1P-10

時系列的話題追跡のためのツイートの特徴語を用いた探索的閲覧支援システムの開発

糸川 翔太† 白松 俊† 大囿 忠親† 新谷 虎松†
名古屋工業大学大学院工学研究科情報工学専攻†

1. はじめに

近年，Twitter における話題分析のための技術に対する社会的要請は高まってきている．しかし，従来，過去のタイムラインをアーカイブ化して振り返るような使い方は想定されていなかった．そこで本研究ではある話題に関する過去のツイート群における探索的閲覧手法を提案する．

大量に存在するツイート集合において，ユーザは過去にあった大量のツイートの内容を知るはずもなく，何をクエリとして検索すればいいのか分からない．特定の話題の背景を把握するためには，その話題に関するツイートの傾向がどのような流れで移り変わっていったか，どのような出来事がツイートの傾向を変えたかという時系列的な流れを知ることが望ましい． そのために本研究では，異なる期間毎の特徴語を用いてツイート集合の概観をユーザに提示し，時系列に沿って探索的にツイートを閲覧可能な支援システムを提案する．

2. ツイート群に対する探索的閲覧

ツイートに対する探索的閲覧に関する関連研究の中でも，よく用いられる手法としてバースト検出を利用したトピックの可視化がある[2]．しかし，ツイートというデータの性質上，時期毎にデータ量に差がある．例えば，何かイベントがあった時期にはツイート数が著しく増加する場合が考えられる．そのため，従来の手法では，盛り上がった期間の特徴のみ抽出し，それ以外の期間の特徴が欠けてしまう．また，ある固定された区間の特徴のみをユーザに提示するだけでは，その区間内の特徴の推移など，詳細な情報を把握することができない．検索対象に関する知識の乏しいユーザにとって，これらは探索的閲覧における網羅性を満たしていない．そのため，特定部分のトピック抽出や，固定区間の特徴抽出だけでは，ツイートを探索的に閲覧する上で不十分である．従って，本研究では，月毎，週毎，日毎の異なるタイムスケールのツイート集合からの特徴語抽出を実現し，それらをユーザが動的にタイムスケールを調整しながら閲覧することで，時系列に沿ったツイートの探索が可能な探索的閲覧支援システムを開発する．

Implementing an Exploratory Browsing System using Feature Terms of Tweets for Tracking Topic along Time Series
Shouta ITOKAWA, Shun SHIRAMATSU, Tadachika OZONO, and Toramatsu SHINTANI
†Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

3. 特定期間における特徴語抽出手法

本研究では，ツイート集合の概観を提示するために期間を表す特徴語を利用する．単純な頻度を利用した特徴語抽出では本研究における適切な特徴語を抽出することは難しい．そこで対象の期間とそれ以外を分類するのにふさわしい特徴を抽出するために情報利得(Information Gain: 以下 IG)と自己相互情報量(Pointwise Mutual Information: 以下 PMI)を組み合わせた特徴語抽出手法を用いる[1]．

しかしながら，IG で上位にくる語には，特定のユーザのみが頻繁に発言する語が含まれてしまう場合がある．ある期間を表す特徴語として，これらは不適切なものが多い．逆に，複数のユーザが共通して発言する語はその期間の特徴を表す語としてふさわしいと考えられる．そこで，本研究では，IG を拡張したバイアス罰則付き情報利得(Bias-Penalized Information Gain: 以下 BPIG)[1]を用いる．これは，特定のユーザ固有の特徴に対し罰則を与え，除外する手法である．以下に BPIG の式を示す．

$$BPIG(C \mid F_i) = IG(C \mid F_i) - \alpha \max_{k \in K_i} IG(M_k \mid F_i)$$

$$K_i = \{k \mid PMI(m_k, f_i \mid c^+) > 0\}$$

$$PMI(m_k, f_i \mid c^+) = \log \frac{p(m_k, f_i \mid c^+)}{p(m_k \mid c^+) p(f_i \mid c^+)}$$

$$M_k = \{m_k^+ \mid m_k^-\}$$

特定の期間に現れるツイート集合 c^+を正例，それ以外のツイート集合 c^-とする．m_kは c^+に現れるユーザであり，m_k^+は m_kに現れるツイート集合，m_k^-はそれ以外のツイート集合を表す．この手順を，特徴語を発言している全ユーザに対し実行することで，複数のユーザで共通して利用されている語を抽出する事が可能となる．

