Twitterにおける観点に基づいた意見文クラスタリング

鷹栖 弘明 †　　小林 聡 †　　内海 彰 ††

† 電気通信大学 電気通信学部 情報工学科

†† 電気通信大学 電気通信学部 システム工学科

takasu@utm.se.uec.ac.jp, satoshi@cs.uec.ac.jp, utsumi@inf.uec.ac.jp

1　はじめに

近年，Web 上のサービスとして，マイクロブログが急速に普及している．マイクロブログには，1 記事あたりの文字制限や，簡単な情報発信，ユーザ間のやりとりといった特徴がある．その中でも，本研究では Twitter を取り扱う．Twitter では各記事をツイートと呼ぶ．

Twitter のようにユーザが発信する情報コンテンツにおいては，意見や評判を述べることが多い．これらの意見や評判を自動分類する研究では，肯定的か否定的かを判定するものがほとんどである [1, 2]．しかし，例えば,「原発」に関する意見に，経済・エネルギー・科学技術・健康など複数の観点が存在するように「どのような箇所に焦点を当てているか」という観点ごとの情報を得たいという場面は多々ある．そこで本研究では，観点ごとに意見文を分類するシステムの構築を目的とする．

観点ごとに意見文を分類するために文書クラスタリングを行うが，1 ツイートあたり 140 文字の制限がある Twitter では，意見文どうしの類似度を計算するには情報量が少ないため，適切にクラスタリングができない．本研究では，意見文どうしの類似度を適切に計算するのに十分な情報を得るために，意見文に関連するユーザのツイートを考慮し，意見の観点に基づいてクラスタリングを行う手法を提案する．

2　関連研究

Twitter を対象とした意見や評判の自動分類の先行研究として，橋本ら [1] は詳細な評判傾向の抽出を目的として，意見文中に用いられる感情表現の違いから意見文をその感情ごとに分類している．しかし，これは感情表現しか見ていないため，意見の観点から分類する本研究とは異なる．また，Jiang ら [2] は，ユーザの関連ツイートを考慮して意見文を肯定/否定/中立に分類している．本研究でもユーザの関連ツイートを考慮するが，肯定/否定/中立の分類が目的ではない．

ツイートに限らず，文書クラスタリングでは文どうしの類似度を計算する必要があり，一般的には単語の出現頻度を用いて計算される．しかし，ツイートは従来のブログ記事に比べ短文であるために，既存の文間類似度を適用するのが難しいという問題点がある．そこで，一般的なツイートを対象としたクラスタリングの研究において，青島ら [3] は日本語 WordNet を用いて，各ツイートに含まれる単語の概念類似度を計算し，これをツイート間の類似度計算に利用している．しかし，ツイート中の単語が WordNet に含まれていない場合が多いために，クラスタリングの精度が悪くなるという結果が報告されている．

3　提案手法

本章では，本研究で提案する関連ツイートを考慮した意見ツイートのクラスタリング手法について述べる．また，本手法におけるアルゴリズムを図 1 に示す．

3.1　意見ツイートの抽出

松本ら [4] の手法を援用し，Twitter のキーワード検索結果のツイート集合 S から意見ツイートを抽出する．松本らの手法では，文末を「文の終端から遡って自立語が出現するまで」と定義し，文末表現に含まれる助動詞や終助詞，表明思考動詞（思うや考えるなど）の出現頻度を素性として SVM で分類器を学習し，ウェブページの主観/客観を判定している．本研究では，文末の助動詞，助詞全て，表明思考動詞に加えて,「こと」や「もの」など非自立語扱いとなる単語を利用する．

しかし，文の終端が自立語の場合は文末表現が抽出できず，SVM による判定ができない．また，文末表現に含まれる情報が少ない場合には，特定の素性の情報のみによって意見かどうかの判定がされ，精度が低下してし

