3Q1-8in1

Twitterからの情報抽出による災害時の情報共有アプリケーションの開発

An information sharing application for disasters by extracting information from Twitter

鈴木 雄大 *1 Yudai Suzuki　小川 和晃 *2 Kazuaki Ogawa　中嶋 航大 *2 Kodai Nakajima　田村 哲嗣 *1 Satoshi Tamura　速水 悟 *1 Satoru Hayamizu

*1岐阜大学工学部電気電子・情報工学科
Department of Electrical, Electronic and Computer Engineering,Faculty of Engineering,Gifu University

*2岐阜大学工学研究科応用情報学専攻
Department of Information Science,Graduate School of Engineering,Gifu University

Twitter in recent years has attracted attention as a means of transmitting information at the time of a disaster. However, it is difficult to obtain only the necessary information out of enormous data. In this research, we tried to construct a system which can extract information efficiently at the time of a disaster by extracting keywords from tweets and classifying them. It displays the obtained information on the map. We also subjectively evaluated keyword extraction and category classification and examined the validity of the method. As a result of the evaluation, it was found that it is possible to extract important words from the tweet text by keyword extraction, and by category classification, it is possible to classify tweets about earthquakes into earthquake categories. Furthermore, by using the keywords and categories, we showed the possibility of collecting information more visually.

1. はじめに

災害時には，被災者にとって有用な情報を共有することが重要である．情報を共有する手段の一つとして，Twitter*1 を始めとする SNS が存在する．Twitter は 2016 年 9 月の時点で 4000 万人以上の利用者がおり，現在でも盛んにツイートが行われている．さらに Twitter にはツイートに GPS による位置情報を含める機能が存在しており，位置情報を活用することで，テレビやラジオでは難しい，ローカルな情報の共有が期待できる．しかし，Twitter へ投稿されるツイートは膨大な量であり，その中から有用なツイートを探すことは難しく，十分に情報を活かしきれているとは言い難い状態である．

[風間 12] は，東日本大震災発生前後のツイートより,「地震」と「原発」の 2 つの単語とその関連語間の時系列変化を分析することで，現実の事件により Twitter がどのような影響を受けるかの分析を行った．また，[坂巻 14] は，震災時に Twitter の情報把握の手間を省くため，Information Value や単純ベイズ分類器を用いて震災と関連するツイートのみを機械的に抽出する手法を提案した．さらに，[六瀬 13] は，震災発生時に Twitter から情報を収集しユーザーの状況に応じて適切な情報を提供するシステムの構築を目指しており，その内の情報収集・整理の部分について検討を行った．このように，災害時に Twitter の情報を活用するため，情報の調査・収集を試みる研究は盛んに行われている．また，現在公開されている Twitter の位置情報を活用した情報収集システムとして,「ちずツイ」*2 と呼ばれるサービスが存在する.「ちずツイ」は GoogleMap 上に位置情報付きツイートを表示できるサービスで，地震などの検索ワードを指定することにより，震災時にローカルな情報を得ることができる．しかし「ちずツイ」では，地図上には Twitter アカウントのアイコンしか表示されない．そのため地図を見るだけではツイートの内容がどのようなものか分からないという欠点がある．

そこで本論文では，情報共有の支援に向けた，GoogleMap 上にツイートを表示する Android アプリケーションの開発を行った．さらに，自然言語処理を用いてツイートのキーワード抽出とカテゴリ分類を行い，キーワードとカテゴリを GoogleMap 上に表示することを検討した．これにより,「ちずツイ」と比較して，キーワード等の視覚的な情報を用いたより効率のよい情報共有が期待できる．なお，本アプリケーションは災害一般を想定しているが，本稿では地震を例に開発を行った．

連絡先: 鈴木雄大，岐阜大学工学部電気電子・情報工学科，〒 501-1193 岐阜市柳戸 1 番 1，yudai@asr.info.gifu-u.ac.jp

∗1 https://twitter.com/

∗2 http://chizutwi.jp/

2. 提案手法

2.1 概要

本研究で提案するアプリケーションシステムの概要を図 1 に示す．具体的な流れを以下に述べる．

1. ツイートの収集
   Twitter Streaming API を用いてツイートを収集する．このとき，位置情報付きツイート,「地震」のキーワードをハッシュタグに含むツイート，ランダムに選ばれたツイートの 3 種類のツイートに着目して，収集する．なお，ツイートの文字数が 100 字以上かつ，日本語が含まれるツイートのみに限定することとする．

