O-010

大規模災害時の情報提供を目的としたツイート分類手法
Tweets Classification Method to Provide Useful Information at the Time of Large Scale Disaster

六瀬 聡宏† 内田 理†† 鳥海 不二夫‡
Toshihiro Rokuse Osamu Uchida Fujio Toriumi

1. はじめに

大規模災害が発生した際，被害を最小限に食い止めるには，災害発生後の迅速かつ的確な情報収集・伝達が重要である．例えば，東日本大震災が発生した際には，速報性の高い情報の受発信が行われるという特徴を有する Twitter が多数の被災者に利用されたことが判明しており[1]-[3]，Twitter 利用者の約 8 割が情報収集に役立ったとの調査結果も報告されている[4]．また，災害発生時に Twitter を利活用する試みも既に多くの事例が知られている[5]-[7]．例えば，Micro Mappers[5]というプロジェクトでは，視覚的にわかりやすい災害情報地図を作成する手段の一つとして Twitter が利用された．しかし，Twitter のタイムラインを通して流通する情報は膨大であり，また重複する情報やノイズが多数存在しているため，被災者が自分の状況の適した情報を的確，かつ簡便に得ることは容易ではない．例えば，東日本大震災が発生した 2011 年 3 月 11 日には，約 3,300 万件のツイートが投稿された事が判明している[8]．そのような背景から，我々は大規模災害時に Twitter から情報を収集・整理し，ユーザの属性や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いる[9]-[11]．現在は，ツイートを情報提供に適したカテゴリに分類し，地図上にマッピングするシステムを構築中である．本稿では，災害関連ツイートのカテゴリ分類について報告する．

2. 災害関連ツイートのカテゴリ分類

2.1 東日本大震災時のツイートの分類

本研究では，東日本大震災発生直後のツイート（2011 年 3 月 11 日 14 時 46 分 49 秒～同日 23 時 59 分 59 秒）を対象としてツイートのカテゴリ分類を試みる（我々の研究グループは，2011 年 3 月 7 日 0 時から 3 月 23 日 24 時までの期間に収集された約 4 億の日本語ツイートを保有している．これは，TwitterAPI の制限などから，該当期間の日本語ツイートの 8 割程度を収集したものである）．本研究では，「津波に関する情報」「避難（避難所・避難施設）に関する情報」「ライフライン（電気・ガス・水道など）に関する情報」「交通機関（鉄道・バス・飛行機）の運行状況に関する情報」「道路の状況」の合計 5 カテゴリへの分類を試みる．今回は，代表的な機械学習法である SVM (Support Vector Machine)[12]，Random Forest[13]，及び Naive Bayes[14]で分類器を作成し，分類結果の F 値により性能評価を行う．

2.2 実験用データセットの作成

実験用のデータセットを作成するため，地震発生時に災害状況をツイートする際に多く用いられると予想されるキーワード（「地震」「津波」「避難」「停電」「運休」「通行止」など）によるフィルタリングでツイートを絞り込んだ．また，情報提供に有益なツイートであることを前提としているため，地震関連のハッシュタグ（#jisin, #jishin, #jishin_jp, #earthquake, #eqjp, #311jisin, #saigai のいずれか）を含む 50 文字以上のツイートのみを抽出し，先頭が RT，もしくは QT で始まるものは削除した．さらに，地図上にツイートをマッピングすることを目的としているため，国土数値情報ダウンロードサービス[15]を利用して作成した地名・施設名リスト上に存在する名称を含むツイートのみを抽出した．最後に重複ツイートや不適切なツイートを手動で削除して残された 1655 ツイートをデータセットとして利用することとした．

次にこの 1655 ツイートに著者 1 名が手動で正解カテゴリのラベルを付与した．各カテゴリのツイートの例を表 1 に示す．なお，今回は，1 つのツイートに対し複数のラベルを付与することとした．例えば，

・停電。東京都 12 万件、神奈川 130 万、千葉 35 万、埼玉 36 万件との事。ＴＯＫＹＯＦＭから。東京都晴海、16 時 30 分頃津波到達予想。仙台は停電中。

