

マイクロブログの時系列情報を利用した 関連語発見手法に関する研究

和泉 諒†

西山 裕之†

† 東京理科大学理工学研究科

1 はじめに

Webの検索技術の進歩により検索精度は向上し、自分が調べたい事柄を検索語として入力するだけで様々な情報を得ることが可能となった。近年注目されているTwitter上でもユーザの気になった情報をキーワードとして検索することができるが、ユーザがいつでも検索目的に適した検索語(関連語)を思いつくとは限らない。そこで関連語を検索するためのWebサービスも存在する。本研究で発見する関連語とは、辞書的に定義される意味的な関係の単語ではなく、単語と関連性の高い単語のことである(同じイベント内で使用される単語同士など)。関連性の高い単語が発見できることにより、ユーザの情報検索支援を行うことができる。またこのような関連語は、時代とともに変化していくものである。過去では「アンドロイド」といえば、「ロボット」に密接な関係があったが、現在ではアンドロイド携帯などの登場により、「携帯電話」の方が、関係が深いといえる。そこで、このような関連語を発見するために本研究では、時系列情報を利用した関連語発見手法を述べる。情報源として、マイクロブログに代表されるTwitterを利用する。Twitterを利用することで、よりリアルタイム性の高い情報を得ることができ、さらにTwitterのアカウント情報を用いることでユーザの趣向がわかる。また、関連語を発見する際は、単語間の類似度を計算するが、その計算方法として、二つの単語のTwitter内ではつぶやかれている時系列情報に着目した。ある一定時間内につぶやかれた数の増減から二つの単語の相関を求め、その相関が高いものを関連語として発見する。時系列情報やユーザのアカウント情報を利用することで、ユーザに適したトレンド語や関連語を発見することを目的とする。本研究でのトレンド語とは、一定時間内につぶやかれた数が瞬間的に上昇した単語である。

2 ツレンド語・関連語の発見

現在、トレンド語を発見するためには、Googleトレンドというサービス[1]がある。ある単語の検索された回数を時系列順に並べ増減を調べることで、話題な単語を発見することができる。このように、時系列情報の増減を調べることで、トレンド語の抽出が可能となる。しかし、このサービスは、インターネットユーザ全体でのトレンド語を抽出することは可能だが、特定の分野を指定したトレンド語の抽出は困難である。個人によって、興味のある分野は違っており、それに合わせた

情報提供が必要となる。そこで、本研究では、トレンド語を発見するための情報源として、個人のアカウントのタイムライン上からトレンド語を発見する。タイムライン上から、トレンド語を抽出することで、個人の興味のある分野から発見することが可能となる。また、従来の関連語を発見する手法として、TF-IDF法を用いて類似度を計算し、類似度の高い記事の頻出単語をトレンド語や関連語として発見し、話題の変遷を抽出する研究[2]がある。本研究では、よりリアルタイム性を求めるため、情報源をブログではなく、マイクロブログを利用している。マイクロブログは、字数制限があり、平均Tweet文字数は30-40文字と言われている。従来の比較方法であるTF-IDF法では、文章の文字数が少なく、記事の比較にTFIDF値を利用するのは困難である。さらに、単語の係り受け関係を用いて、関連のある語の抽出を行う研究[3]がある。これも上記の理由により、適用できない場合がある。従って、本研究では、マイクロブログの特徴であるタイムラインという時系列情報を利用することで、関連語を発見できるシステムを実装した。

2.1 システム概要

本システムの流れは以下の通りである。

- タイムライン上からトレンド語候補の抽出
- ツレンド語の時系列情報の増減を調べる
- 増減が閾値を超えるトレンド語候補をトレンド語と登録
- ツレンド語から関連語候補の抽出
- 関連語候補とトレンド語の時系列情報の相関を解析
- 相関が閾値を超える関連語候補を関連語と登録

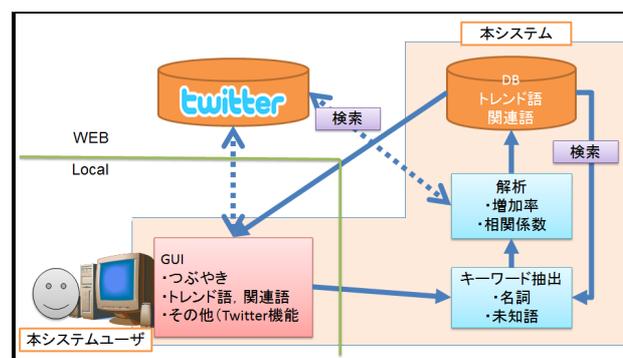


図1: システム構成図

Research on related word discovery technique using time series information on micro-blog.

