F-003

SVM 手法を用いた Twitter トレンドのトピック追跡
Topic Tracking in Twitter Trends using Support Vector Machine

増田 真也† 石川 孝†
Shinya Masuda Takashi Ishikawa

1. はじめに

Twitter 上で現在注目を集めているキーワードを見つける手段としてトレンドトピック機能がある．この機能は，Twitter 上で最近の出現頻度が多いキーワードを選んで表示している．このトレンドトピックは，現在注目されている Twitter 上の複数のキーワードを表示するので，Twitter のユーザがそのキーワードによってトレンドトピックに関する情報収集を行うことを可能にする．しかし，トピックが意味的に同じでも表示されるキーワードが時間的に変化することがあるため，ユーザがトピック追跡を行うにはキーワードの対応付けが必要となる．

そこで本研究は，機械学習の SVM 手法 [1]を用いて，トレンドトピックに対するツイート集合から，ツイートに含まれるトレンドトピックを表わすキーワード以外の特徴語を学習して，トレンドトピックのキーワードの対応付けを構文解析のみで行うことを目的とする．一般的なトピックの定義は，形式的には“影響力のある出来事または活動，およびそれらと一緒に直接関係のある出来事と活動” [2]であるが，本研究におけるトピックの定義は，ソーシャルメディア上の会話において話のタネになるようなキーワードであるとする．

本論文は，2 章でトピック追跡の方法，3 章で評価実験，最後に 4 章で結論と今後の課題について記述する．

2. トピック追跡の方法

2.1 問題定義

本研究のトピック追跡は，時刻 t におけるトレンドトピック i_t に対するツイート集合 M_t^i を，時刻 $t-1$ での既知のトレンドトピックのクラス j_{t-1} （$1 \le i, j \le k$， k はクラス数）に分ける排他的多値分類問題として定式化する．ここで，時刻 t におけるトレンドトピック i_t はキーワード w_t^i で表され，そのトレンドトピックについてのツイート集合 M_t^i は，すべて w_t^i を含むとする．そして，時刻 $t-1$ のトレンドトピックを表わすキーワード $w_{t-1}^j \in W_{t-1}$ を分類のラベルとする．ここで，キーワード集合 $W_{t-1}=\{w_{t-1}^1, w_{t-1}^2, \cdots, w_{t-1}^k\}$ は，時刻 $t-1$ におけるすべての w_{t-1}^i を要素とする集合である．この分類問題の解 $w_t^i = w_{t-1}^j$ すなわち $i_t = j_{t-1}$ は，時刻が異なるトレンドトピック同士を対応付けする．このとき，ツイート集合 M_t^i の要素は w_t^i を含むツイート m_t であるため，ツイート集合 M_t^i を j_{t-1} で分類する問題は，次の解法に示すように，時刻 t のツイート m_t を j_{t-1} で分類する問題に帰着する．

† 日本工業大学 Nippon Institute of Technology

2.2 解法

上述の分類問題の解法は，つぎの特徴語抽出，分類の学習，ツイート集合の分類のステップからなる．

- 特徴語抽出は，ツイート m から特徴語の組 $X(m)$ を求める．
- 分類の学習は，ツイート集合 M_{t-1}^j （$1 \le j \le k$ ）に含まれる特徴語 x の重みを SVM 手法で求める．
- ツイート集合の分類は，時刻 t のツイート集合 M_t を時刻 $t-1$ におけるトレンドトピックのクラス j_{t-1} に排他的多値分類する．

