1B1-R-3-7

Twitterにおける語の関連性に着目したユーザ興味語抽出手法の提案

Extracting User's Preference Focused on Word Relationship in Twitter

渡邊 恵太　　加藤 昇平
Keita Watanabe　　Shohei Kato

名古屋工業大学工学研究科情報工学専攻
Department of Computer Science and Engineering Graduate School of Engineering Nagoya Institute of Technology

Recently, Twitter becomes widly spread rapidly and attracts much attention as large information sources. One of the feature of Twitter is that many information is contributed in real time. However, conventional searching methods like keyword search are not enough to acquire information about user's interest from Twitter immediately, because user must type a keyword every time when they want to search. Consequently, the system which recommends information based on user's interest in real time is required. It is important for such a system to analyze user's contribution and extract user's preference from Twitter. The length of contribution is limited to less than 140 words at Twitter. Therefore information is scare for extracting user's preference. In this paper, we propose a method of extracting user's preference based on relationship between words in Twitter. We confirmed the validity of the system by user evaluation experiment.

1. はじめに

近年，Twitter が急速に普及し，Twitter から情報を得る機会が増えている．ユーザが Twitter から自身の嗜好にあった情報を得る方法としては，あらかじめ登録した他ユーザの投稿が時系列順に表示されるログ画面を監視する方法が一般的に用いられている．この方法では，Twitter の特徴である即時性の高さが活かされる一方で，投稿を誰の投稿かという情報で選別するためユーザの嗜好にあった情報が Twitter に投稿されていても必ずしもユーザに提示されないという問題が存在する．また，別の方法としてキーワード検索が挙げられるが，この方法では情報を得るためにその都度検索を行わなければならず，Twitter の特徴である即時性が損なわれてしまう恐れがある．このため，Twitter においてユーザに対しその嗜好に合わせて情報を選別する研究が行われている [桑原 09, 後藤 10]．このような情報推薦機能を実現するためには，ユーザがどのような嗜好を持っているか分析し，明らかにすることが重要である．

これまでブログなどにおいて，ユーザの嗜好を分析する方法としてユーザの投稿を tf·idf 法により重み付けし，興味語を抽出する手法が用いられてきた．Twitter においても，ユーザ自身の投稿だけでなくユーザの友人の投稿も関連付けて tf·idf 法によりユーザの嗜好分析を試みた研究が行われている [早川 11]．しかし，Twitter のようなマイクロブログでは投稿が短文に制限されており分析の対象となる情報が少ない．また，ブログに比べ投稿が手軽に行うことができるとされており，投稿中で同じ意味の別や略語が用いられることが多い．tf·idf 法では，関連性がある語であっても表記の異なる語は別の情報としてしか扱うことができない．このような場合には，同一の分野に関する語や同じ意味で用いられた語に対してもそれぞれ単体に重みが与えられユーザ嗜好分析の性能低下を招く．これらの問題に対処するために，Twitter の投稿からユーザの嗜好を分析するには語と語の関連性を考慮することが必要であると考える．

Twitter において語の関連性を考慮してユーザの嗜好を分析した研究としては，Wikipedia の情報からオントロジー辞書を構築し，その辞書を用いて Twitter ユーザの興味を分析した研究 [宮城 09] やソーシャルブックマークのタグの共起関係から語の関連性を分析し，その情報を用いてユーザの興味分析を行った研究 [齋藤 11] がある．しかし，Twitter では，ユーザが手軽に投稿でき，新しい情報が次々と寄せられるため，新語の発生や語の関連性の変化の速度が非常に早い．そのため，語の関連性の分析に Wikipedia やソーシャルブックマークを利用した場合，Twitter における語の関連性の変化に十分に対応することができないと考える．したがって，新語や語間の関連性の変化に対して高速に対応するためには，Twitter に寄せられた投稿から語と語の関連性を分析する必要があると考える．