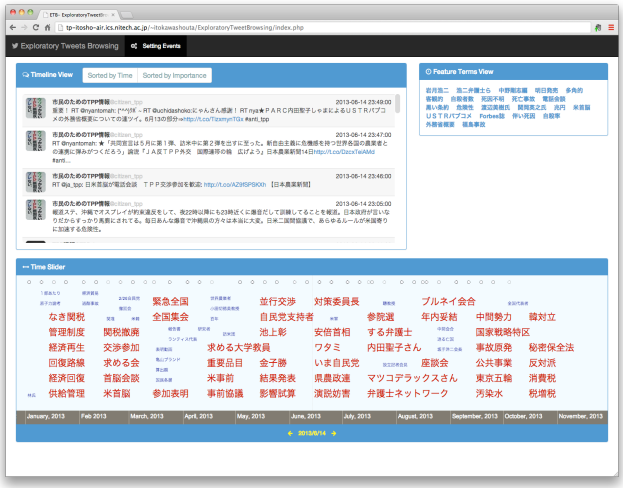

図 1　システムのインタフェース

図 2　タイムスケールの調整

4. ツイートの探索的閲覧支援システム

図 1 に本システムのインタフェースを示す．右上には，指定日のツイートの特徴語が表示される．下部にあるタイムスライダー上には，月毎，週毎の特徴特が表示される．タイムスライダーを左右にスライドすることで日付を指定すると，その日付に対応したツイート一覧が左上に表示される．月毎の流れから，週毎の流れを見たい場合，タイムスケールをズームし，調整することで指定の月の中での細かいツイートの特徴の流れを閲覧することができる (図 2)．また，表示されている特徴語をクリックすることで，その特徴語の期間に対応したツイートの一覧の中で，クリックした特徴語が含まれるもののみを一覧表示することができる．上記の機能を実現することで，ツイート集合全体の概観を提示し，必要に応じてズームやフィルタリングを行いながら探索的なツイートの閲覧を可能とした．

表 1　TPP に関するツイートの 4 月の特徴語

特徴語	Rank (IG)	特徴語	Rank (BPIG)
交渉 参加	1	米 議会	1
事前 協議	2	組織 協議会	2
米 事前	3	参加 撤回	3
日 米	4	協議 合意	4
重要 品目	5	交渉 脱退	5
農産 品	6	歳入 委員会	6
意見 交換	7	世界 農業者	7
自動車 分野	8	女性 組織	8
百 年	9	評論 家	9
研究 者	10	意見 交換	10

5. 評価・考察

本節では，期間毎の特徴語抽出の実験結果およびその考察を述べる．「TPP」という話題に関する 2013 年 4 月から 10 月までのツイートを対象に，本研究での提案手法を用いて，4 月の特徴語抽出実験を行った．4 月のツイート 1,175 件を正例，それ以外の 23,277 件のツイートを負例として特徴語抽出を行った際の上位 20 件を表 1 に示す．TPP に関する情報として，実際 4 月には，日米事前協議が開かれ，日本の TPP 交渉参加が決定している．実験結果からも，「交渉参加」，「事前協議」「日米」「重要品目」など，TPP に関して 4 月の期間を表すような特徴語が抽出できていることが確認できた．実際にシステムを利用しツイートを探索したところ，4 月の第 3 週で日本の交渉参加が決定し，それに伴い第 4 週で TPP 反対に関するツイートが増加したことなど時系列に沿った話題と意見の因果関係を把握することができた．

6. おわりに

本研究では，ツイートを対象とし，時系列に沿った話題や意見の追跡を実現するため，ツイート集合の期間毎の特徴後を利用した探索的閲覧支援システムを開発した．本システムを利用することで，期間を表すのにふさわしい特徴語の抽出，および時系列に沿った話題や意見の因果関係の把握が可能であることを確認した．

参考文献

[1] S. Itokawa, et al: “Estimating Feature Terms for Supporting Exploratory Browsing of Twitter Timelines,” Proc. of the 4th International Conference on E-Service and Knowledge Management, pp. 62-67, 2013.
[2] S. Bhulai, et al: “Trend Visualization on Twitter: What’s Hot and What’s Not? ”, Proc. of the DATA ANALYTICS 2012, pp. 43-48, 2012.