表 1 各カテゴリのツイート例

津波	・14:46 三陸沖で大きな地震発生。最大震度 7（宮城県栗原市）。震源深さ 10km、M7.9。大津波警報範囲拡大。北海道太平洋沿岸から千葉県九十九里・外房までの太平洋沿岸に大津波警報。津波、余震に最大級警戒を！ ・大阪府にも津波警報が出ています！津波到着予想は 17 時 10 分 0.5m 海などの近隣の方は気を付けください ・鹿児島県東部、西部にも津波注意報発令！
避難	・まだ余震が続いています。。江東区大島 1 丁目、2 丁目の方は猿江恩賜公園が避難場所になっています。 ・東京都心部の無料開放避難場所。【池袋】立教大学【御茶ノ水】明治大学リバティタワー 【新橋】新橋第一ホテル 【品川】品川プリンスホテル 【新宿】高島屋タイムズスクウェア【渋谷】青山学院大学 【上野】東京文化会館 ・北区滝野川第七小学校開放してました。田端と駒込の間を歩いてる方近いです
ライフライン	・NHK 発： 東京電力 相模原、川崎でも停電している地域がある模様。 復旧のめどたたず ・《NHK 実況》福島県内約 30 万世帯 停電。仙台市全域で都市ガス供給停止。 ・新庄市は停電が続いているようです。戸沢村のライフラインは大丈夫です。
交通機関	・東海道新幹線上りはとりあえず名古屋まで到着。ここでしばらく停車するとのこと。東京静岡間の停電は復旧したが、被害状況確認のため ・阪急バスサイトより。大阪一新宿・池袋・千葉線が大阪発、新宿・池袋・千葉発とも運休決定。 ・東京メトロ銀座線、渋谷浅草間全線、半蔵門線九段下、押上間で運転再開。渋谷九段下間は引き続き見合わせ
道路	・関越道の新潟から東京方面。高速通行止めで水上から下道です。ホテルもないし、渋滞がすごいので泊まれるとこで泊まることを推奨。 ・甲州街道 20 号渋滞なう。 徒歩帰宅者多数 ・拡散希望 埼玉県 幸手情報 市内にある踏切は全て遮断機が降りて通過出来ません 4 号線などに迂回して下さい

† 東海大学大学院工学研究科情報理工学専攻
†† 東海大学情報理工学部情報科学科
‡ 東京大学大学院工学系研究科システム創成学専攻

表 2　各データセットのカテゴリラベル数

	dataset1	dataset2	dataset3	合計
津波	213	213	212	638
避難	110	108	109	327
ライフライン	63	64	63	190
交通機関	167	168	168	503
道路	41	41	41	123

表 3　各データセットの素性数

dataset1	dataset2	dataset3
2363	2379	2385

というツイートには津波に関する情報とライフラインに関する情報の両方が含まれるため，「津波」と「ライフライン」の 2 つに正解ラベルを付与した．

最後に，データセットを 3 つ（dataset1: 552 ツイート, dataset2: 552 ツイート, dataset3: 551 ツイート）に分割した．分割の際，データセットごとにカテゴリに偏りが極力生じないように考慮した．データセットごとのカテゴリラベル数を表 2 に示す．

2.3　機械学習を用いたカテゴリ分類器の生成

本研究では，機械学習（SVM，Random Forest，及び Naive Bayes）により各カテゴリへの分類器を生成する．今回の分類は複数ラベルテキスト分類に該当するため，各カテゴリへの分類を行う 5 つの分類器を生成する．分類に用いる素性は Mecab[16]により抽出された名詞（ただし，代名詞，数詞，アルファベットのみの名詞，及び非自立名詞は除く）とし，素性の値は出現したか否かの二値とした．各データセットの素性数を表 3 に示す．