Twitter に寄せられる投稿から語間の関連性を分析する手法には，投稿数の時系列的な変化に着目した研究 [和泉 11] がある．この研究では，ある語が単位時間において出現頻度が上昇した際，その語の共起語が同様に上昇していた場合に関連語とする手法をとっている．しかし，この方法では時系列的な変化に大きく依存しており，ノイズによる影響を受け易いと考える．また，他の研究には n-gram を用いて頻出単語解析を行い，共起関係から語の関連性を分析した研究 [岩井 11] がある．この研究では n-gram を用いて投稿からの単語抽出を行うため，単語抽出の安定性が低い．また，共起関係からのみ語の関連性を分析するため，一般的に多く用いられる語が関連語として捉えられやすい問題がある．

本研究では，Twitter に寄せられる投稿を分析し，語の共起関係と共に逆文書出現頻度を用いて関連語の辞書を作成する．この辞書を用いることで語の関連性を考慮したユーザ嗜好の分析を行い，ユーザの嗜好を表す興味語を抽出する．これにより，Twitter における新語の発生や語の関連性の高速な変化に対応することができ，また語の表記のぶれや分析の対象とする情報が少ないという問題にも対処してユーザの興味語を抽出することができると考える．

2. 興味語抽出システム概要

本研究では，Twitter におけるユーザの投稿を収集，分析することでユーザがどのような嗜好を持っているか分析し，ユーザの興味を表す興味語の抽出を行う．提案システムでは，特に

連絡先: 加藤昇平，名古屋工業大学，愛知県名古屋市昭和区御器所町，052-735-5625，shohey@katolab.nitech.ac.jp

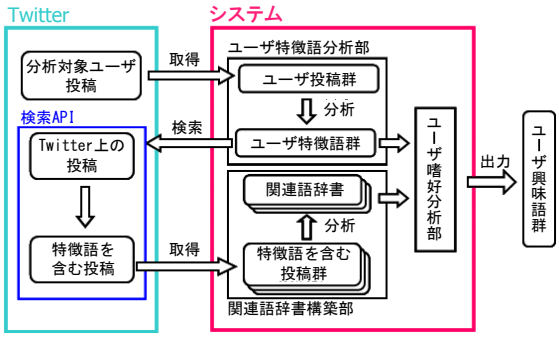

図 1: システム概要図

語の関連性に着目しており，Twitter に寄せられた投稿を分析することで関連語辞書を構築し，この辞書をユーザの興味語抽出に利用する．

提案システムは以下の 3 つの要素からなる．

- ユーザ特徴語分析部
- 関連語辞書構築部
- ユーザ嗜好分析部

提案システムの概要を図 1 に示す．ユーザ特徴語分析部では，Twitter におけるユーザの投稿からそのユーザに特徴的な語を分析する．関連語辞書構築部では，Twitter に寄せられる投稿から語間の関連性を分析し，その情報を格納した辞書を構築する．ユーザ特徴語抽出部では，ユーザの特徴語情報と関連語辞書を用いて，語の関連性を考慮してユーザの興味を表す興味語の抽出を行う．以下に各部について詳しく述べる．

2.1 ユーザ特徴語分析部

ユーザ特徴語分析部では，tf·idf 法を用いてユーザの投稿に出現する各語に重み付けを行うことで，ユーザに特徴的な語を分析する．

tf·idf 法は，文章に出現する語に重み付けを行う手法であり，情報検索分野において索引語の重み付け手法として用いられてきた [天野 07]．この手法では，文章を特徴付ける語とはその文章に多く出現し，また他の文章にはあまり出現しないような語であるという考えに基づいており，文章に特徴的な語ほど大きな重みが与えられる．近年では tf·idf 法をブログの投稿に適用することで，その投稿者がどのような嗜好を持っているか分析するのに用いられている．これは，ブログにおける投稿は筆者の嗜好が反映されている場合が多く，tf·idf 法による重み付けがユーザの興味分野を反映したものになると考えられるからである．本研究においても，この考えに基づき Twitter の投稿に対して tf·idf 法を適用し，ユーザに特徴的な語の分析を行う．

