Twitter からの天気と食べ物の関係性抽出

伊藤 拓　　深澤 佑介　　太田 順

東京大学大学院工学系研究科　NTT Docomo　東京大学大学院工学系研究科

ito@race.u-tokyo.ac.jp　*fukazawayu@nttdocomo.com*　*ota@race.u-tokyo.ac.jp*

概要　天気は，食べ物の嗜好に大きく影響を及ぼすと考えられる．我々は，Twitter 上で投稿される食べ物の種類と，天気との関係を調べた．先行研究により，天気の中でも特に Twitter の投稿内容と関係の深い気温と湿度を天気コンテキストとして採用した．また，気温や湿度は時期によって大きく変動する．時期の影響を考慮した天気―食べ物間の関係性抽出のためのモデル化を行い，その評価を行った．

キーワード　トピックモデル，コンテキストアウェア，天気，時期，ウェブマイニング

1 はじめに

ツイートを投稿する際の文書生成プロセスは，天気によって大きく影響を受けると考えられる．天気コンテキストと Twitter の文章との関係が抽出できれば，天気と関係の深いトレンドの抽出，商品の需要予測に応用することが期待される．天気は季節によって大きく変動するため，天気コンテキストを考慮する際には，時期コンテキストを考えることが重要である．

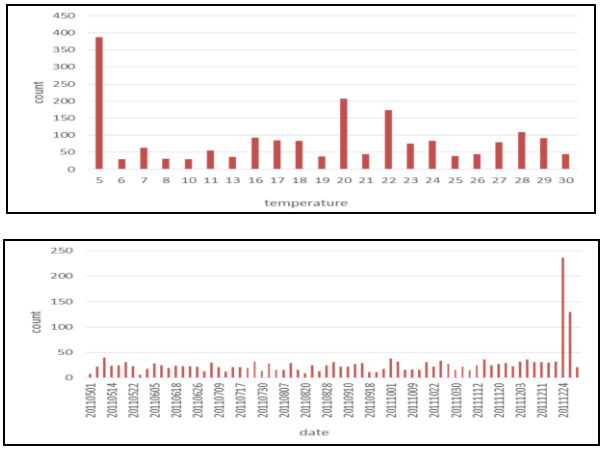

図 1「ケーキ」という単語を含むツイート数（上：気温，下：日）

図 1 は，Twitter で「ケーキ」という単語を含むツイートが投稿された数を表したものである．上図が気温ごとに，下図が日付ごとのヒストグラムを表している．気温のヒストグラムを見ると，5℃付近にピークが来ており，寒さとケーキとの間に深い関係があるように見えるが，日付のヒストグラムを見てみると，12 月 24 日付近にピークが来ており，「ケーキ」という語は気温ではなく，クリスマスという時期的な要因によって多く投稿されていたことが分かる．このように，時期コンテキストを考慮しないと，誤った結論に陥ってしまう可能性がある．

2 提案モデル

図 2 は，グラフィカルモデルを用いて，本研究で提案する文書生成過程を表したモデルである．このモデルは，トピックモデルとよばれる文書生成モデルの 1 つである LDA[1]に，天気クラスと時期クラスという 2 つのクラスを導入したものである．天気クラスとは，天気コンテキストに関するコンテキストクラスである．たとえば，あるツイート d について，そのツイートがなされた日の気温(a_{1d})と湿度(a_{2d})に基づいて，天気クラス(m_d)と紐付ける．そして，似たような気温(a_{1d})と湿度(a_{2d})のときに投稿されたツイートは，同じ天気クラス(m_d)に割り振られる．同じように，時期クラスは，近い日付(y_d)に投稿されたツイートは，同じ時期クラス(t_d)に割り振られる．

ユーザは，天気クラスに基づいて天気トピック(z_{di})を生成し，時期クラスに基づいて時期トピック(k_{di})を生成する．最終的に，天気のトピックか時期のトピックかどちらか一方を選択し，トピックに応じて単語(w_{di})を出力するという流れになっている．例えば，あるユーザが 8 月下旬の気温 33℃のときにツイートするとき，そのツイートは「気温 30℃付近のクラス」「8 月付近のクラス」に分類され，クラスに応じて，天気のトピックであれば「避暑トピック」「スポーツトピック」，時期のトピックであれば「甲子園トピック」「新学期トピック」などが選択される．そのうち 1 つのトピック（例えば「避暑トピック」）が選択され，トピックに基づいて「カキ氷」などの単語が出力される．ある単語が天気トピックから生成されるか，時期トピックから生成されるかは確率的に決まり，学習を重ねることである単語が天気に関係ある語なのか時期に関係ある語なのかが学習される．データを用いた学習には，Collapsed Gibbs Sampling[2]を用いている．

3 評価実験

今回，提案モデルの評価実験を行うにあたって，まずデータセットの構築を行った．実験の対象としたのは，2011 年 5 月~12 月に東京都で投稿された日本語のツイートであり，928051 件であった．これらのツイートについて，ツイートされた日付から，気象庁でダウンロードした天気データとツイートデータとの紐付けを行い，データセットを構築した．なお，語彙統語パターンを用いてツイートから食べ物関連の単語を抽出し，それを単語特徴量として用いる．