上述した素性を用いて分類器を生成し，テストデータを分類した結果（F 値：適合率と再現率の調和平均）を表 4 に示す（Random Forest は学習ごとに生成される分類器が異なるため，3 回試行した結果の平均値である）．今回，SVM, Random Forest, Naive Bayes は python の scikit-learn ライブラリ[17]を利用して実装した．SVM は線形カーネル（ペナルティパラメータ C=10）と RBF カーネル（ペナルティパラメータ C=1000, カーネル関数 $\exp(-\gamma\|x-y\|^2)$ の $\gamma=0.001$）の 2 パターンで学習した．Random Forest の特徴選択基準には Gini 係数を利用した．また，Naive Bayes は Bernoulli モデルで学習した．表 4 からわかる通り，全てのカテゴリにおいて実用上十分と思われる分類精度を得ることができた．また，全てのカテゴリにおいて SVM が最も高い分類精度を有している．

2.4　素性選択（特徴ベクトルの次元削減）

素性の削減（特徴ベクトルの次元削減）による分類精度の変化について考察する．今回は χ^2 値を素性の選択指標として利用する．カテゴリ c を対象としているとき k 番目の素性 t_k の χ^2 値 $\chi^2(t_k,c)$ を式(1)により算出する[18].

$$\chi^2(t_k,c)=\frac{|T|\left(p(t_k,c)\cdot p(\bar{t}_k,\bar{c})-p(t_k,\bar{c})\cdot p(\bar{t}_k,c)\right)^2}{p(t_k)\cdot p(\bar{t}_k)\cdot p(c)\cdot p(\bar{c})} \quad (1)$$

ここで $|T|$ は訓練データの全素性数である．訓練データが dataset1, 2 の場合の素性を $\chi^2(t_k,c)$ の降順でソートした結果（上位 10 素性）を表 5 に示す．表 5 より，カテゴリ特有の素性が上位となっていることがわかる．また，上位 n 件のみを利用して生成した分類器による F 値の例を図 1~5 に示す（縦軸は F 値，横軸は素性数 n を表す．横軸は対数目盛りである）．図 1~5 では，n = 10, 20, 30, 40,50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 及び全素性を利用した場合をプロットしている．図 1~5 からわかるように，多くの場合において素性選択により分類精度（F 値）が向上している．

表 4　各カテゴリの分類性能（F 値）

	訓練用 dataset	1, 2	1, 3	2, 3	平均
	テスト用 dataset	3	2	1	
津波	SVM（線形）	**0.97852**	**0.98812**	**0.98585**	**0.98416**
	SVM（RBF）	**0.97852**	**0.98812**	**0.98585**	**0.98416**
	Random Forest	0.96571	0.97434	0.96667	0.96890
	Naive Bayes	0.95833	0.95833	0.98148	0.96605
避難	SVM（線形）	**0.96296**	**0.97222**	**0.96804**	**0.96774**
	SVM（RBF）	**0.96296**	**0.97222**	**0.96804**	**0.96774**
	Random Forest	0.93854	0.95188	0.94198	0.94413
	Naive Bayes	0.91000	0.89796	0.90099	0.90298
ライフライン	SVM（線形）	**0.96774**	**0.96970**	**0.96875**	**0.96873**
	SVM（RBF）	**0.96774**	**0.96970**	**0.96875**	**0.96873**
	Random Forest	0.84726	0.90077	0.92713	0.89172
	Naive Bayes	0.41463	0.45783	0.58696	0.48647
交通機関	SVM（線形）	0.93093	**0.95385**	0.95122	0.94533
	SVM（RBF）	**0.93413**	0.95062	**0.96073**	**0.94849**
	Random Forest	0.93073	0.92583	0.94080	0.93245
	Naive Bayes	0.93168	0.90735	0.94375	0.92759
道路	SVM（線形）	**0.88889**	**0.87500**	**0.87671**	**0.88020**
	SVM（RBF）	**0.88889**	**0.87500**	**0.87671**	**0.88020**
	Random Forest	0.64921	0.74478	0.82037	0.73812
	Naive Bayes	0.09302	0.09302	0.04762	0.07789