まず，Twitter から TwitterAPI を用いて分析対象ユーザの投稿を新しいものから最大で 3200 件取得する．次に取得した各投稿を形態素解析基 Sen を用いて形態素解析を行う．ここで，本研究では重み付けの対象とする語を名詞に限定する．これは，本研究の目的がユーザの嗜好を表す興味語を抽出することにあり，このような語としては名詞が最も適していると考えたためである．形態素解析によって取り出された各名詞に対しては，tf·idf 法により重み付けを行う．tf·idf 法によって語に与えられる重みを**重要度**とすると，ユーザ投稿に含まれる語 i の重要度 w_i は以下の式によって与えられる．

$$w_{i,u} = tf_{i,u} \cdot \log \frac{|U|}{df_i} \tag{1}$$

ここで，$tf_{i,u}$ はユーザ u の投稿における語 i の出現頻度を表し，Twitter ユーザの集合を U，またその総数を $|U|$，U のうち語 i を含む投稿を行ったユーザ数を df_i とする．重要度はユーザの投稿に多く現れ，他ユーザの投稿にあまり現れない語ほど大きな値を与えられる．つまり，重要度が大きな語ほどユーザに特徴的な語であるということができる．ただし，tf·idf 法で与えられる重要度は語の関連性を考慮しないため，これ単体では Twitter における投稿からユーザの興味語を抽出するのには十分でないと考える．

2.2 関連語辞書構築部

関連語辞書構築部では，Twitter に寄せられた投稿を分析することで語間の関連性を分析し，その情報を格納した関連語辞書の構築を行う．ここで構築した関連語辞書を用いてユーザ嗜好の分析を行うことで，tf·idf 法では考慮されない語の関連性を考慮することができる．

Twitter に寄せられる投稿を基に語間の関連性を分析する利点としては，Twitter のリアルタイム性の高さから，新語の発生や語間の関連性の変化に高速に獲得，辞書に反映できるという点が挙げられる．また，Twitter では投稿が 140 文字以下の短文に制限されるため，ニュース記事やブログ等に比べ関連性が低い語が同一投稿中に出現しにくいという特徴がある．これにより，Twitter の投稿は辞書構築に適した情報源であると考えられる．

関連語辞書を構築するために，まず辞書を構築したい語を含む投稿を Twitter から収集する．収集には TwitterAPI を用い，収集可能な投稿を直近のものから収集していく．TwitterAPI において収集可能な範囲に 2000 件以上の投稿があった場合は直近から 2000 件のみを収集対象とし，2000 件以下の場合は収集できる投稿を全て収集した．次に，収集した投稿を形態素解析し，名詞を抽出する．ここで抽出された名詞は，辞書の構築対象語と同一投稿中に出現した語であることを意味する．Twitter では投稿は短文に制限されるため一つの話題から構成される場合が多い．従って収集した投稿中に出現する頻度が高い語ほど辞書の構築対象語と関連性が高い語といえる．ここで，辞書の構築対象語を語 i とすると収集した投稿中に語 j が出現する頻度 $c_{j,i}$ は以下の式によって与えられる．

$$c_{j,i} = \frac{m_{j,T_i}}{|T_i|} \tag{2}$$

ここで，m_{j,T_i} は T_i に含まれる投稿のうち語 j を含む投稿の数を表すものとする．

$c_{j,i}$ は語 j が Twitter 上で一般的に多く用いられる語であった場合，語 i がどのような語でも高い値を取ることが考えられる．しかし，実際には一般的に多く用いられる語が全ての語と関連性が高いとは言えず，$c_{j,i}$ をそのまま語の関連性の強さを表す指標として用いるのは適切でない．したがって本研究では辞書構築対照語 i に対する語 j の関連性の強さ $r_{j,i}$ を以下の式によって算出する．

$$r_{j,i} = c_{j,i} \cdot \log \frac{|D|}{m_j} \tag{3}$$

ここで，$|D|$ は Twitter から収集したユーザの過去の投稿郡の集合であり，m_j は D のうち語 j を含む投稿を行ったことがあるユーザの数である．