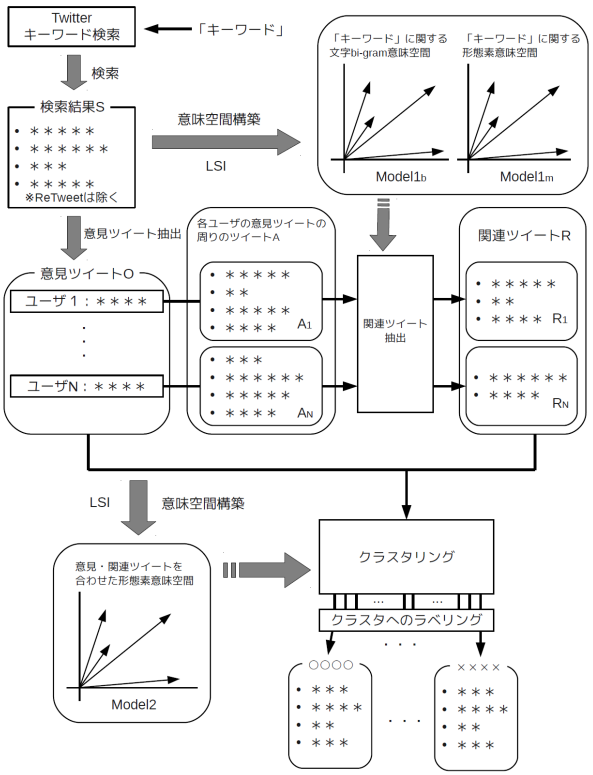

図 1: 本手法におけるアルゴリズム

まう恐れがある．そのため，李 [5] と同様に文末定義の拡張として，文末表現の形態素が 2 つ以下の文については形態素を 4 つまで追加し，素性に使用する情報を増やすこととする．

3.2 関連ツイートの抽出

3.2.1 関連ツイートの定義

Twitter では，文字数の制限や投稿のしやすさから，従来のブログでは 1 回の投稿で済む内容を複数回に分けて投稿することがある．そのため，あるユーザが特定のトピックに対する意見をツイートしたとき，複数回の投稿で 1 つの意見となることがある．

そこで，ユーザ x の意見ツイート o_i の周りにある，ユーザ x のツイート集合 A_i の中で意見に関連するツイート集合を関連ツイート R_i と定義する．

3.2.2 意味空間の構築

関連ツイートの抽出にあたり，LSI[6] を用いて意味空間を構築する．本研究では，キーワード検索の結果 S から形態素単位の意味空間 $Model1_m$ と文字 bigram 単位の意味空間 $Model1_b$ を構築する．この際，以下の条件を含む bigram を省くことにする．

- S 内でツイートに出現する頻度の高い文字列（漢字・カタカナ・英字の文字種ごとに切り出した文字列）に含まれるもの
- ひらがなのみで構成されるもの

3.2.3 類似度計算と抽出

ある意見ツイート $o_i \in O$ の周りのツイートを $a_{ij} \in A_i$ とするとき，関連ツイートの抽出には以下の 3 つの類似度を利用する．

- sim_m：$Model1_m$ 上でベクトル表現された o_i と a_{ij} の cos 類似度
- sim_b：$Model1_b$ 上でベクトル表現された o_i と a_{ij} の cos 類似度
- sim_t：o_i と a_{ij} の投稿時間による時間類似度

時間類似度は，戸田ら [7] の手法を利用する．戸田らは,「文書間のタイムスタンプが離れるごとに，一定の割合で文書内容の類似度が減少する」という仮定に基づいて時間類似度を式 (1) のように定義している．

$$\mathrm{sim}(t_P, t_Q) = T_0 \times \exp\left(-\frac{0.693}{t_{1/2}}|t_P - t_Q|\right) \quad (1)$$

t_P, t_Q はそれぞれ記事 P，Q のタイムスタンプ（単位は日），T_0 は $t_P = t_Q$ のときの重み，$t_{1/2}$ は時間類似度が 50%になるタイムスタンプの差（半減期）を示すパラメータである．本研究においても，複数のツイートから構成される意見は，まとまった時間内にツイートされると考えられるので，戸田らの手法を利用し，タイムスタンプの単位は秒とする．

先に挙げた 3 つの類似度の重みとして，パラメータ $,\beta,\gamma$ ($+\beta+\gamma=1$) を用いて全体の類似度 $\mathrm{sim}(o_i, a_{ij})$ を式 (2) のように定義する．類似度が閾値 T を超えた a_{ij} を関連ツイートとして抽出する．

$$\mathrm{sim}(o_i, a_{ij}) = \quad \times \mathrm{sim}_b + \beta \times \mathrm{sim}_m + \gamma \times \mathrm{sim}_t \quad (2)$$

3.3 クラスタリング

意見ツイートのクラスタリングには，階層的クラスタリング手法である Ward 法を用いる．

意見ツイート集合 O と関連ツイート集合 R から構築した形態素ベースの意味空間 *Model2* を構築し，意見ツイート o_i の特徴ベクトル $\boldsymbol{u}_i$ と関連ツイート $r_{ij} \in R_i$ の特徴ベクトル $\boldsymbol{v}_{ij} \in V_{R_i}$ を生成する．$\boldsymbol{u}_i$ と V_{R_i} の重心ベクトル $\boldsymbol{w}_i$ を意見 o_i の特徴ベクトルとして再定義し，意見ツイート o_i, o_j の類似度を $\boldsymbol{w}_i, \boldsymbol{w}_j$ の cos 類似度により計算する．