表 5　χ^2 値によりソートされた素性

順位	津波	避難	ライフライン	交通機関	道路
1	津波	開放	停電	運転	渋滞
2	警報	場所	全域	線	通行止め
3	沿岸	困難	ガス	再開	道路
4	県	避難	東京電力	メトロ	国道
5	運転	帰宅	電気	全線	通行
6	東京	者	世帯	銀座	道
7	宮城	高校	信号	半蔵門線	信号
8	線	都立	神奈川	押上	高速
9	再開	品川	停止	九段下	号線
10	高台	都内	山形	運行	車道

2.5　素性の追加による分類精度の向上

これまではツイートの形態素（名詞）のみを素性として利用してきたが，形態素以外の素性を利用することにより分類器の精度向上が図れると考えられる．今回は，交通機関カテゴリについて素性の追加を検討した．交通機関の運行状況に関連するツイートの例を以下に示す．

・東京メトロ銀座線は渋谷駅の混雑のために運転を全線で見合わせています。都営大江戸線は全線運行しています。

この例からもわかる通り，交通機関カテゴリに分類されるツイートには，鉄道路線名や駅名，バス運行会社名，空港名などが頻出するが，形態素解析器でこれらの名称が適切に抽出できないケースも多く，また頻度が少ない名称は学習に効果的でない可能性がある．そこで，ツイ