表 1: 「伏見稲荷」の関連語辞書

順位	関連語	関連度
1	鳥居	0.481
3	千本	0.240
4	稲荷	0.230
5	京都	0.227

構築する関連語辞書では，この $r_{j,i}$ を正規化した値を用い，これを**関連度**とすると辞書構築対象語 i に対する語 j の関連度 $s_{j,i}$ は以下の式によって与えられる．

$$s_{j,i} = 2\varsigma_\alpha(r_{j,i}) - 1 \tag{4}$$

$$\varsigma_\alpha(x) = \frac{1}{1 + \mathrm{e}^{-\alpha x}} \tag{5}$$

なお，本稿では $\alpha = 5$ とした．関連度は，語間の関連性が大きいほど 1 に近い値を取り，関連性が小さいほど 0 に近い値となる．関連語辞書構築部では，収集した辞書構築対象語を含む投稿に出現するすべての名詞に対し関連度を計算し，関連度が大きい上位 50 語について，その関連度を格納した関連語辞書を構築する．実際に構築された関連語辞書の例を表 1 に示す．

2.3 ユーザ嗜好分析部

ユーザ嗜好分析部では，ユーザ特徴語分析部においてユーザの投稿に出現する各語に対し算出された重要度と，関連語辞書構築部において構築した関連語辞書を用いることで語と語の関連性を考慮したユーザ嗜好の分析を行う．ユーザ嗜好分析部で用いる関連語辞書は，ユーザ特徴語分析部において算出された重要度が大きい上位 150 語のものとする．この際，関連語辞書構築部で関連語辞書を作成するために用いる投稿数は，語によって異なる．辞書の構築に用いる投稿数は数が多いほど辞書の信頼性が高くなると考えられることから，各辞書について構築に用いた投稿数に基づく辞書信頼度を設定する．語 i の関連語辞書の辞書信頼度 p_i は以下の式によって与えられる．

$$p_i = \left(\frac{|T_i|}{|T_{max}|}\right)^{\frac{1}{2}} \tag{6}$$

ここで，$|T_a|$ は語 a の関連語辞書構築のために用いた投稿数，T_{max} は収集対象とする投稿数の最大数であり本稿では 2000 とした．また，収集できた投稿数が 100 より少なかった場合については信頼出来る辞書の構築ができないと判断し，全ての語に対して関連度が 0 であるものとして扱うこととした．ここで，ユーザのある語に対する興味の強さを**興味度**とすると，語 i に対する興味度 v_i は以下の式によって与えられる．

$$v_i = w_{i,u} + \sum_{k \in L} w_{k,u} \cdot s_{i,k} \cdot p_k \tag{7}$$

ここで，$w_{i,u}$ はユーザの投稿集合 u における語 i の重要度であり，L は関連語辞書が構築された語の集合，$s_{i,k}$ は関連語辞書を持つ語 k に対する語 i の関連度である．

また，関連語辞書を用いることでユーザの投稿に含まれなかった語（以降「未投稿語」とする）であっても，関連語辞書内に含まれる語であれば興味度が与えられる．未投稿語 j の興味度 v_j は以下の式によって与えられる．

$$v_j = \sum_{k \in L} w_{k,u} \cdot s_{j,k} \cdot p_k \tag{8}$$

表 2: 興味語抽出の例

順位	推定興味語	興味	未投稿
1	棋聖	○	
2	米長	○	
3	永世		
4	電王	○	
5	資料		
6	ボンクラーズ	○	○
7	将棋	○	○
8	棋士	○	○
9	対局	○	○
10	古紙回収	○	

式 (7),(8) に示したとおり興味度は関連語辞書を用いることによってユーザの投稿中に含まれる語同士の関連性を考慮して算出がされる．つまり，ユーザが関連性のある複数の語を用いて投稿をしていた場合，tf·idf 法では各語が別々に重み付けされるのに対し，提案手法の興味度は語同士を結びつけて算出がされる．これにより，Twitter のようにユーザの嗜好を分析するのに使用できる情報が少ない場合でも，