2. 不要なツイートの除去
   災害時に役立つとは考えにくい不要なツイートは，予め除去しておく．方法として，事前に不要なツイートに含まれるであろう単語を登録し，その単語が含まれるツイートは不要であると判断する．また，半角文字が一定の割合以上である場合，文字数に対して含まれる情報量は少ないと考え，不要ツイートであると判断する．

3. ツイートからのキーワード抽出
   各ツイートにおいて重要であると考えられる重要語を，キーワードとして抽出する．

4. ツイートからの地域名の抽出
位置情報付きツイートを使用するにあたり，位置情報付きツイート数が比較的少ないことが問題として挙げられる．そこで，位置情報付きツイート数を増やすために，ツイート内から地域名を抽出し，地域名から緯度経度の情報を取得することで，位置情報付きではないツイートでも位置情報を付与する．これにより，擬似的に位置情報付きツイートの数を増加させることを検討する．なお，最終的に地図上に表示を行うため，このときまでに位置情報が付与できなかったものは除くこととする．

5. ツイートのカテゴリ分類
ツイートの内容から，ツイートのカテゴリ分類を検討する．具体的には，ナイーブベイズ分類器を用いて，「地震などの危険情報を含むツイート」，「生活や家庭に関する情報を含むツイート」，「その他のツイート」の 3 種類に分類することとした．

6. ツイートの重要度の算出
ツイートにおける信頼度を表す指標として，以下の情報を用いる．

   - ツイートしたユーザーの「フォロワー数」
   - ツイートの「リツイート数」
   - ツイートの「いいね数」

   今回は実験的に，各値がその時点での全ツイートにおける平均値以上である場合は，信頼性が高いと仮定する．

7. アプリ内での投稿の処理
本システムでは，アプリ内における投稿も可能である．そこで，アプリ内で投稿された情報に対しても，同様の処理を行う．その際，アプリ内投稿にはフォロワー数などの情報が存在しないため，1. と 6. を除いた 2. から 5. までの処理を行う．

8. ツイート・アプリ内投稿を地図上に表示
一定時間の間収集し，各処理を施したツイート・アプリ内投稿を，吹き出し形式で地図上に表示する．その際，キーワード，カテゴリ情報，信頼度情報なども同時に表示する．

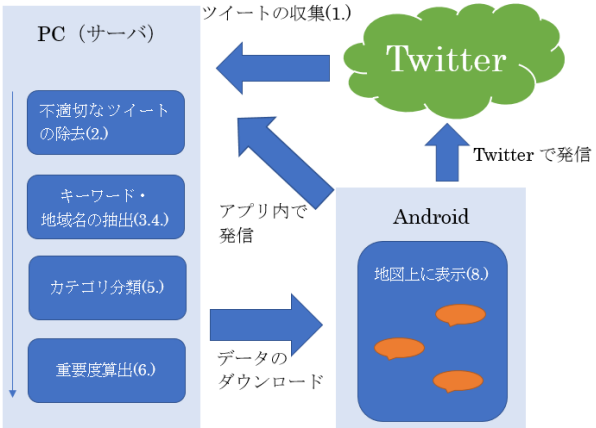

図 1: アプリケーションシステムの概要図

2.2 キーワード抽出

重要語の抽出には TF-IDF を用いる．TF-IDF とは，高頻度かつ一般性の低い単語は重要な単語であるという考えに基づき，単語に重みを付与することで，重要語を抽出するための手法である．TF は，各文書における単語の出現頻度である．本研究では，ツイートごとに名詞のみを抽出し，その頻度を TF として算出する．一方で，各単語が出現する文書の頻度を表すのが DF である．本研究では，収集したツイート群と毎日新聞コーパスを用いて DF を算出する．収集したツイート群とは，その時点までに収集したツイートをコーパスとして保存したものであり，少し前の話題に対応するために用いることとする．また，毎日新聞コーパスとは，毎日新聞東京・大阪本社の朝夕刊最終版を対象とした，毎日新聞 1991 年度以降の全文記事データ集のことである．社会性の高い話題に対応するために，本研究では毎日新聞「1 面」「2 面」「3 面」コードの記事から各 30000 記事ずつを抽出し，合計で 90000 記事の中から名詞のみの DF を算出する．以上，既存のツイート群を用いて得られた DF と，毎日新聞のコーパスを用いて得られた DF を合算し，その逆数をとった値を IDF として使用する．最終的に，TF と IDF を掛けた値を TF-IDF として，単語に対する重み付けを行う．そして，各ツイート内で最も高い TF-IDF の値をとる単語をキーワードとして抽出する．