なお，特徴語抽出には mecab [3]を，分類の学習とツイート集合の分類には SVM-Light [4]を用いる．

2.3 特徴語抽出

特徴語抽出は，ツイート m に対し，そのテキスト内に含まれる特徴語 x の出現有無[0,1]を要素とする特徴ベクトル $X(m)=\{x_1, x_2, \cdots, x_n\}$ （n は語数）を出力する [5]．抽出する特徴語は，単体で何らかの実体を表わす意味を持ち，文の主語となれる語（文法上の体言）とする．この定義に該当する語の具体例は，本研究では名詞，複合名詞（要素名詞も含む）とし，Twitter に特有な語として URL，スクリーンネーム，ハッシュタグを含める．ただし，ここでの名詞は IPA 辞書による mecab での出力を指し，名詞の詳細のうち，ナイ形容詞語幹・引用文字列・形容動詞語幹・接続詞的・接尾・代名詞・動詞非自立的・特殊・非自立に属するものは，上述の特徴語の定義に適さないと考えられるため特徴語抽出の対象外とする．

2.4 分類の学習

分類の学習は，ツイート $m_{t-1} \in M_{t-1}^j$ から求めた特徴ベクトル $X(m_{t-1})$ に対して，SVM 分類器 $f(X(m)) = a \cdot X(m) - b$ （a は重みベクトル，b はバイアス項）を出力する．このとき，特徴ベクトル $X(m_{t-1})$ に対する正例・負例のラベル付けは，j_{t-1} （$1 \le j \le k$ ）のうち任意の一つのクラスを正例とし，その正例以外の全クラスを負例とする．トレンドトピックのキーワードが複数クラスのツイートに含まれることがあるので，一つの $X(m_{t-1})$ が二つ以上のクラスについて正例となることがありうる．分類の学習においては，トレンドトピックを表すキーワードの影響を除くため，特徴ベクトル $X(m_{t-1})$ の要素には，時刻 $t-1$ におけるすべての $w_{t-1}^j \in W_{t-1}$ を除外する．

2.5 ツイート集合の分類

ツイート集合の分類は，時刻 t のツイート集合 M_t と複数個の SVM 分類器 $f(X(m))$ を用いて， M_t に対する時刻 $t-1$ でのトレンドトピックのクラス j_{t-1} を出力する．この処理は，まず M_t の要素であるツイート m_t の特徴ベクトル $X(m_t)$ に対する関数距離 $f_{j_{t-1}}(X(m_t)) = a \cdot X(m_t) - b$

を各クラスについて求め，one-versus-rest 法 [5]を用いて $X(m_t)$ をクラス j_{t-1} に排他的多値分類する．その後，各クラスについて分類された $X(m_t)$ の個数を求め，その個数が式(1)に示すしきい値 θ 以上であるクラスを集合 M_t の分類 $w_{t-1}(M_t^i) \in W_{t-1}$ とする．しきい値 θ は，集合を排他的多値分類する際に，分類のクラスが一つに定まる十分条件となっている．

$$\theta = \frac{1}{2}\left|M_t^i\right| \qquad (1)$$

one-versus-rest 法は，二値分類器である SVM 分類器を多値分類器へ拡張する手法であり，分類段階において，一つの特徴ベクトル $X(m_t)$ に対して一つのクラスラベルのみが許されているとして，関数距離 $f_{j_t}(X(m_t))$ が最も大きいクラス l_{t-1} （$l = \arg\max\{ f_{j_{t-1}}(X(m_t^i))\}$）に分類する．

3. 評価実験

3.1 目的

本研究においてトピック追跡における最も基本的なタスクは，トレンドトピックの対応付けである．したがって評価実験の目的は，本手法によってトレンドトピックの対応付けが可能であることを示すことである．ここで対応付けが可能となる条件は，ある時刻 t と $t-1$ の間においてトレンドトピックを表すキーワードが変化していない状態 $w_t^i = w_{t-1}^j$ とする．

3.2 方法

評価実験で使用するツイート集合に対するトレンドトピックのキーワード w とツイート集合 M は，Twitter が提供する各種の API によって取得する．まず，GET trends/place を用いて日本のトレンドトピックを表わすキーワード集合 W を取得する．その後，GET search/tweets を用いて各トレンドトピックを表すキーワード $w \in W$ をクエリとして，w に属する日本語で投稿された直近のツイートを要素とするツイート集合 M を取得する．データ収集の時間間隔は，訓練データのツイート集合 M_{t-1} と分類するツイート集合 M_t の重なりを考慮して 10 分とする．