3. ユーザ興味語抽出実験

本研究で提案したユーザ嗜好抽出システムを用いてユーザの嗜好を表す興味語を抽出できるか，ユーザ評価実験を行い調査を行った．評価実験では，被験者を Twitter ユーザ 10 名とし，実際に被験者の投稿を収集，提案システムを用いて各ユーザの嗜好を表す興味語を抽出した．また，抽出された興味語について，ユーザにアンケートを行い興味語が適切に抽出できているか，未投稿語でも語の関連性を考慮することでユーザの興味語として正しく推定，抽出することができるか検証を行う．

3.1 興味語抽出の一例

ある被験者の投稿に対して提案手法により興味語抽出を行った結果を表 2 に示す．表中の順位は算出された興味度の大きい順であり，推定興味語は抽出された興味語，興味は被験者が抽出された興味語に興味があったかどうか，未投稿は抽出された興味語がユーザの投稿に含まれたかどうかを示している．本実験で用いた投稿は 2011 年 1 月 18 日から 23 日の間に収集した．表中にあるように，将棋に関する興味語が多く抽出された．また，興味度 6 位の「ボンクラーズ」は未投稿であるが被験者が興味を持っている語であった．これは，他の将棋に関する語の関連語として抽出されており，特に 2012 年 1 月 14 日の日本将棋連盟会長の米長氏がコンピュータ将棋ソフト「ボンクラーズ」に敗北したニュースを反映したものと考えられる．このように，Twitter から語の関連性を推定することで高速に語の関連性の変化を捉えることが可能である．

3.2 アンケート 1：興味語抽出実験

3.2.1 概要

アンケート 1 では，語の関連性を考慮することで興味語抽出の性能が向上するか検証を行う．各被験者について，提案手法により算出された興味度の上位 50 語と比較手法として採用した tf·idf 法による重み付けの重み上位 50 語を抽出する．これら全 100 語をランダムに並べ替え，どちらの手法によって抽出された語か区別できない状態にし，各被験者に提示する．被験者には提示された語をそれぞれ「興味がある」「興味がない」「どちらでもない」のいずれかに分類させた．なお，提案

表 3: 被験者別興味語抽出正答数

被験者 id	A	B	C	D	E	F	G	H	I	J
提案手法	**25**	**38**	**21**	19	25	**45**	**32**	36	25	**32**
比較手法	24	28	17	**21**	25	44	30	**38**	25	29
差分	**1**	**10**	**4**	-2	0	**1**	**2**	-2	0	**3**

手法及び比較手法の両方で抽出された語については一度のみ提示する．このような語は，最大の被験者で 35 個，最小の被験者で 23 個，平均で 29.1 個存在した．

3.2.2 結果

提案手法及び比較手法により抽出した各 50 語に対し，被験者が「興味がある」と答えた語の数を表 3 に示す．表中の差分は，被験者が「興味がある」と答えた語のうち提案手法により抽出された語の数から比較手法により抽出された語の数を引いた数である．つまり，差分が正の値であれば提案手法のほうが多くの興味語を抽出できたと言える．表 3 に示したとおり，提案手法のほうが多くの興味語を抽出できた被験者は 6 人，提案手法と比較手法で同数だった被験者が 2 人，比較手法のほうが多かった被験者が 2 人であった．つまり，過半数の被験者において興味語抽出性能の向上が見られており，語の関連性を考慮してユーザの嗜好を分析する手法の有効性が確認された．

性能の改善が見られなかった被験者について考察する．性能が低下した原因として，本稿では関連語辞書の構築対照語を各被験者についてそれぞれの重要度上位 150 語のみに限定したことが考えられる．この条件下では，語の関連性が考慮されるのは重要度上位 150 語のみとなり，150 位より下位の語は被験者が興味をもっている語であっても関連性の考慮がなされない．関連語辞書の数が少ない場合，興味度は重要度からの影響を大きく受けることになる．つまり，被験者が興味を持っていない語に大きな重要度が与えられていた場合，興味度も大きな値を取る場合が多くなる．逆に，被験者が興味を持っている語に大きな重要度が与えれていなければ，下位の関連語の影響力が発揮されず，興味度が適切に算出されない場合が存在する．性能の低下した被験者はこのようなことが起きていたのではないかと考えられる．したがって，関連語辞書の構築範囲を被験者の投稿に含まれる全ての語にまで拡大すれば，本稿で提案した手法でも性能を向上させることができると考える．

