2N-9

Twitter発言の時系列解析に基づくハッシュタグの内容説明

黒木 陽介† 倉門 浩二†† 大石 哲也††† 越村三幸†††† 藤田博†††† 長谷川 隆三††††
† 九州大学工学部電気情報工学科 †† 九州大学大学院システム情報科学府
††† 九州大学情報基盤研究開発センター †††† 九州大学大学院システム情報科学研究院

1 はじめに

近年 Twitter というインターネット上のコミュニケーションサービスが急激に普及し始めた．Twitter とは, 個々のユーザーが「ツイート」と呼ばれる 140 文字以内の短文を投稿するミニブログの一種である．本研究では,Twitter のハッシュタグという機能を用いてミニブログの内容説明を行う．ハッシュタグとは, 投稿時に「#英字列」を入力したタグを付けることで発言をグループ化できる機能である．この機能を用いることで Twitter を疑似的な電子掲示版に見立て, そのグループ化されたツイート群の内容要約を試みる．また,Twitter にはリツイート (RT) という機能がある．これはあるユーザの発言を引用形式で自分のアカウントから発言することである．

Twitter は前述したように手軽に発言できるため, ユーザはその瞬間にしていること, 感じたことを記述することが多い．そのため Twitter を用いて内容説明することができれば, 電子掲示版を用いた際よりも詳細で且つ瞬時に要約された文書を得ることが期待できる．ハッシュタグの内容説明が可能であると, 例えばある離れた地域で講演会等が行われている際, その場にいなくてもそのイベントの内容を知ることができる．

2 関連研究

文献 [1] の研究は, 電子掲示板の要約を行った．掲示板に書きこまれた重要とされる投稿をスコアリングした後, そのスコアの高い上位 4 つを時系列順に表示させることで内容を説明した．

文献 [2] の研究は, ワールドカップでのサッカー日本代表の試合中継のハッシュタグの内容説明をした．扱っている内容は本研究と同じくハッシュタグの内容説明であるが, この研究では選手名もしくはチーム名と指定したサッカー用語の両方を含む発言だけを用いている．これに対して本研究では特定のハッシュタグに限らず内容を説明できるような汎用的なものを目指す．

文献 [3] では, 同じ内容の記事・事件に関して複数ユーザが Twitter でつぶやいた発言から, その記事・事件の要約を行った．この論文では各ツイートを形態素解析したものを合成してある木構造を生成し, スコアの高い枝のみを抽出して要約文とした．扱っている言語が英語ならば効果的だが, 日本語での実現は難しい．

3 提案手法

本節では内容説明の具体的な手法について説明する．まず, 既存手法 [2] の要約手法を基本として採用する．この手法ではハッシュタグ中の発言を時系列別にクラスタリングして, 各クラスタの中から重要と見なせるツイートを抽出する．各クラスタには代表発言が決められているが, その代表発言は各クラスタで頻出する名詞を多く含む発言とする．本研究の手法はまず以下を行う．

1. 同一のハッシュタグが付いたツイートを抽出する．
2. 取り出したツイートを MeCab を用いて形態素解析を行い名詞を取り出す．
3. tf/idf 値を計算して文書スコアを算出する．

MeCab は形態素解析エンジンのひとつである．上の手順の後, 文書スコアを基にスコアの高い順に 5 つのツイートを抜き出し, 時系列順に表示させ内容説明とする．

この手法では RT された発言に対して特定の処理を行っていない．これは RT されたツイートは RT したツイート中に出現するので,RT された文章中に出現した名詞は自動的にある一定の重みが付くと見なせるからである．

D をハッシュタグの発言群,di をその中の各発言とする． またある dt の中に出てくる名詞を $Wt1, Wt2 \ldots, Wtn$ とし, 文書 di のスコアを $S(di)$, 文書 dk 中に現れる名詞の価値を $Va(Wk)$ とする以下のようなモデルで idf 値を算出し文書スコアを計算する．

$$D = \{d1, d2, d3, \ldots, dn\} \tag{1}$$

$$D = \begin{bmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{m,n} \end{bmatrix} \tag{2}$$

$$S(di) = 1\text{：初期条件} \tag{3}$$

$$Va(Wk) = \sum_{di \in D} (S(di) \times TF(d(i), w_{k,i})) \tag{4}$$

$$S(di) = \sum_{w \in di} Va(Wk) \tag{5}$$

この (3)〜(5) を繰り返し, 算出した文書スコアの高い方から 5 つの発言を抽出して内容説明とする．

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

3.1 比較検討手法

上の提案手法とは別に [2] の手法の改良版を作成した. [2] では Twitter の重要な特徴である RT に関しては特に触れていなかった. RT された発言や,RT した発言をそのままの状態で要約手法に適用すると RT された発言に重みが偏る傾向があった. また,RT された発言は要約に有用と言えるが,RT した発言は重要でないものが多い. 実行結果 2 では [2] の手法に RT を考慮するため,RT した発言を抜き取り処理を行った. また前述したように [2] の手法は時系列毎に発言をクラスタリングし, 各クラスタに代表発言を設ける. 実行結果 3 では RT した発言を抜き取る処理に加え, その代表発言を決める際に tf/idf をかけその値の大きいものを代表発言とした. また, そのままでは長い発言が重要な発言となる傾向があったが, パラメータを設定して発言長に左右されないようにした.