Model2 の構築には，意見ツイート o_i および関連ツイート R_i に出現する単語の頻度を利用する．意見に関連している単語に重みを置くために $r_{ij} \in R_i$ に出現する単語の頻度は，その頻度と式 (2) の類似度 $\mathrm{sim}(o_i, r_{ij})$ の積とする．

3.4 ラベリング

クラスタへのラベリングを行い，意見の観点を抽出する．意見の観点を表す単語の抽出手法は検討中であるが，本研究ではクラスタ内の意見ツイート（関連ツイートを含む）の特徴語をそのクラスタのラベルとする．各クラスタに含まれる名詞句（形態素解析したときの名詞の連続）に対して TF-ICF（Inverse Cluster Frequency）により特徴語を抽出する．単語 w のクラスタ c 内での出現頻度を $tf_{w,c}$，単語 w が出現するクラスタ数を cf_w，全クラスタ数を C としたとき，クラスタ c における単語 w の TF-ICF 値 $tf\ \ icf_{w,c}$ は式 (3) のように求まる．

$$tf\ \ icf_{w,c} = tf_{w,c} \times \log(C/cf_w) \quad (3)$$

TF-ICF 値を計算する際には，クラスタリングのときと同様の手法を取り，クラスタ内に含まれる名詞句の出現頻度を利用する．

4 評価実験

4.1 意見ツイート抽出

4.1.1 手順

分類器の学習には,「原発」,「増税」,「防災の日」,「代表選」,「衆議院選挙」,「笹子トンネル」,「ノロウィルス」の Twitter キーワード検索で得られたツイート 1075 件（リツイートを除く）を使用する．得られた各ツイートに対して 2 人ずつ意見/非意見のラベル付けを行い，2 人のラベルが一致したツイート 599 件（意見 270 件，非意見 329 件）を訓練データとして利用した．

なお，SVM のツールには LIBSVM を使用し，SVM のタイプは C-SVM（緩和制約下における SVM モデル）を指定し，分類における誤りをどれだけ許容するかを示す指標であるコストパラメータ C を $C = 2048$ とした．また，カーネル関数には非線形な分類が可能である RBF（Radial Basis Function）を設定し，RBF で使用するカーネルパラメータ γ を $\gamma = 3.05e^{-5}$ とした．

訓練データとは別に，ツイートが重複しないように

表 1: 意見ツイート抽出評価の結果

キーワード	評価対象	正答率	再現率	適合率	F 値
原発	33/35	0.721	0.818	0.675	0.740
衆議院選挙	12/35	0.745	0.750	0.500	0.600
尖閣諸島	26/92	0.873	0.731	0.704	0.717
平均		0.780	0.766	0.626	0.686

表 2: 関連ツイート抽出評価の結果

キーワード	評価対象	パラメータ							再現率	適合率	F 値
		$d1_b$	$d1_m$	$t_{1/2}$		β	γ	T			
原発	37/198	15	20	120	0.45	0.25	0.30	0.50	0.351	0.929	0.510
衆議院選挙	29/69	10	10	120	0.45	0.25	0.30	0.55	0.759	1.00	0.863
尖閣諸島	20/193	10	10	180	0.25	0.45	0.30	0.55	0.500	0.385	0.435

新たに「原発」,「衆議院選挙」,「尖閣諸島」の Twitter キーワード検索で得られたツイート 270 件（リツイートを除く）に対して先ほどと同様に，人手で意見/非意見のラベル付けを行い，評価用データ（233 件）を作成した．

4.1.2 結果

訓練データで学習した分類器を用いて，評価データから意見ツイートの抽出を行った結果を表 1 に示す．なお，評価対象の列にある数値は「意見の数/非意見の数」を指す．

F 値の平均値が 0.686 とおおむね良い結果となり，松本ら [4] の手法を援用した意見抽出は有効だと言える．

4.2 関連ツイート抽出

4.2.1 手順

4.1 節において，人手で意見だと判定された意見ツイート集合に対して，各意見ツイートの前後のツイート（リツイートを除く）5 件ずつを意見の周りのツイートとして取得した．意見の周りのツイートが，その意見に関連するものかどうかを 1 ツイートにつき 2 人ずつラベル付けを行い，2 人のラベルが一致したツイートを評価用データとして作成した．