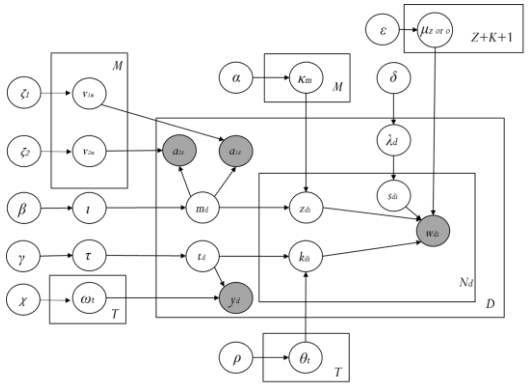

図 2 提案モデル

3.1 量的評価

量的評価として，提案モデルにおいて時期クラスを考慮しなかった場合のモデルとの文書予測精度の比較を行う．単語予測精度として Perplexity を用いる．Perplexity は，ある単語のモデルにおける尤度の対数を表したものであり，値が低いほどモデルの単語予測精度が高い．なお，このモデルにおいてはクラス数，トピック数がパラメータとなっており，Perplexity を比較することでチューニングを行った（クラス数 10，トピック数 80 を利用）．表 1 に結果を示す．天気のみを考慮したモデルよりも，時期を考慮した提案モデルの単語予測精度の方が高くなっていることが分かる．

表1 天気のみのモデルと提案モデルの Perplexity 比較

天気のみのモデル	提案モデル
4943.4	4663.0

3.2 質的評価

表 2 は，提案モデルによって学習された単語を表している．この表には，各天気クラスに割り振られたツイートの平均気温，湿度が載っている．また，各天気クラスで頻出したトピックと，そのトピックに対応している単語を載せている．例えば，天気クラス 2 の平均気温は 33.69℃，湿度は 46.63%であり，この天気クラスに割り振られたツイートを見てみると，「アイス」「納豆」「お茶」といった単語が多く含まれていることが分かる．

まず，気温の低いクラスであるクラス 3 の単語について考察していく．第 1 章で述べたように，「ケーキ」という単語は気温 5℃付近で頻度がピークとなっているため，時期コンテキストを考慮しない場合，「ケーキ」が気温の低いときの単語として学習され，クラス 3 の特徴語となっていたと考えられるが，提案モデルでは，時期依存の単語は時期トピックとして分離されるため，天気クラスの特徴語としては学習されなかったと考えられる．次に，気温の高いクラスであるクラス 2 の単語について考察する．もっとも多く登場するトピックの中で，さらに最も頻出の単語は「アイス」である．この「アイス」という単語について，図 1 と同じく気温と湿度のヒストグラムを図 3 に載せる．

表 2 提案モデルによって学習された，天気クラスごとのトピックと単語

天気クラス	2	3
気温平均	33.69±1.11℃	9.09±1.21℃
湿度平均	46.63±5.21%	44.98±9.84%
出力された語	アイス	カレー
	納豆	うどん
	お茶	クレープ

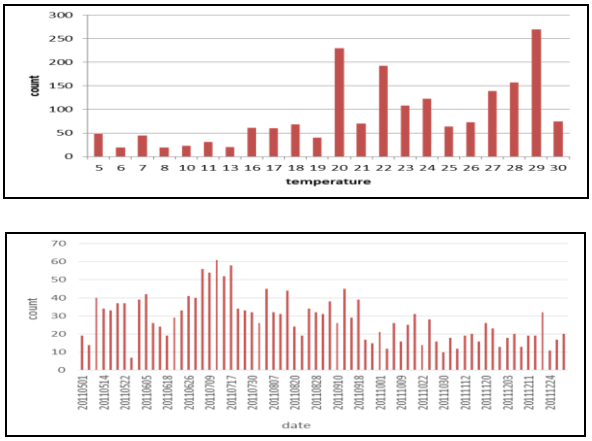

図 3 「アイス」という単語を含むツイート数（上：気温，下：日）

ここで，気温のヒストグラムを見てみると，ピークが 29℃付近と気温の高い時期にピークが来ており，気温の高いクラスの単語として学習されたのは妥当であると言える．一方，20℃付近でも登場頻度が高くなっている．「アイス」という単語が登場するツイートを見てみると，29℃付近でのアイスは「アイスクリーム」という意味でアイスを用いているのに対し，20℃付近では「アイスコーヒー」「アイスカフェオレ」という意味で用いられていた．今回天気クラス 2 で学習されたのは，前者の意味での「アイス」であると考えられる．このことからアイスクリームは夏だからという季節的な要因と比較し，気温による影響をより強くうけるということが分かる．

参考文献

[1] Blei, D., Jordan, A. Ng. M., and Lafferty, J.: Latent Dirichlet Allocation, JMLR, Vol. 3, pp. 993-1022, 2003.

[2] Griffiths, T. L and Steryvers, M.: Finding scientific topics, PNAS, 2004.