Twitter におけるつぶやきの関連性を考慮した改良相関ルール抽出による話題抽出

Topics Extraction Using Twitter Tweet Replies By Improved Association Rule

鈴木　啓太 †　新美　礼彦 †

システム情報科学部 情報アーキテクチャ学科

1　はじめに

インターネットの発展に伴ない情報発信メディアが多様化し、われわれが日々入手できる情報量は増大している。その中であらゆる情報源をチェックして最新の話題をチェックする事は困難であり、世間の関心を集めている情報を簡単に知りたい、注目されている話題をまとめて知りたいなどのニーズの高まりから世間で話題になっているトピックをキーワードで表現し、ユーザーに提示するサービスも生まれている [1]。このようなサービスで注目されている話題のキーワードを得ることはできるが、なぜ話題になっているのかやどういった意味で使われているのかということは分からず、それらを知るためにそのキーワードに関して自分で調べる必要がある。

そこで、本研究ではユーザーが指定したをトピックワード説明する単語もしくは文書を抽出し提示するシステムを提案する。解析するデータとしては Twitter[2] という Web サービスを対象として用いることにした。

図 1　Twitter のインタフェース

Twitter は、「いまどうしてる？」に対する解答を 140 文字の短いメッセージ（ツイート:Twitter に投稿するメッセージのこと。Twitter のデータの基本単位である。）という形でつぶやくサービスである。2009 年 6 月時点で、Twitter にアクセスしているユニークユーザー数は 1 億 1000 万人を超えており、急成長を遂げている Web サービスであると言える。リアルタイム性の高い Twitter サービスには多くのユーザが参加しており、Twitter 上で話題になっているワードから関連する類推ワードを抽出することで、日々生まれている新語・略語の用途について、理解を促すシステムを構築することが可能である。

Twitter のメッセージは 140 文字で投稿するという制限がかけられていることが大きな特徴であると言える。SMS のような感じで気軽につぶやける反面、ひとつのツイートの長さが短いため、解析する際うまく単語ベクトルを生成できないなど問題がある。そこで、本研究ではツイートのリプライ（あるツイートに対してなされる返信行為。リプライを投稿しあうことによって会話や議論を行っているユーザもいる。）や ReTweet（あるツイートを自分の発言として再投稿して、情報を拡散する行為。）の関係を見てツイートをまとめる事で本文の長さが少ない Twitter のデータに対して、リプライやリツイートからあるトピックワードを説明するワード（類推ワード）や関連する文書を自動生成することを可能とする手法を提案する。

2　関連研究

話題抽出の関連研究としては時間情報を含む文書集合から burst 分析を使って話題を抽出する藤木らの研究 [3] がある。この研究では時間情報を含む文書集合を定義し、その文書集合中のある文書とその次に来る文書の到着間隔を使って話題を抽出している。文書の到着間隔が短い状態を burst 状態と呼び、burst 状態にあるということは、それだけよく情報として発信されているということになる。つまりは話題になってい

る文書列であると言うことがいえる。

短いテキストの例では菊池らの研究 [4] の電子番組表 (EPG) を使った研究があげられる。これは時系列文書集合を話題ごとでクラスタリングし、各話題クラスタに属する文書集合から話題のキーワード群とキーワードの推移を表すグラフを生成してユーザーに提示する手法をとっている。

これらの 2 つの研究では、文書集合から話題を表すトピックワードを抽出する事を目的としている。話題を類推させるためには話題を表すトピックワードだけでは情報不足であるため、本研究ではトピックワードを説明する類推ワードと関連する文書を抽出する。

Twitter の研究としては、松村らの研究 [5] がある。この研究では Twitter のツイートデータから、盛り上がっている場所を抽出している。具体的には場所のキーワードを含むツイートを抽出し、単位時間あたりのツイート数を使って盛り上がっているかどうかを判断する手法を取っている。

本研究では、話題を表すトピックワーが理解可能なように類推ワードや関連ツイートを提示するシステムを提案する。

3 提案手法

ある話題を表す単語をトピックワードと定義し、またトピックワードを類推または説明するような単語を類推ワードと定義する。またトピックワードを説明するような文書を関連文書と定義する。