2.3 カテゴリ分類

ツイートのカテゴリ分類には，テキスト分類の代表的な手法とされるナイーブベイズ分類器を使用する．ナイーブベイズ分類器とは，単語間の出現確率に独立性を仮定し，ベイズの定理を応用することで分類を行う手法である．まず，あるツイートを d，そのツイートが属するカテゴリを c としたとき，ツイート d がカテゴリ c に属する確率を $p(c|d)$ と表現する．この確率を用いて，ツイート d がどのカテゴリに属するかどうかを判断する．このとき，ベイズの定理を適用すると，$p(c|d)$ は次の式 1 で表される．

$$p(c|d) = \frac{p(c)p(d|c)}{p(d)} \propto p(c)p(d|c) \tag{1}$$

ここで，ツイート d に含まれる i 番目の単語を w_i としたとき，単語間の出現確率に独立性を仮定したとすると，確率 $p(d|c)$ は次の式 2 で表現することができる．

$$p(d|c) = p(w_1 \wedge w_2 \wedge ...w_i... \wedge w_K|c) = \prod_{i=1} p(w_i|c) \tag{2}$$

これより，元の式は次の式 3 に書き換わることとなる．

$$p(c|d) = p(c)\prod_{i=1} p(w_i|c) \tag{3}$$

続いて，確率 $p(w_i|c)$ を定式化することを考える．これは，カテゴリ c という条件の下で単語 w_i が出現する条件付き確率であり，カテゴリ c のツイート全体における単語 w_i の出現確率で表される．確率 $p(w_i|c)$ は式 4 で表現することができる．

$$p(w_i|c) = \frac{t(c, w_i)}{\sum_{w' \in V} t(c, w')} \tag{4}$$

しかし，分類対象において，学習データに含まれない単語が存在する場合，この確率はゼロとなってしまうゼロ頻度問題が考えられる．そこで，ゼロ頻度問題に対処するためにスムージングと呼ばれる手法が用いられており，本研究ではその中でも

よく用いられるラプラススムージングを使用する．ラプラススムージングは，各単語の出現頻度に 1 を加えることでスムージングを行う手法であり，式 5 で表現できる．

$$p(w_i|c) = \frac{t(c, w_i) + 1}{\sum_{w' \in V}(t(c, w') + 1)} \tag{5}$$

以上より，最終的には確率 $p(c|d)$ が最大となるカテゴリ c に分類することとなる．なお，$p(w_i|d)$ は非常に小さな値をとるため，アンダーフローを起こす可能性が存在する．よって，実際には対数をとった値を最大化することを考える．(式 6)

$$\hat{c} = \arg\max_{c} \log p(c|d) \tag{6}$$

2.4 アプリケーションの外観

図 2 の左側に，アプリケーションのメイン画面を示す．ツイートは，取得した位置情報を元に，吹き出し状のアイコンとして GoogleMap 上に配置される．このとき，吹き出しに表示されている文字が抽出されたキーワード，吹き出しの形が分類されたカテゴリ，吹き出しの薄さがツイートの行われた時間を示す．吹き出しの薄さは，ツイートが行われてから時間が経過するほどより薄くなっていくため，新しいツイートが強調されて表示されることとなる．また，左上のボタンを操作することで，古いツイートをどの範囲まで表示させるかを選択することができる．さらに，右下のボタンで表示させるツイートのカテゴリを絞り込むことができる．

図 2 の右側に，吹き出しをタップした際の画面を示す．吹き出しをタップすることで，そのツイートの詳細情報が表示される．吹き出しに表示されていたキーワードは，ツイートの本文中で赤色に強調して表示される．左上に表示されている数字はそれぞれツイートのいいね数，つぶやいたユーザーのフォロワー数，ツイートのリツイート数を表している．また，各値がその時点での平均値を超えている場合は，信頼性が高いとし，強調して表示される．右上の円グラフは，カテゴリ分類の際の，各カテゴリにおける分類の信頼度を割合で表示している．