4.2.2 結果

人手で意見だと判定されたツイートに対して本手法で関連ツイートを抽出した．予備実験の結果から，F 値が最大を取ると推測された各パラメータの範囲を以下に示す．

- $(d1_b, d1_m) = \{10, 15, 20\} \times \{10, 15, 20\}$
- 時間類似度の半減期 $t_{1/2} = \{60, 120, 180\}$
- 閾値 $T = \{0.5, 0.55, 0.6\}$

なお，$d1_b$，$d1_m$ は $Model1_b$，$Model1_m$ の次元数を指す．3.2.3 節で述べた β, γ については，$\gamma = \{0, 0.1, 0.2, 0.3\}$ と変化させたとき，$\delta = (1-\gamma)/2$ として $(\beta) = \{(\delta-0.1, \delta+0.1), (\delta, \delta), (\delta+0.1, \delta-0.1)\}$ と変化させた．

以上の組み合わせを変化させ，F 値の最大値およびそのときのパラメータを表 2 に示す．

キーワード「衆議院選挙」のときは F 値が 0.863 と良い結果となったが，他の 2 つのキーワードにおいては，良くない結果となった．

4.3 クラスタリング

4.3.1 手順

4.2 節において，人手で関連ツイートだと判定されたツイートとその意見ツイートに対して，本手法で予めクラスタリングを行う．そのクラスタリング結果をもとに人手（各キーワード 2 人ずつ）で観点ごとにツイートがまとまるようにクラスタの修正を行った．その後，人手により修正されたクラスタリング結果と同じクラスタ数でシステムでクラスタリングを行った．

表 3: クラスタリング評価の結果

キーワード	$d2$	クラスタ数	F 値	平均
原発	10	7	0.430	0.413
		9	0.396	
衆議院選挙	10	5	0.794	0.734
		5	0.674	
尖閣諸島	15	10	0.381	0.448
		11	0.514	

4.3.2 結果

人手による正解クラスタ群 $\boldsymbol{L}$ とシステムによるクラスタ群 $\boldsymbol{S}$ が，どの程度近いかの指標として F 値 [8] を用いた．これは，$\boldsymbol{L}$ と $\boldsymbol{S}$ に含まれるクラスタどうしで F 値を計算し，F 値の重み付き平均が最大となるようなクラスタの組み合わせを決める方法である．

$Model2$ を構築する際に行う行列の次元圧縮後の次元数を $d2$ とし，10〜20 の 5 刻みで変化させた．各キーワードにおいて 2 人の平均 F 値が高い時の $d2$ および結果を表 3 に示す．

キーワード「衆議院選挙」のときは F 値が 0.734 と良好な結果となったが，他のキーワードにおいては，4.2 節と同様に良くない結果となってしまった．

4.4 ラベリング

4.4.1 手順

4.3 節で人手でクラスタリングした結果の各クラスタに対して，TF-ICF 値の高い名詞句を 3 つまで出力した．出力された各クラスタのラベル 3 つそれぞれに対し，4 段階（適切：4，不適切：1）でクラスタに含まれる意見の観点として適切かどうかをクラスタリングのときと同じ参加者で評価を行った．

4.4.2 結果

各クラスタの，すべてのラベルに対する平均スコアを表 4 に示す．なお,「スコアが 3 以上の割合」の列は，やや適切：3 以上と評価されたラベルを含むクラスタ数の割合を指す．スコアが 3 以上の「やや適切である」「適切である」のラベルを含むクラスタは全体の 80%近くとなっていたので，TF-ICF による名詞句を特徴語としたラベル付けは有用であると思われる．また，キーワード「衆議院選挙」では 4 段階の中間値 2.5 を上回る良い結果となったが，他のキーワードにおいては，中間値を下回る悪い結果となった．

5 考察

5.1 意見ツイート抽出

システムが誤って意見と判定したツイートの中には，ツイート中に意見を述べている引用文を含むものや，ニュースサイトからの投稿で，単にニュース記事の紹介をしているツイートが特に目立った．

引用文を含むツイートについては，引用文を削除する

表 4: ラベリング評価の結果

キーワード	平均スコア	平均	スコアが 3 以上の割合
原発	1.95	1.90	42.9%
	1.85		66.7%
衆議院選挙	2.47	2.80	100%
	3.13		100%
尖閣諸島	2.13	2.27	80.0%
	2.42		90.9%

などの処理を施す必要があると考えられる．また，ニュース記事を紹介している投稿に対してはフィルタをかける必要があると思われる．

5.2 関連ツイート抽出

各キーワードにおいて，$\gamma = 0$ のときに比べ，$\gamma > 0$ のときの方が F 値が高くなったことから，ツイート間の類似度計算に時間類似度が有効であることが分かる．しかし，γ の範囲を決める際に，γ を大きくしすぎると逆に F 値が下がってしまっていた．そのため，形態素や文字 bigram で計算した類似度を時間類似度で補正する程度が良いと言える．

