Vol.2009-DBS-149 No.18
2009/11/21

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

桑原 雄[†1] 稲垣陽一[†1] 草野奉章[†1]
中島伸介[†2] 張建偉[†2]

近年，ブログや SNS 等，ユーザの生活体験が直接反映されたコンテンツが数多く配信されるようになってきている．我々は，これらのコンテンツを解析することで，ユーザの生活体験に基づいたシソーラスを半自動的に作成するシステムの研究開発を行っている．本稿では，このシステムを用いて作成したシソーラスを用いて，マイクロブログ上でユーザが発信した情報を分析し，特徴的なトピックやそれに対する感情を抽出することで，興味，志向が類似したユーザを発見する方法，及び類似ユーザを推薦する手法を提案する．

キーワード Web マイニング，テキスト解析，マイクロブログ

User Recommendation Method For Micro-Blogging Services Based on User Posting Analysis

Yu Kuwabara,[†1] Yoichi Inagaki,[†1]
Tomoaki Kusano,[†1] Shinsuke Nakajima[†2]
and Jianwei Zhang[†2]

Recently, the internet has experienced an explosion of personal media from social media sites such as blogs, Facebook, Twitter and so on. Much of this persona media describes a person's life experiences and events, ranging from the mundane to the intriguing. As it is such, we are developing a life experience thesaurus system which can automatically recognize the life experiences based upon textual analysis. In this paper, we describe the use of this thesaurus in our micro-blogging user recommendation system. This system analyzes user postings, extracting the topics and sentiments, and lastly recommends similar users.

Keywords: web mining, text analysis, micro-blogging

1. はじめに

近年，ブログや SNS 等，CGM（Consumer Generated Media）と呼ばれるコンテンツが多数配信されるようになっている．我々は，このようなコンテンツを解析することで，ユーザの生活シーンや体験，感情に基づいた言葉の意味付けやその変化を抽出し，半自動的にシソーラス化するシステムの研究開発を行っている．

本稿では，現在急速に普及しているマイクロブログ上のコンテンツを解析し，共通の話題を発信しているユーザを発見し，推薦する手法を提案する．

ユーザ間の類似性分析やユーザへの情報推薦の関連研究としては 1)，2) などがある．これらの研究はハイパーリンクやトラックバックを利用している．我々の手法では，ユーザの発信内容に基づき，類似性を判定する．また，単に共通の話題というだけではなく，その話題に対してどのような感情を持っているかを考慮した判定手法についても検討する．

2. 提案手法

2.1 ユーザ特性の分析

はじめに，ユーザの個々の投稿を解析する．解析には，我々が研究開発を行っている「生活体験シソーラス・システム LETS （Life Experience Thesaurus System）」を用いる．

LETS は，ブログやニュースなどのテキスト内で表現される生活体験を人手で体系的に分類，整理したシソーラスと，シソーラスに登録されているカテゴリの周辺語彙から自動生成される重み付きの共起語辞書，およびシソーラスと共起語辞書を用いて，入力テキストを分類するシステムで構成されており，現在 14,000 強のカテゴリが登録されている．このシステムを用いてマイクロブログの投稿を解析し，

- 投稿内で語られているトピックとスコア
- 投稿から読み取れる感情とスコア

を抽出する．ここでのスコアとは，その投稿がトピックおよび感情に対してどれくらい深く述べているかを示す値である．文章を実際に解析した例を**表 1** に示す．

ユーザの全ての投稿に対して，トピックと感情の抽出を行い，抽出された各トピックに対

†1 株式会社きざしカンパニー
kizasi Company,inc
†2 京都産業大学
Kyoto Sangyo University

表 1 文章の解析例

入力文	トピック (スコア s_i)	感情 (スコア f_j)
猫を飼っています。可愛いです。	猫 (26.78)	かわいい (27.09)
テニスは疲れるけど楽しい。	テニス (24.46)	楽しい (22.53)
		疲れた (22.00)