図 2: アプリケーションのメイン画面 (左) とツイートを表示した画面 (右)

2.5 アプリ内投稿機能

本研究で提案するアプリケーションは，情報を受け取る機能だけでなく発信する機能も持つ．とりわけ，Twitter を用いて位置情報付きツイートを発信する際の個人情報を考慮した結果，アプリ内のみで位置情報が共有される投稿を検討した．このアプリ内投稿機能により発信された情報は，本アプリケーションの，災害時情報共有という意図に沿った情報である可能性が高いと考えられる．

3. 実験

本研究における手法の妥当性を検証する実験を行った．

3.1 実験条件

キーワード抽出とカテゴリ分類の妥当性を検証するため，2016 年 10 月 21 日 14 時 7 分頃に発生した鳥取県中部地震の直後のツイートから収集した 213 ツイートを用いて，提案手法の検証を行った．その中から例としてツイートを 2 つ抜き出し，提案手法を適用した結果を示すこととする．抜き出したツイートを以下に示す．

ツイート 1 :『【高速道路通行止め】 日本道路交通情報センターによりますと、この地震の影響で、午後２時２０分現在、中国自動車道は、岡山県内の落合ＩＣと新見ＩＣの間で、またそれに接続する岡山自動車道の北房ジャンクションと有漢ＩＣの間でいずれも安全点検のため通行止めになっています』
ツイート 2 :『＜地震＞鳥取で震度６弱＝午後２時７分（毎日新聞） - Yahoo!ニュース https://t.co/hHfxlt3xj6 # Yahooニュース この影響で新幹線が停電している。新大阪まであと一歩なので焦る』

また，カテゴリ分類の実験では，ナイーブベイズ分類器の学習データとして以下のものを使用した．

- 「地震に関する危険情報のカテゴリ」: 毎日新聞「1 面」コードから,「地震」の単語を含む記事 1000 件
- 「生活・家庭に関する情報のカテゴリ」: 毎日新聞「家庭」コードから 1000 件
- 「その他の情報のカテゴリ」: 毎日新聞「総合」コードから 1000 件

3.2 キーワード抽出

3.2.1 概要：キーワード抽出

ツイート 1，ツイート 2 のそれぞれに対し，キーワード抽出を行った結果を表 1，表 2 に示す．なお，DF の算出に関して，コーパス 1 は収集したツイート 213 件のみを用い，コーパス 2 は毎日新聞コーパスと，本実験時点までに蓄積された 711701 ツイートを用いた．また，アプリケーションの吹き出しアイコンには，ツイート内の名詞における TF-IDF の値が最も高かったキーワードのみが表示される．

表 1: ツイート 1 に対して提案手法を用い抽出されたキーワード

重要度の順位	抽出されたキーワード	
	コーパス 1	コーパス 2
1	自動車	通行止め
2	通行止め	自動車
3	道路	道路

表 2: ツイート 2 に対して提案手法を用い抽出されたキーワード

重要度の順位	抽出されたキーワード	
	コーパス 1	コーパス 2
1	毎日新聞	新大阪
2	新幹線	停電
3	新大阪	毎日新聞

3.2.2 考察：キーワード抽出

表 1 より，コーパス 2 を用いた場合,「通行止め」という単語が最も重要であると判断された．このツイートの内容は，交通情報センターによる通行止め情報について述べているツイートであるため，コーパス 2 による抽出ではツイート本文から重要な単語を抜き出すことができたと推測される．また表 2 を見ると，コーパス 2 では,「新大阪」という単語が重要であると判断された．ツイート 2 は新大阪付近で新幹線が停電したことを述べているツイートであるため，コーパス 2 を用いることで，重要語を抜き出すことができたと考えられる．以上より，提案手法を用いることでツイート本文から重要であると考えられる単語の抽出が可能であることを示した．一方で，毎日新聞コーパスと過去のツイート群であるコーパス 2 を DF の算出に使用することで，最重要語がツイート 1 では「自動車」から「通行止め」に，ツイート 2 では「毎日新聞」から「新大阪」となることが確認された．