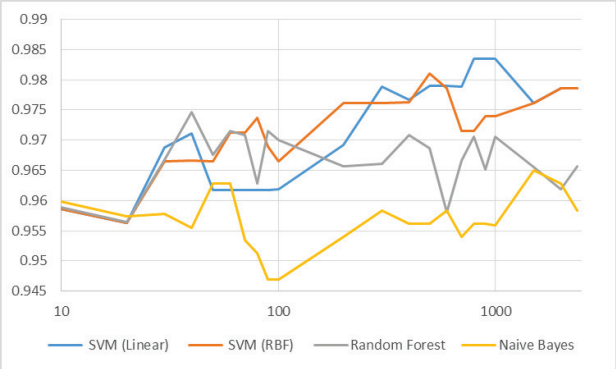

図 1 素性選択による F 値の変化（津波）
（訓練データ: dataset1, 2, テストデータ: dataset3）

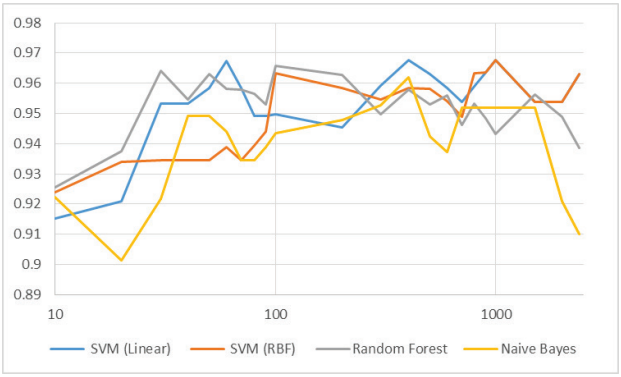

図 2　素性選択による F 値の変化（避難）
（訓練データ: dataset1, 2, テストデータ: dataset3）

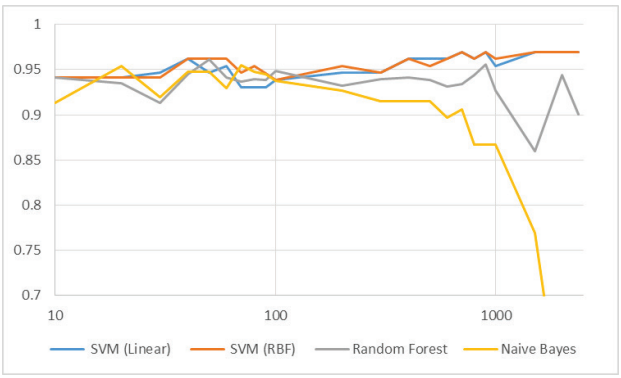

図 3　素性選択による F 値の変化（ライフライン）
（訓練データ: dataset1, 2, テストデータ: dataset3）

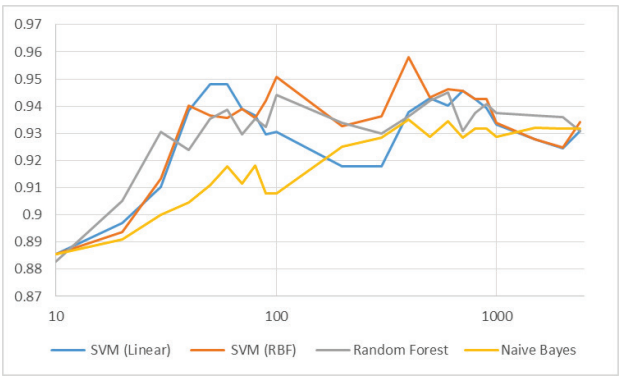

図 4　素性選択による F 値の変化（交通機関）
（訓練データ: dataset1, 2, テストデータ: dataset3）

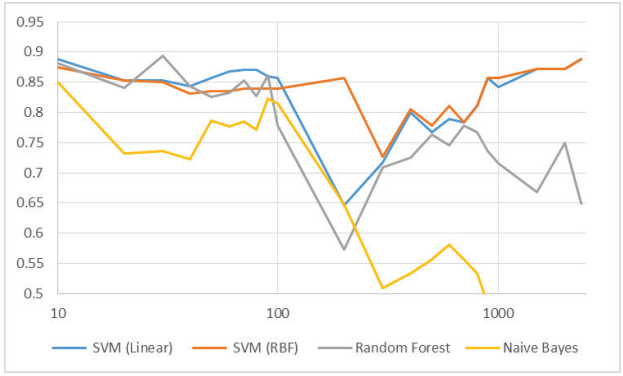

図 5　素性選択による F 値の変化（道路）
（訓練データ: dataset1, 2, テストデータ: dataset3）

ートに鉄道路線名，及び駅名が含まれるか否かを表す素性を 2 次元追加する．鉄道路線名や駅名が含まれるかの判定は，別途鉄道路線名リスト，及び駅名リストを作成し，単純な文字列マッチングにて行なった．これらのリストの作成には，国土数値情報ダウンロードサービス[15]や Wikipedia「日本の鉄道路線一覧」[19]を利用した．素性の追加による F 値の変化を図 6～9 に示す（縦軸は F 値，横軸は素性数 *n* を表す．横軸は対数目盛りである）．図 6～9 より，鉄道路線名，及び駅名が含まれるか否かを表す素性（2 次元）を追加することにより，分類性能（F 値）が向上していることがわかる．

今回は，交通機関カテゴリのみを対象に素性の追加を検討したが，その他のカテゴリにおいても適切な素性を追加することにより，分類性能の向上が図れると考える．例えば今回の分類にはハッシュタグを素性として利用しなかったが，それぞれのカテゴリに対応したハッシュタグ[20][21]を素性に加えることで精度向上が図れる可能性があると考えている．

3. まとめと今後の課題

我々の研究グループでは大規模災害時の被災者支援のため，災害関連ツイートを情報提供に適したカテゴリに分類し，地図上にマッピングするシステムを構築中である．本稿では，東日本大震災発生直後のツイートを用いて，機械学習によるカテゴリ分類を検証した．SVM，Random Forest，Naive Bayes により分類器を生成したところ，実用上十分と思われる精度で分類できることが確認できた．また，素性選択や形態素以外の素性の追加について検討を行なった．