して，スコアの合計値を計算する．ここで，ユーザ A における，トピック t に対するスコアを $score(A,t)$ とすると，

$$score(A,t) = \sum_{i=1}^{n} s_i \times w_t \tag{1}$$

と表すことができる．ただし，n はユーザ A の投稿総数，s_i は投稿 i におけるトピック t のスコア，w_t はトピック t の珍しさを用いた重みで，

$$w_t = 1 + \log_{10}\left(\frac{\text{解析対象とするユーザ数}}{\text{トピック t に対するスコアが 0 でないユーザ数}}\right) \tag{2}$$

として表す．また，トピック t に対する感情 e のスコアを $score(A,t,e)$ とすると，

$$score(A,t,e) = \sum_{j=1}^{m} f_j \tag{3}$$

と表すことができる．ただし，m はトピック t が出現した投稿数，f_j は投稿 j におけるトピック t に対する感情 e のスコアである．ここで，投稿から読み取れる感情が必ずしもトピックに対する感情とは限らないことが問題となるが，同じトピックが抽出された複数の投稿に対する感情のスコアを合計することで，トピックに対するユーザの感情を推測できると考えられる．例えば，「猫」というトピックが抽出されたある投稿に対して「好き」「嫌い」という相反する感情が抽出された場合でも，それ以外の「猫」が抽出された投稿に対して「好き」ばかりが抽出されていればユーザの「猫」に対する感情としては「好き」が妥当であると推測できる．$score(A,t,e)$ は，トピック t をユーザの特性として採用するかしないかの判定に用いる．

全てのトピックに対して $score(A,t)$ および $score(A,t,e)$ を計算し，$score(A,t)$ が高いもの上位 10 件を用いてユーザの特性を表現する．このときユーザの特性は，$score(A,t)$ を要素とする 10 次元ベクトルとして表される．これをユーザの特性ベクトルとする．ただし，最も値が大きい $score(A,t,e)$ の e が否定的な感情であったトピックは，特性ベクトルとして採用しない．これは，例えば猫が嫌いなユーザに対して「猫」を特性ベクトルとして付与するのは意味がないと考えられるためである．

2.2 類似度の計算

類似度の計算には，特性ベクトルのコサイン類似度を用いる．ここで，ユーザ A の特性ベクトルを C_A，ユーザ B の特性を C_B，ユーザ A と ユーザ B の類似度を $sim(A,B)$ とすると，

$$sim(A,B) = \frac{C_A \cdot C_B}{|C_A| \times |C_B|} \tag{4}$$

と表すことができる．ただし，$C_A \cdot C_B$ は C_A と C_B の内積，$|C_A|$，$|C_B|$ はそれぞれ C_A，C_B の長さであり，$sim(A,B)$ の取り得る値は $0 \leq sim(A,B) \leq 1$ である．なお，C_A にしか存在しない要素については，C_B での値は 0 として計算する．逆も同様である．

3. 評　　価

本稿では Twitter[3] を対象に実験データを作成した．Twitter から約 3,500 ユーザの投稿を収集し，特性ベクトルを作成した上で，ユーザ A，B に対してそれぞれ類似度の高いユーザ 5 名の投稿内容を読み，類似ユーザと判断できるかを評価した．結果，ユーザ A に対しては 4 名，ユーザ B に関しては 1 名が類似ユーザと判断できた．

表 2 に，ユーザ A に対して類似度が高く，かつ類似ユーザと判断できたユーザの特性ベクトルの例を示す．両者に共通している特性ベクトルの要素は「登山」「携帯電話」「iPhone」「仕事」「家族」だが，いずれのユーザにも趣味の登山に関する投稿や iPhone に関する投稿が多く見られ，類似ユーザとして妥当であった．しかし，「仕事」や「家族」に関しては，例えば “そろそろ仕事を始めよう” などといった，ユーザの特性を表現しているとはいえない投稿が多く，ユーザ特性の分析精度に関しては改善の必要があるといえる．

表 3 に，ユーザ B に対して，類似判定に失敗していた例を示す．両者に共通している特性ベクトルの要素は「母」「家族」「パソコン」「学校」だが，先述した「仕事」の例と同様に，いずれもユーザの特性を示すトピックとしては不適切であり，投稿内容にも類似性はみられなかった．