3.3 カテゴリ分類

3.3.1 概要：カテゴリ分類

カテゴリ分類の精度を評価するため，closed 条件における正解率の算出を行った．本研究では，分類対象のデータ全てに対してカテゴリ分類を行い，そのうち正解データと分類結果が一致した数を，分類に用いた全データ数で割った値を正解率とした．本実験では分類対象のデータとして，学習に用いた毎日新聞の記事 3000 件を使用し，そのうち 3 カテゴリに分類した結果から正解率を算出した．その結果，closed 条件における正解率は 0.933 であった．

一方で，今回収集した鳥取県中部地震直後の 213 ツイートを用いて，実験的にツイートに対するカテゴリ分類を評価した．

3.3.2 考察：カテゴリ分類

closed 条件における正解率が 0.933 であったことから，高い正解率であることが確認できる．しかし，closed 条件ではツイートではなく学習に用いた毎日新聞コーパスの分類を行っているため，ツイートに対しても高い正解率を出すことができるとは限らない．

そこで，実験に用いた 213 ツイートに対しカテゴリ分類を適用した．なお，213 ツイートは全て「地震」の単語を含むツイートであった．結果として 212 ツイートが「地震」のカテゴリに分類され，1 ツイートのみ「生活や家庭」のカテゴリに分類された．このツイートを見ると，地震の記述とラーメンの記述が混在しており，危険情報を含むツイートではなかった．これより，地震のカテゴリに関しては実際のツイートに対しても分類を行うことができると推測される．

4. 結論

本研究では，ツイートに対してキーワード抽出とカテゴリ分類を行い，その結果を含めて情報を地図上に表示させることで，被災時に効率的な情報共有を支援するアプリケーションシステムの構築を試みた．また，キーワード抽出とカテゴリ分類に関しては主観的な評価を行い，手法の妥当性を検証した．その結果，キーワード抽出においては，ツイート本文から重要であると考えられる単語の抽出が可能であることを示した．また，カテゴリ分類においては，分類器として用いたナイーブベイズ分類器が正しく学習されたこと，さらに実際のツイートに対して，地震発生時の地震に関する危険情報を含むツイートを，正しいカテゴリに分類できたことを確認した．以上のキーワード抽出とカテゴリ分類の結果を本システムに用いることで，より視覚的かつ効率の良い情報共有支援の可能性を示した．

5. 今後の課題

本研究では，ツイートの可視化を行い，効率の良い情報共有を支援するシステムの開発を目指した．しかし，開発したアプリケーションの客観的な評価を十分に行うことはできていないといえる．よって，実際にアプリケーションを使用したユーザーに対し調査を行う必要があると考えられる．また，カテゴリ分類では「地震に関する危険情報を含むツイート」「生活や家庭に関する情報を含むツイート」「その他のツイート」の 3 カテゴリへと分類を行ったが，その際「生活や家庭」「その他」に関するカテゴリの定義が曖昧であるという問題が存在する．これにより，この 2 つのカテゴリには意図していたツイートが集まっているとは言い難い結果となった．よって，カテゴリの再検討やコーパスの調整，最適化などが必要であると考えられる．本稿では地震を例に開発を行ったが，他の災害に関しても検証を行っていきたい．

参考文献

[Twitter] https://twitter.com/

[ちずツイ] http://chizutwi.jp/

[風間 12] 風間一洋, 鳥海不二夫, 榊剛史, 篠田孝祐, 栗原聡, 野田五十樹 (2012).“東日本大震災時の Twitter データを用いた単語間の関係の時系列変化の分析”, 人工知能学会全国大会, IC3-OS-12-2.

[坂巻 14] 坂巻英一, 亀井悦子 (2014),”Twitter 上のツイートに関するテキストマイニングの事例研究 -Twitter 上のツイートに関するテキストマイニングの事例研究-”, 日本経営工学会論文誌 Vol. 65 No.1,pp.39-50

[六瀬 13] 六瀬聡宏, 長島俊, 内田理, 鳥海不二夫 (2013)”Twitter を用いた大規模災害時における情報提供システム”, 第 12 回情報科学技術フォーラム,O-055,pp.651-652

[山本 12] 山本修平, 佐藤哲司 (2012)”Twitter からの実生活情報の抽出法の提案” 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2012) F3-4

本論文は NEXT COMMUNICATION FORUM2016 に応募した作品をまとめたものである．