Ryo Izumi†, Hiroyuki Nishiyama†

† Graduate School of Sci. and Tech, Tokyo University of Science

この流れを実現するためのシステム構成図を図1で示す。本システムは、GUIとキーワード抽出と解析とDBで成り立っている。GUIはつぶやき(タイムライン)などのTwitterの標準機能が備わっている。キーワード抽出では、文章から名詞や未知語を取り出す。解析では、増加率や相関係数を計算する。DBにはトレンド語や関連語を保存している。このようなシステム構成図の上で流れを説明する。

2.1.1 タイムライン上からトレンド語候補の抽出

トレンド語候補を抽出するために、個人のTwitterのタイムラインを利用する。タイムライン上でつぶやかれている、友人のつぶやき、著名人のつぶやき、ニュース情報を形態素解析し、未知語と名詞を抽出する。タイムラインから得られた単語は、個人の興味のある単語といえるので、その単語をトレンド語候補として登録する。

2.1.2 トレンド語の時系列情報の増減

Twitter内全体で単位時間あたりのトレンド語候補の単語の入ったつぶやき数を調べる。単位時間あたりにつぶやかれた数を増減は、次の式を利用する。

$$\text{単位時間 } t \text{ あたりにつぶやかれた数} : X_t \quad (1)$$

$$\text{勢いの増加量} : F_t = \frac{X_t - X_{t-1}}{\sqrt{X_t^2 + X_{t-1}^2}} \quad (2)$$

上記の式で、求めた値が閾値以上のものをトレンド語として登録する。

2.1.3 トレンド語から関連語候補の抽出

トレンド語をTwitter内全体で検索を行い、得られたつぶやきを形態素解析し、関連語候補を抽出する。全ての関連語候補の出現頻度を登録する。

2.1.4 関連語候補とトレンド語の時系列情報の相関を調べる

トレンド語と関連語候補の単位時間あたりにつぶやかれた数を比較する。比較する関連語候補は、頻度の高い単語とする。下記の式に示す。

$$\text{二つの単語の時系列情報} : (x, y) = \{(x_i, y_i)\} \quad (3)$$

二つの単語の相関係数 :

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

式の結果が閾値を超えるものを関連語として登録する。以上の流れにより、タイムライン上からトレンド語と関連語を抽出する。

2.2 GUI

本システムでは、ユーザがTwitter内のタイムラインを利用するためのGUIを用意した。図2に示す。ユーザは、タイムライン表示、つぶやき、フォローしているユーザの閲覧、検索が行える。

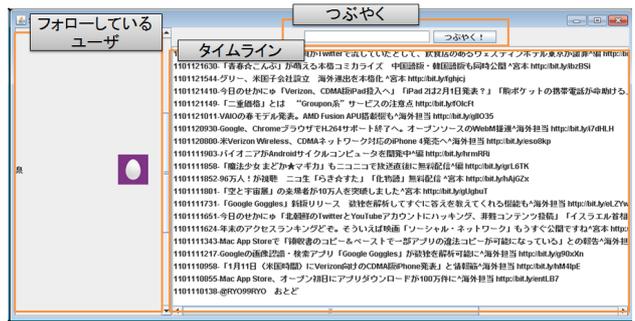


図2: GUI (タイムライン表示)

3 事例

本システムを利用して、スポーツの記事のみをフォローしたアカウントでのトレンド語と関連語を抽出した。抽出結果は1月13日におけるスポーツのトレンドを取り出したものである。抽出した結果を表示した画面を図3に示す。結果はいつでもスポーツに関係した

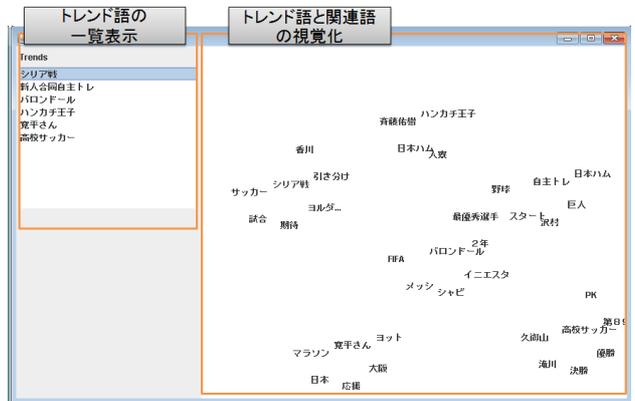


図3: GUI (トレンド表示)

ものとなっており、それぞれのトレンド語と関連語が抽出できた。GUIの画面では、左側にリストで一覧表示され、右側の部分では、トレンド語と関連語が一目でわかるような視覚化を行った。

4 おわりに

本研究では、マイクロブログの時系列情報を利用することで、関連語を発見するための手法を示した。また、本システムを利用することで、ユーザの興味のあるトレンド語や関連語を視覚化でき、簡単な操作で、関連語検索が行えることがわかる。今後の展望として、ブログの関連語検索と組み合わせた新たな関連語発見手法が提案できる。

参考文献

[1] “Googleトレンド”, <http://www.google.co.jp/trends>
 [2] 戸田 智子, 福田 直樹, 石川 博 “Blog記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出”, DEWS2007, A8-3, 2007.
 [3] 数原 良彦, 戸田 浩之, 櫻井 彰人 “ブログにおけるイベントマイニングのための適切なキーワード抽出”, DEWS2007, A2-6, 2007.