のトレンドワード w の出現時刻集合を $\tau(w) = \{t \in \tau\,;\, w \in V_t\}$ とすれば，ワード w に関する単語頻度ベクトル間での平均類似度は次式となる．

$$\mu(w) = \frac{1}{|\tau(w)|^2} \sum_{s\in\tau(w)} \sum_{t\in\tau(w)} \rho(\mathbf{x}_{s,w}, \mathbf{x}_{t,w}) \tag{2}$$

よって，長期的にも頻繁に出現し，話題の一貫性が高い単語の評価尺度として，トレンドワード w に対する凝縮度 [3] は $\phi(w) = |\tau(w)|(\mu(w) - \mu)$ と導ける．提案法では，凝縮度 $\phi(w)$ の高い上位 M 個のワード集合 $W = \{w_1, \cdots, w_M\}$ を抽出する．

次に，最小全域ツリー構築法について述べる．ワード w に対し各時刻 t での単語頻度ベクトル $\mathbf{x}_{t,w}$ を合成したベクトルを $\mathbf{y}_w = \sum_{t\in\tau(w)} \mathbf{x}_{t,w} = (\mathbf{y}_{w,1}, \mathbf{y}_{w,2}, \cdots, \mathbf{y}_{w,|\eta|})$ とする．提案法では，ベクトル $\mathbf{y}_w$ のコサイン類似度により，ワード集合 W に対して最小全域ツリー S を構築する．

最後に，アノテーションワード抽出法 [2] について述べる．いま，S から $(K-1)$ 本のリンクをカットして得られる K 個のサブツリーの集合を $\{S_1, \cdots S_K\}$ とする．サブツリー S_k での第 h 番目の単語の出現数を $n_{k,h} = \sum_{w\in S_k} y_{w,h}$，出現単語総数を $n_k = \sum_{k=1}^{|\eta|} n_{k,h}$ としたとき，それぞれの確率を $p_{k,h} = n_{k,h}/n_k$ と $q_k = n_k/N$ とする．ここで $N = \sum_{k=1}^{K} n_k$ である．提案法では，次式のエントロピーを最小にするように貪欲法と局所改善法を組合せた手法でサブツリーを求める．

$$E(S_1, \cdots S_K) = -\sum_{k=1}^{K} q_k \sum_{h\in\eta} p_{k,h} \log p_{k,h} \tag{3}$$

また，ツリー S での第 h 番目の単語の出現確率を $r_h = \sum_{k=1}^{K} n_{k,h}/N$ とする．各サブツリーに対して，次式の Z

Visualizing trend words with annotation by term extraction based on cohesiveness
†Marina Fujino †Anna Sato †Takayasu Fushimi †Kazumi Saito †Seiya Okubo †Tetsuo Ikeda
†School of Management and Information , University of Shizuoka