3.3 アンケート 2：未投稿語評価

3.3.1 概要

アンケート 2 では，未投稿語について，語の関連性を考慮することで正しく抽出が行えるか検証する．各被験者に対し，アンケート 1 で提示された語の中から自身の興味が強い上位 10 語を選択させた．選択された上位 10 語に未投稿語が含まれるかどうか調査する．

3.3.2 結果

アンケート 2 で選択された 10 語の構成を表 4 に示す．未投稿語はユーザの投稿に含まれなかった語，上昇語は提案手法によってのみ抽出された語を指す．表 4 に示されたとように，興味上位 10 語に未投稿語を含めた被験者は 10 名中 7 名であった．ただし，被験者 J はアンケート 1 において未投稿語が提示されていなかった．多くの被験者で未投稿であるが被験者が興味を持っている語を抽出することに成功しており，語の関連性を考慮することで tf·idf 法ではできない未投稿だが興味のある語の抽出が可能となることが示された．

表 4: 興味上位 10 語の構成

被験者 id	A	B	C	D	E	F	G	H	I	J
未投稿語	**1**	**3**	**1**	**1**	0	0	**1**	**1**	**1**	0
上昇語	1	2	3	1	1	3	2	3	2	3

4. おわりに

本稿では，Twitter から語の関連性を分析し，その情報を基にユーザの嗜好を表す興味語を抽出する手法を提案した．ユーザ評価実験の結果，語の関連性を考慮することで tf·idf 法に比べ興味語抽出性能の向上が確認された．また，未投稿であるユーザの興味語を抽出することが可能であることも確認した．今後の課題として，本稿で重要度の上位 150 語に限定した関連語辞書の構築対象語をすべての語に拡大することが挙げられる．これにより性能の向上が期待される．また，関連語辞書をリアルタイムに更新することでその変化を詳しく検証することも必要である．今後は，本システムにより抽出されたユーザの興味語をもとにした情報推薦システムの構築を目指したい．

参考文献

[岩井 11] 岩井 一晃, 鈴木 優, 石川 佳治 : マイクロブログにおける関連語の自動抽出, 全国大会講演論文集, Vol. 2011, No. 1, pp. 679–681 (2011)

[宮城 09] 宮城 良征, 當間 愛晃, 遠藤 聡志 : 日本語オントロジー辞書システム Ontolopedia の構築と興味抽出手法への応用検討, 知能と情報 : 日本知能情報ファジィ学会誌 : journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 21, No. 5, pp. 815–826 (2009)

[桑原 09] 桑原 雄, 稲垣 陽一, 草野 奉章, 中島 伸介, 張 建偉 : マイクロブログを対象としたユーザ特性分析に基づく類似ユーザの発見および推薦方式, 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2009, No. 18, pp. 1–3 (2009)

[後藤 10] 後藤 清豪, 高田 秀志 : ソーシャルメディア上での行動に基づく「意外な情報」の提供者になり得る人物の推薦手法, 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2010, No. 41, pp. 1–8 (2010)

[早川 11] 早川 豪, 岡部 誠, 尾内 理紀夫 : Twitter を利用したソーシャルニュース記事推薦システム, 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2011, No. 16, pp. 1–4 (2011)

[天野 07] 天野 真家, 石崎 俊, 宇津呂 武仁, 成田 真澄 : 自然言語処理, オーム社 (2007)

[和泉 11] 和泉 諒, 西山 裕之 : 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, 全国大会講演論文集, Vol. 2011, No. 1, pp. 681–683 (2011)

[齋藤 11] 齋藤 準樹, 湯川 高志 : ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価, 情報処理学会研究報告. DD, [デジタル・ドキュメント], Vol. 2011, No. 2, pp. 1–8 (2011)