システムが関連ツイートでないと誤って判定したツイートでは，そこに含まれている単語が意味空間 $Model1_b$，$Model1_m$ に含まれていないことが多かった．ツイートだけでなく，Wikipedia など外部の情報を，ツイートに含まれる単語間の類似度計算に利用することで改善できるのではないかと考えられる．

キーワード「尖閣諸島」においては，システムが誤って関連ツイートだと判定したツイートが多かった．意味空間の構築に使用したツイート数が多く，次元圧縮をしたことでキーワードに関連する多くの単語が，意見の観点となる単語と同等に扱われ，意見ではなくキーワードに関連したツイートを抽出してしまったためと考えられる．そのため，意味空間の構築に使用するツイート数が多い場合には，ツイートに対しフィルタリングを行う必要があると考えられる．

5.3 クラスタリング

本研究では，引用リプライについては引用部より前の情報のみを抜き取っており，引用部の情報については考慮していない．キーワード「原発」において，正解クラスタでは引用部の情報も見てツイートが分類されていても，システムでは引用部を考慮しないために誤って別のクラスタにツイートを分類しているケースが目立った．

また，意見ツイートが短く，関連ツイートの方が情報量が多いために，関連ツイートの内容を中心にクラスタリングされてしまっていたケースも目立った．これは，本手法では意見/非意見を問わず，意見に関連する内容を含むツイートを関連ツイートとして抽出しているためと考えられる．つまり，関連ツイートを意見ツイートの観点に基づいて抽出する必要があり，関連ツイートの定義・抽出手法について再考しなければならないと言える．

5.4 ラベリング

観点が同じでもクラスタ中で同じ単語が何度も繰り返されることがないために，他の単語と TF-ICF 値の差がなくなってしまい，クラスタ内にラベルとなりうる単語があるにも関わらず出力されていないケースがあった．また，クラスタ内のツイートからではラベルとなりうる単語が取りづらいケースもあったため，観点の抽出には改善点があると考えられる．

6 おわりに

本論文では，Twitter 上に存在する意見ツイートに対して，ユーザの意見ツイートに関連するツイートを考慮することで情報量を増やし，観点ごとに意見ツイートをクラスタリングする手法を提案した．意見ツイートの抽出についてはおおむね良い結果が得られた．また，関連ツイート抽出の評価における最大 F 値が 0.863，クラスタリング評価の最大平均 F 値が 0.734 ということで，精度にばらつきはあるもののキーワードによっては関連ツイートを考慮したクラスタリングが有用であると推測された．

今後の課題としては，関連ツイート抽出において Wikipedia など外部の情報などを使ったツイート間の類似度計算の改善による精度の向上や，他のクラスタリング手法による検討，TF-ICF 以外による観点の抽出方法の考案が挙げられる．また，本研究ではリツイートを一切考慮しなかったが，リツイートをしてから意見を述べるというケースもよくあるため，リツイートも考慮したクラスタリング手法の考案も今後の課題である．

参考文献

[1] 橋本 和幸，中川 博之，田原 康之，大須賀 昭彦：センチメント分析とトピック抽出によるマイクロブログからの評判傾向抽出, 電子情報通信学会論文誌-D, J94-D(11), 1762-1772 (2011).

[2] L.Jiang, M.Yu, M.Zhou, X.Liu and T.Zhao: Target-depedent twitter sentiment classi cation, *Proc. of ACL2011*, 151-160 (2011).

[3] 青島 傳隼，福田 直樹，横山 昌平，石川 博：マイクロブログを対象とした制約付きクラスタリングの実現, 第 2 回データ工学と情報マネジメントに関するフォーラム（DEIM2010） 論文集, B1-3 (2010).

[4] 松本 章代，小西 達裕，高木 朗，小山 照夫，三宅 芳雄，伊東 幸宏：文末表現を利用したウェブページの主観･客観度の判定, 第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM2009） 論文集, A5-4 (2009).

[5] 李 笈鎬：機械学習による意見文抽出，電気通信大学 平成 23 年度システム工学科卒業論文 (2011).

[6] S.Deerwester et al.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, 41, 6, 391-407 (1990).

[7] 戸田 浩之，北川 博之，藤村 考，片岡 良治：時間的近さを考慮した話題構造マイニング, 電子情報通信学会 第 18 回データ工学ワークショップ（DEWS2007） 論文集, L6-4 (2007).

[8] 折原 大，内海 彰：HTML タグを用いた Web ページのクラスタリング手法, 情報処理学会論文誌, 49(8), 2910-2921 (2008).