4 実験結果

今回の実験では 2010 年 11 月 28 日に開催されたトヨタのモータースポーツイベントに関するハッシュタグを用いた.

提案手法による説明文例

11 月 28 日 富士スピードウェイで開催する「TMSF2010」に今年も参加します。昨年好評だった、当日限定のスペシャルボディーを用意しております。
品川駅なう。ＪＡＦ表彰式は無事終了！関係者の皆様お疲れ様でした！お世話になったスタッフの皆さん、ありがとうございました！明日からは富士スピードウェイへ参りますっ！TMSF ですよー !!
TMSF で富士スピードウェイなう
２００９年の TMSF で初めてイベントに参加しました。スーパー GT 参戦チームの計らいで GT マシンと同じカラーリングの紙ボディーを用意。人気は、TOM's、WedsSports でした。

昨日の TMSF のイベントで用意したモックカーのスペシャルシート、一番早くになくなったのはトムス。その次に人気なのが、GT300 の WedsSports!! このチームは、昨年もなくなるのが早かった。

比較的イベント開催日以前の発言が目立つ. tf/idf 値が高い単語を含んだ発言だけでは, イベントの告知など内容とは関係のない発言も多く含まれてしまう. そのため的確にイベントの内容を抽出したとは言い難い.

文献 [2] の改良版 (1) による説明文例

今日は TMSF！ これから富士スピードウェイに向けて出発！
GT ドライバートークショーに行くか、コース上のエヴァを見るか、それが問題だ…
あはははははw QT @tatebou 自転車のイベントに行ってもモータースポーツのイベントに行っても、いつも片山右京さんがいるので、右京さんの追っかけ状態になっている…。
パドックに行けなかったので、先生・ピス兄・いっとちゃんのトークショーに来てみた
http://twitpic.com/3awqod 歴代のトヨタ F1 マシンとあたし。壮観!!
石浦選手、ナスカーとか…すげー喜んでそう
TMSF でソープボックスダービーコーナーに来てくれた方、本当にありがとうございました。午前中で予定数に達してしまったために、手に入らなかった人すみません。またスタッフに不手際等があったかと思いますが、楽しんでいただけたでしょうか？ また来年 TMSF であいましょう

TMSF 楽しかった！ちょっと遠いけど、行ってよかった。関係者の皆様、お疲れさまでした！そして、ありがとうございました !!

本研究の提案手法よりもイベント開催中の発言は多いが, イベントの内容が明解だとは言い難い.

文献 [2] の改良版 (2) による説明文例

TMSF 準備がすすんでます。
GT ドライバートークショーに行くか、コース上のエヴァを見るか、それが問題だ…
あはははははw QT@tatebou 自転車のイベントに行ってもモータースポーツのイベントに行っても、いつも片山右京さんがいるので、右京さんの追っかけ状態になっている…。
パドックに行けなかったので、先生・ピス兄・いっとちゃんのトークショーに来てみた
歴代トヨタ F1。こうして見ると、なんか悲しいなぁ(涙)
SGT スペシャルバトルはーじまーるよー！
TMSF でソープボックスダービーコーナーに来てくれた方、本当にありがとうございました。午前中で予定数に達してしまったために、手に入らなかった人すみません。またスタッフに不手際等があったかと思いますが、楽しんでいただけたでしょうか？ また来年 TMSF であいましょう

六本木到着なう。TMSF にご来場頂きました皆様ありがとうございました。また明日からは、六本木店宜しくお願いします！ これは [2] の手法に以下の処理を加えた結果である.

1. 代表発言を決める際,tf/idf をかけてその値が高い発言を代表発言にする.
2. パラメータを設定し, 文書長が長い発言が tf/idf 値が高くなることを緩和させる.
3. RT した発言を省く.

この結果を見ると,[2] の手法よりもイベントの具体的な内容を含んでいる.

5 おわりに

以上実験の結果より, 要約文としてイベントの内容を説明するには具体的内容についても適度に抽出できるよう考慮する必要がある. 今後はその具体的な手法についてさらに研究を進めていきたい.

謝辞 本研究は科研費 (21500102) の助成を受けたものである.

参考文献

[1] 松尾 豊, 大澤 幸生, 石塚 満, “ 電子掲示板における会話からのトピックの発見と要約 ”, The 16th Annual Conference of Japanese Society for Artificial Intelligence, 2002

[2] 高村 大地, 横野 光, 奥村 学, “ Summarizing microblog stream ”, 人工知能学会研究資料, 2010

[3] Beaux Sharifi, Mark-Anthony Hutton, Jugal Kalita, “ Summarizing Microblogs Automatically ”,University of Colorado at Colorado Springs

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.