本研究の目的は、トピックワードを説明する類推ワードまたは関連文書を抽出して、ユーザーに提示しする事でトピックワードを類推させる手法を提案することである。提案手法では、話題類推情報抽出を行うことにより、この目的を達成可能なシステムの構築を目指す。本手法は文書集合一般に適用可能な手法であるが、Twitter データの特徴に合わせた対応も合わせて提案する。

3.1 話題類推情報抽出

提案する手法では、アプリオリアルゴリズムによる相関ルール抽出を用い、相関ルールの支持度と確信度を基準として、トピックワードに強い相関がある語を類推ワードとして抽出する。また、抽出した類推ワードを含む文章を文書集合から抽出し、それを関連文書としてユーザに提示する。

以下の手順で処理することによって類推ワードと関連文書を抽出した。

1. トピックワードを含む文書を収集する。
2. アプリオリアルゴリズムを用い、トピックワードが結論部となる相関ルールを抽出し、条件に当たるワード集合からなる候補語集合を作成する。
3. 得られた候補語集合の単語に対し、相関ルールの支持度と確信度をもとにスコアを付ける。
4. スコアが高い単語を類推ワードとして出力する。
5. 類推ワードを含む文書を文書集合から抽出し、関連文書として出力する。

3.2 Twitter データへの対応

Twitter のデータを使う場合、ツイートのテキスト長が短いため、うまく相関ルールが抽出できないという問題がある。これに対し、複数のツイートをまとめてひとつの文書集合として扱うことにより、テキスト長の短さ問題を解決する。あるツイートに対するリプライや ReTweet にはリプライ元のツイートに対する説明、やコメントなどが書かれていることが多い。このため、提案手法では複数のツイートをまとめる際に、リプライと ReTweet の関連に着目して、リプライや ReTweet 関係を持ったツイートをまとめる処理を行う。具体的にはリプライや ReTweet 関係を持つ前後数ツイートをひとつの文書として扱うことにした。これにより、ツイートの長さの問題を解決できるだけでなく、前後のツイートから話題の流れに沿った分析を行うことも可能となる。

4 実験

提案手法の有効性を検証するため、Yahoo!ニュースにあるニュース記事での性能をチェックした。実験で使用したデータセットは、Yahoo!ニュースの全トピックスからランダムに選んだ 20 トピックス分の記事データであり、それぞれのトピックに対して提案手法を用い、類推ワードと関連文書を抽出した。また、比較のために単純頻度が高いワードも抽出した。

4.1 高頻度語との比較

トピックに対して抽出した類推ワードと単純頻度を計算し、高頻度に出現するワードのどちらがよりトピッ

クワードを説明しているかを被験者 30 名に回答してもらった。その結果を表 1 に示す。表中の数字はそれぞれのトピックスのに対して、高頻度後の方が説明していると思った、提案手法による類推ワードの方が説明していると思ったかを回答した人数を表している。

表 1　高頻度語と類推ワードとの比較

	高頻度	類推ワード
エヴァンゲリヲン	13	17
学生の就職活動	23	7
Google	17	13
検索エンジン	14	16
WiMAX	14	16
ドラゴンクエスト	16	14
ファイナルファンタジー	23	7
スマートフォン	12	18
電子書籍	10	20
クラウドコンピューティング	16	14
合計	158	142

表の合計から高頻度語の方が、トピックを説明しているもしくは最新の話題を掴んでいるとの回答が多いことがわかる。トピックごとの結果を分析してみると、学生の就職活動やファイナルファンタジーなどトピックを指すワードが漠然としすぎている場合に高頻度語の方が良いと回答する傾向にあった。しかし、実験結果に対し、両側 5 パーセントでの t 検定を行ったところ、トピックをを説明しているもしくは最新の話題を掴んでいる単語に関しては高頻度語と提案手法で抽出した単語に対する回答の差は見られないことがわかった。

4.2　ランダム文書との比較

トピックの記事からランダムに抽出した文章と提案手法で抽出した文章のどちらがより説明しているか、もしくは最新の話題をつかんでいるかを比較してもらった。その結果を表 2 に示す。表中の数字はどちらがよく説明しているかを回答した人数である。