4. 今後の課題

(1) 特性解析手法の改善

一般的すぎるトピックが特性として解析されることで，類似度の判定に失敗している失敗例

表 2 類似ユーザの特性ベクトルの例

ユーザ A の特性ベクトル		類似ユーザ X の特性ベクトル	
トピック	スコア ($score(A,t)$)	トピック	スコア ($score(X,t)$)
登山	313.313	登山	274.398
携帯電話	164.243	携帯電話	223.775
iPhone	83.970	iPhone	188.836
仕事	70.798	仕事	354.217
家族	69.260	家族	269.691
灰皿	201.247	酒	317.458
ベランダ	121.113	インターネット	269.364
読書	86.690	日本酒	219.435
ミネラルウオーター	86.391	社会	216.398
おでかけ	62.236	タイ	176.750

表 3 類似ユーザ判定の失敗例

ユーザ B の特性ベクトル		類似ユーザ Y の特性ベクトル	
トピック	スコア ($score(B,t)$)	トピック	スコア ($score(Y,t)$)
母	224.499	母	190.587
家族	179.286	家族	338.045
パソコン	147.729	パソコン	278.567
学校	89.316	学校	273.461
照明	147.318	釜飯	459.134
梅田望夫	137.907	ぬいぐるみ	437.593
家電	121.512	ドラゴンクエスト	266.123
カレー	68.339	恋愛	237.247
生活家電	66.443	もやし	195.189
堀江貴文	57.646	抹茶	185.520

が多く見られた．例えば毎朝出勤前に “会社に行きます” などと投稿してから出かけるユーザがみられたが，この投稿から抽出される「会社」というトピックはユーザの特性を反映しているとはいえないため，ユーザの特性ベクトルの解析手法を改善する必要がある．案として，特性として採用するトピックをある程度具体的なものに限定することを検討している．また，本稿では否定的な感情が強いトピックを特性から除外したが，例えば「サッカーの試合に負けてくやしい」などのように，興味の対象であるがゆえに否定的な感情が検出される場合もあるため，より有効な感情属性の使い方を検討する．

(2) **類似ユーザの定義の検討**

本稿では類似ユーザを，特性ベクトルのコサイン類似度によって定義したが，例えば全ユーザの中である 2 名だけが言及しているトピックがあったとすると，そのトピックは，スコアの大小によらず 2 者だけの類似点であると考えられる．情報推薦のための類似度として有用な指標を検討する必要がある．

(3) **マイクロブログ特有の情報の考慮**

Twitter におけるフォローは，ユーザの興味を直接表していると言える．既に，自分がフォローしているユーザがフォローしているユーザを推薦するサービス[4),5)] も存在するが，フォロー関係のつながりだけでなく，どのような理由でつながっているのかを解析することができれば，トピックとしては直接は現れない類似点を見つけられるのではないかと考えている．また，多くのユーザにフォローされているユーザは，それだけフォローする価値のあるユーザだとも考えられる．これらのようなマイクロブログ特有の情報も考慮することを検討している．

(4) **ユーザによる評価実験**

フォロー対象の候補として推薦されたユーザに興味をもつかどうかは，実際には推薦された本人にしか判断できない．そのため，実際に Twitter 上で動作するユーザ推薦システムを作成し，ユーザからのフィードバックを得て評価，改善を行う必要がある．

5. おわりに

本稿では，トピックに対する感情を考慮した類似ユーザの判定方法，およびそれに基づいた情報推薦方法について述べたが，検討すべき課題は多い．今後は実際にシステムを作成し，評価，改善を行っていく予定である．

参考文献

1) 古川忠延，松澤智史，松尾豊，内山幸樹，武田正之：Weblog におけるユーザの繋がりと閲覧行動の分析，電子情報通信学会論文誌，Vol. J88-B, No.7, pp.1258–1266 (2005).
2) 小原恭介，山田剛一，絹川博之，中川裕志：Blogger の嗜好を利用した協調フィルタリングによる Web 情報推薦システム，The 19th Annual Conference of the Japanese Society for Artificial Intelligence, 2C2–02, 2005.
3) Twitter, http://twitter.com/
4) Twubble, http://crazybob.org/twubble/
5) ふぉろわのふぉろわー, http://followernofollower.com/