本研究のトピック追跡は排他的多値分類であると定義したため，分類結果の評価指標は式(2)に示す正解率（*accuracy*）を用いる．評価事例 $W_{test} \in W_t \cap W_{t-1}$ は，時刻 t と $t-1$ の間で $w_t^i = w_{t-1}^j$ であるトレンドトピックの組を要素とする集合とする．正解事例 $W_{correct}$ は W_{test} のうちで正しく分類できたトレンドトピックの集合である．分類における正解は，分類問題の解 $i_t = j_{t-1}$ から w_t^i と w_{t-1}^j を比較し同じキーワードであった場合とする．

$$accuracy = \frac{\left|W_{correct}\right|}{\left|W_{test}\right|} \qquad (2)$$

実験で使用するデータは，2012 年 10 月 4 日 ~ 10 月 5 日において取得した時刻数 144 におけるトレンドトピックとツイートとした．

3.3 結果

実験で使用するデータの内からランダムに 10 個の時刻を選択して本手法によるトピック追跡を行った結果，評価事例 W_{test} は 40 件存在し，正解事例 $W_{correct}$ は 38 件であり，分類の正解率は 95%となった．

3.4 考察

分類が不正解となった原因を調べるため，ツイート集合 M_t の分類が正解および不正解となった場合の，それぞれの分類の学習結果の重みベクトル a の特徴量の分布のヒストグラムを比較した．分類が正解となった場合の重みベクトル a の特徴値の分布は，単峰であるが，単峰が存在する値域より大きい値域で外れ値が見られた．一方，分類が不正解となった場合は，正解となった場合と同様に単峰であるが，外れ値が見られないか頻度的に少ないことが見られた．したがって，これらの事実からトレンドトピックを表すキーワードを使わずにトピック追跡を行うためには，訓練データ内に，トレンドトピックを表すキーワードの他に共起性の高い特徴語が必要であることが推測される．さらに，重みベクトル a の要素を特徴量で降順にソートして観察したところ，上位となった要素に対応する特徴語は，本研究におけるトピックの定義に適するような語であることが確認された．

本手法のトピック追跡はツイート集合 M_t 排他的多値分類問題として定義したが，実際のツイートは排他的ではなくに複数のクラスに属することがあり，問題定義から外れてしまうことがある．このため，問題定義を実際のツイートの分類に即して行うならば，排他的という条件を除いた多値問題として再定義する必要がある．

4. まとめ

本論文の手法は，評価実験の結果から，少なくともキーワードが同じであるトレンドトピックを対応付けることが可能である．また，学習にトレンドトピックを表すキーワードを使わないという制約条件のもとで正しくトピック追跡するためには，訓練データ内にトレンドトピックを表すキーワードの他に共起性の高い特徴語が存在することが必要である．さらに，トレンドトピック i を表すキーワード w^i を除外した訓練データを用いた SVM 手法による特徴語の学習により，i に属するツイートを識別できるような，w^i 以外の語を発見することも可能である．

今後の課題は，同じトピックの対応付けだけでなく，トレンドトピックの消滅・発生・融合・分離を同定できるように問題を再定義して手法を改良することである．

参考文献

[1]T. Joachims, “Text categorization with support vector machines”, ECML (1998).
[2]J. Allan, editor. “Topic Detection and Tracking: Event-based Information Organization”, Kluwer, (2002).
[3]mecab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html .
[4]SVM-Light, http://svmlight.joachims.org/.
[5]平博順, 向内隆文, 春野雅彦, “Support Vector Machine によるテキスト分類”, 情報処理学会自然言語処理研究会 NL128-24, pp. 173-180, (1998).
[6]高村大也, 奥村学, “言語処理のための機械学習入門”, コロナ社, (2010).