本研究では，多くの条件でフィルタリングした少数のツイート（1655 ツイート）のみをデータセットとして利用したが，今後はより緩和された条件下でノイズとなるツイートも含むデータセットを作成し，ツイート分類の精度を検証する必要がある．また，本研究では東日本大震災時のツイートを用いてカテゴリ分類を試みたが，地震の発生地域や季節，時間帯によってはツイート内容が大きく異なる可能性がある．また，台風や大雪，ゲリラ豪雨など，災害の種類や規模には様々なものがある．そこで，災害種・災害規模にかかわらず広く適用可能な方法について検討する予定である．また，カテゴリ分類されたツイートを地図上にマッピングして被災者に提示するシステムを実装し，その有用性を検証する予定である．

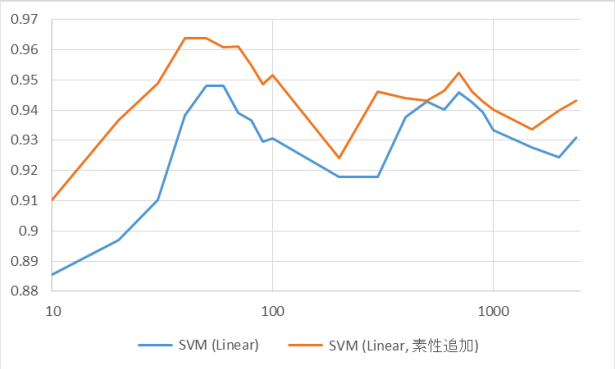

図 6　素性追加による F 値の変化（SVM, 線形カーネル）
（訓練データ: dataset1, 2, テストデータ: dataset3）

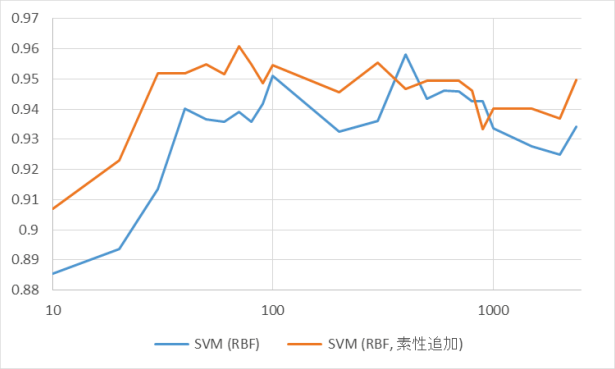

図 7　素性追加による F 値の変化（SVM, RBF カーネル）
（訓練データ: dataset1, 2, テストデータ: dataset3）

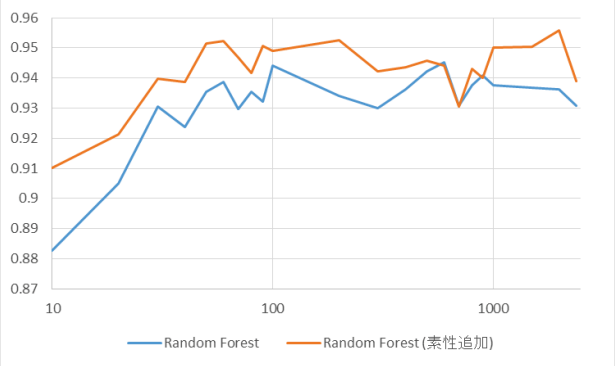

図 8　素性追加による F 値の変化（Random Forest）
（訓練データ: dataset1, 2, テストデータ: dataset3）

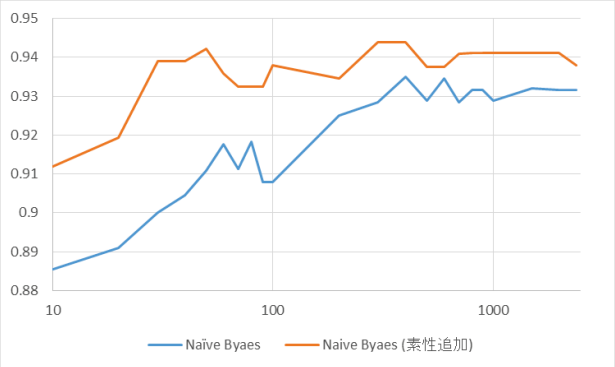

図 9　素性追加による F 値の変化（Naive Bayes）
（訓練データ: dataset1, 2, テストデータ: dataset3）

謝辞

本研究を行うにあたり，ツイートデータの収集に協力していただいたクックパッド株式会社の兼山元太氏に感謝する．本研究は，平成 25 年度文部科学省「地（知）の拠点整備事業」補助金の助成を受けて実施した．