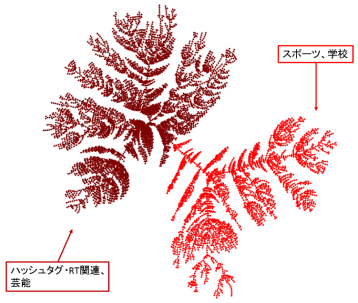

図 1: $K = 2$

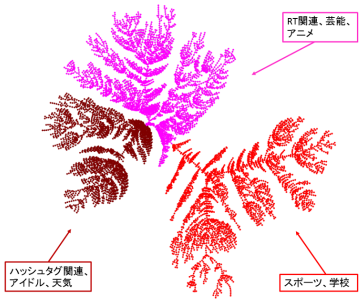

図 2: $K = 3$

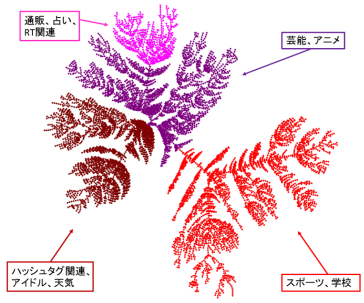

図 3: $K = 4$

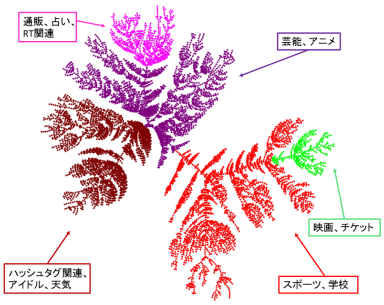

図 4: $K = 5$

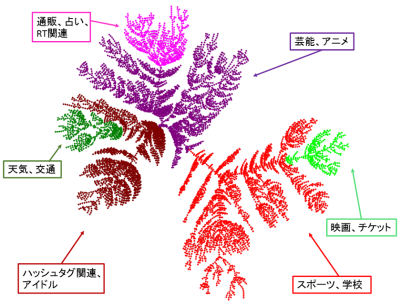

図 5: $K = 6$

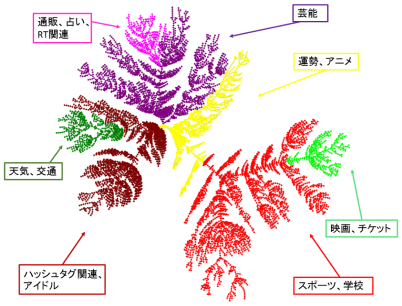

図 6: $K = 7$

スコア $z_{k,h}$ が有意に大きな単語を，アノテーションワードとして選定する．

$$z_{k,h} = \frac{n_{k,h} - n_k r_h}{\sqrt{n_k r_h (1 - r_h)}} \tag{4}$$

提案法では，サブツリー毎に配色して可視化したツリーに，アノテーションワードを付与し，最終結果とする．

3 評価結果

提案手法を評価するにあたって，2014 年 1 月 1 日から 2014 年 6 月 30 日の半年間（181 日間）の Twitter トレンドに出現したトレンドワード群を用いた．総トレンドワード種数 111,029，総共起単語種数 109,250，平均共起単語数 880，平均共起単語種数 308 であった．

$K \in \{2, \cdots, 7\}$ の注釈付き可視化結果を，図 1 から図 6 に示す．図 2 から図 4 において”ハッシュタグ関連・アイドル・天気”とアノテーション付与されたサブツリーが，図 5 において”ハッシュタグ関連・アイドル”と”天気・交通”のサブツリーに分割されるといったように，K が増加するにつれてトピックを階層的に分割していることがわかる．また，色付けによって，サブツリー毎のトレンドワードの量がわかるとともに，図 6 の”スポーツ・学校”と”映画・チケット”といった日常娯楽関係，”運勢・アニメ”と”芸能”といったテレビ関係のように，似たトピックは近くに位置することも確認できる．

これらの結果より，提案法によりトレンドワードを類似するトピックごとのグループに分割し，見やすく妥当なトピックでグループに意味づけをすることができたといえる．よって，提案法の有効性が示唆された．

4 おわりに

本研究では，Twitter のトレンドワードから有用単語を検出し，それらの関係を注釈付きで可視化する手法を提案した．また，計算機実験により提案法によるトレンドワードの可視化結果の特性や有効性を確認した．今後は，既存の可視化法との比較や，提案法を用いてさらに大規模な Twitter データの可視化を行うことで，その有効性の評価を行う予定である．

謝辞 本研究は，総務省 SCOPE(No.142306004)，ふじのくに地域・大学コンソーシアム学術研究，及び，科研費 (No.23500312) の補助を受けた．

参考文献

[1] 佐藤 杏奈，伏見 卓恭，大久保 誠也，斉藤 和巳，風間 一洋，”出現ツイート群の類似度に基づくトレンドワードのタイムライン可視化,” 第 10 回ネットワーク生態学シンポジウム（NETECO2013）, 2013.

[2] 小林 えり，斉藤 和巳，池田 哲夫，大久保 誠也，”L1 埋め込みによるアノテーション付き可視化法,” 第 7 回 Web とデータベースに関するフォーラム（WebDB2014）, 2014.

[3] 小林 えり，斉藤 和巳，池田 哲夫，大久保 誠也，“ 凝縮性に基づく注釈単語検出法とその評価,” 情報処理学会第 77 回全国大会（IPSJ2015）, 2015.