表から提案手法で抽出した文書の方が、話題を説明しているもしくは話題をつかんでいるという回答が多いことがわかる。実験結果を詳細に分析したところ、ダルビッシュ有のトピックだけランダムの方が良いという回答が多く、21 人という結果になった。

理由としては、ダルビッシュ有のトピックスから抽出した文書の中にほぼ同じ単語で構成された文書が複数存在していたからであると考えれる。例えば、下の 2 つの文章はほぼ同じ情報を持っている。

表 2　ランダムに抽出をした文章との比較

	ランダム	手法
Android	11	19
iPhone	5	25
Twitter	4	26
ネット犯罪	0	29
こんにゃくゼリー窒息死事故	8	18
イチロー	7	23
ダルビッシュ有	21	9
3Dテレビ	5	25
クーポンサイト	4	26
サッカー日本代表	1	29
合計	59	229

- 日本ハムのドラフト１位・斎藤佑樹投手（２２）＝早大＝が１５日、東京・江東区の東京ビッグサイトで行われた日本ハムグループ商品展示会に出席

- 日本ハムのドラフト１位･斎藤佑樹投手（２２）＝早大＝が１５日、東京ビッグサイトで行われた日本ハム本社の商品展示会に出席

しかし、現状のシステムでは、この二つは別の物として扱われる。そのため、ランダムに抽出した方が、より説明しているように見えたと考えられる。単に類推ワードを含む文を抽出するだけでは、関連文書として不十分であることがわかる。

また、上記の実験を踏まえ、類推ワードのみを提示した場合と、関連文書を提示した場合について、どちらが話題をとらえやすかったかアンケートしたところ、文書もしくは文書と単語の両方あった方が話題をとらえやすいという結果が結果となった。このことからも、単語のみではなく、文章も同時に出力することが重要であると言える。

4.3　Twitter を用いた実験

Twitter に対して、提案手法の適用を試みた。ハッシュタグでトピックを指定した 1500 件のツイートに対して手法を適用した。

結果、ツイートから抽出した場合でもほぼ同じ単語で構成された複数の文書が関連文書として抽出された。ツイッターの場合、実況など進捗を表すツイート（bot からの投稿をふくめて）や非公式 RT（「RT @user:引用文」の形で投稿するされる非公式の ReTweet) による情報拡散のためのツイートがこの結果に影響していると考えられる。

Yahoo!ニュースによる実験結果と合わせて、関連文書の抽出・提示の仕方を検討する必要があることがわかった。

5 おわりに

本研究では、Twitter から抽出した話題に対し、相関ルールを用いることにより、その話題（トピックスワード）を説明するようなワード（類推ワード）や文書を抽出し提示するシステムを提案した。

ニュース記事での実験では提案手法から抽出した単語や文書がトピックワードの説明、もしくは最近の話題を掴んでいるかどうかを判断してもらった。その結果、提案手法による類推ワードの提示は、高頻度を提示した場合と差がないことがわかった。提案手法による関連文書の提示では、ランダムに分を抽出する場合に比べて、話題を理解しやすいと言うことがわかった。加えて、単語と文書のどちらが話題をとらえやすいかをアンケートしたところ、文書もしくは文書と単語の両方会った方が話題をとらえやすいという結果が出ている。このことからも、単語のみではなく、文章も同時に出力することが重要であると言える。

現在、提案手法を Twitter に適用させる実験を行っている最中であり、分析結果により提案手法の有効性を検証したいと考えている。

参考文献

[1] **kizasi.jp:ブログから、話題を知る、きざしを見つける**, `http://kizasi.jp`, 最終アクセス日 2010 年 11 月 02 日.

[2] **Twitter**, `http://twitter.com/`, 最終アクセス日 2010 年 11 月 02 日.

[3] 藤木稔明, 南野明之, 鈴木泰裕, 奥村学.(2004) **document stream における burst の発見**. 情報処理学会研究報告. 自然言語処理研究会報告

[4] 菊池匡晃, 岡本昌之, 山崎智弘.(2008) **階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出**. 日本データベース学会論文誌. 第 7 巻

[5] 松村飛志, 安村通晃.(2008) **街に着目した Twitter メッセージの自動収集と分析システムの提案と試作**. 電子情報通信学会 WI2 研究会