参考文献

[1] 風間 一洋, “Twitter における情報伝播”, 人工知能学会誌, Vol.27, No.1, pp.35-42, 2012.
[2] 鳥海 不二夫, 篠田 孝祐, 栗原 聡, 榊 剛史, 風間 一洋, 野田 五十樹, “震災がもたらしたソーシャルメディアの変化”, JWEIN11, pp.41-46, 2011.
[3] H. Wilensky, “Twitter as a Navigator for Stranded Commuters during the Great East Japan Earthquake”, Proc. of 11th Int’l ISCRAM Conference, pp.695-704, 2014
[4] 株式会社 IMJ モバイル, “東北地方太平洋沖地震に伴う twitter, facebook 利用実態に関する調査”, http://www.imjp.co.jp/press/release/20110404_000581.html
[5] Micro Mappers, http://micromappers.com/
[6] 相田 慎, 新堂 安孝, 内山 将夫, “「東日本大震災関連の救助要請情報抽出サイト」による救助活動支援”, 自然言語処理, Vol.20, No.3, pp.405-422, 2013.
[7] 後藤 淳, 大竹 清敬, Stijn De Saeger, 橋本 力, Julien Kloetzer, 川田 拓也, 鳥澤 健太郎, “質問応答に基づく対災害情報分析システム” , 自然言語処理, Vol.20, No.3, pp.367-404, 2013.
[8] NEC ビッグローブ株式会社, “東日本大震災におけるツイッターの利用状況について”, http://tr.twipple.jp/info/bunseki/20110427.html
[9] 六瀬 聡宏, 長島 俊, 内田 理, 鳥海 不二夫, “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, 2013.
[10] 高畑 洋貴, 六瀬 聡宏, 榎本 光, 齊藤 大樹, 近藤 直人, 富田 誠, 梶田 佳孝, 山本 義郎, 鳥海 不二夫, 内田 理, “大規模災害時における避難支援情報の可視化”, 言語処理学会第 20 回年次大会発表論文集, pp.82-84, 2014.
[11] 馴田 俊平, 六瀬 聡宏, 榎本 光, 齊藤 大樹, 近藤 直人, 富田 誠, 梶田 佳孝, 山本 義郎, 鳥海 不二夫, 内田 理, “エリア限定型大規模災害時情報提供システム”, 言語処理学会第 20 回年次大会発表論文集, pp.67-69, 2014.
[12] C. Cortes and V. Vapnik, “Support-Vector Networks”, Machine Learning, Vol.20, No.3, pp.273-297, 1995.
[13] L. Breiman, “Random Forests”, Machine Learning, Vol.45, No.1, pp.5-32, 2001.
[14] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification”, Proc. AAAI-98 Workshop on Learning for Text Classification, pp.41-48, 1998.
[15] 国土数値情報ダウンロードサービス, http://nlftp.mlit.go.jp/ksj/
[16] Mecab, http://mecab.googlecode.com/
[17] scikit-learn, http://scikit-learn.org/
[18] F. Sebastiani, “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, Vol.34, No.1, pp.1-47, 2002.
[19] Wikipedia, 日本の鉄道路線一覧 http://ja.wikipedia.org/wiki/日本の鉄道路線一覧
[20] 村井 源, “東日本大震災後の Twitter 利用傾向 —震災関連ハッシュタグの計量的分析—”, 情報知識学会誌, Vol.22, No.2, pp.97-106, 2012.
[21] Twitter 社公式ブログ, “先ほどの地震について”, http://blog.twitter.com/ja/2012/xian-hodonodi-zhen-nituite