

興味領域を考慮した Twitter アカウント推薦

久米 雄介¹ 打矢 隆弘¹ 内匠 逸¹

概要 : 近年, Social Networking Service(SNS) の爆発的な普及により, 社会的ネットワークを介したコミュニケーションが活性化している. 特に, Twitter と呼ばれるサービスが世界各国でシェアを伸ばしており, 社会で約 2 億人のユーザがこのサービスを利用している. Twitter ユーザはツイートと呼ばれる 140 字以内の短文を投稿し, 情報を発信する. また, 興味のあるユーザをフォローすることにより, 他のユーザのツイートをタイムライン上に表示させることで情報を取得する.

Twitter には, 膨大な数のユーザから自分の嗜好に合ったものを探し出すのは非常に手間がかかるという問題がある. 本研究では, Twitter における問題点を解決する為に, ユーザの興味領域と特徴語に基づく Twitter アカウント推薦を提案する.

Recommendation of Twitter account based on category of interest

YUSUKE KUME¹ TAKAHIRO UCHIYA¹ ICHI TAKUMI¹

1. はじめに

近年, Social Networking Service(SNS) の爆発的な普及により, 社会的ネットワークを介したコミュニケーションが活性化している. 特に, Twitter[1] と呼ばれるサービスが世界各国でシェアを伸ばしており, 社会で約 2 億人のユーザがこのサービスを利用している. Twitter ユーザはツイートと呼ばれる 140 字以内の短文を投稿し, 情報を発信する. また, 興味のあるユーザをフォローすることにより, 他のユーザのツイートをタイムライン上に表示させることで情報を取得する. Twitter ユーザは, 友人とのコミュニケーションや著名人へのアクセス, 情報の発信・取得等, 多様な目的に応じて Twitter を利用している.

Twitter において膨大な数のユーザからフォローユーザを探し出すには, 自分で Twitter アカウントを検索し, 検索結果から自分の嗜好に合ったユーザを選択しなければならない. しかし, 現在はこれらの作業全てを手動で行わなければならない為, ユーザに大きな負担が生じ, ユーザ間のコミュニケーションの活性化の大きな障害となっている. これらの問題点の解決の為に, 近年はユーザの嗜好に合った Twitter アカウントをシステムが自動的に推薦する「フォローユーザ推薦サービス」が提案されている.

従来のフォローユーザ推薦手法としては, ユーザのツイートから単語頻度 (TF : Term Frequency) ・逆文書頻度 (IDF : Inverse Document Frequency) のような従来のキーワードベースの指標で興味語を抽出し, 推薦を行っている. しかし, Twitter においてこの手法を用いると, 顔文字等の日常的にツイートに登場するが, 興味を示すキーワードでは無いものがノイズとなり, ユーザがどのような対象や領域に興味を持っているのか検出することが困難である.

そこで本研究ではスポーツ・音楽等のキーワードのカテゴリに着目する. 各キーワードがどのカテゴリに属するかを把握・管理することで, 顔文字等のどのカテゴリにも属さないキーワードを除外する. また, このカテゴリを一つの嗜好情報として取り扱い, 従来の推薦手法に加えることで, より利用者の嗜好に合った Twitter アカウント推薦の実現を目指す.

2. 背景要素

本研究の背景要素として, Twitter の各機能と Twitter からの嗜好抽出に用いられる tf・idf 法について述べる.

2.1 Twitter の機能

Twitter は Twitter, Inc. が運営するマイクロブログサービスである. Twitter ユーザは, 140 字以内の短文を投稿

¹ 名古屋工業大学工学研究科情報工学専攻

し、情報を発信する。主な Twitter の機能として以下の様なものがある。

ツイート (tweet)

ユーザから投稿される 140 字以内の短文であり、ツイートごとに固有の URL が割り当てられる。

タイムライン (timeline)

ツイートが時系列に並べて表示される画面を指す。

フォロー (follow)

他のユーザのツイートを自分で閲覧できるように登録する機能であり、アカウントのプロフィール画面に表示されるフォローボタンを押すことで実行できる。自分がフォローしているユーザをフォロイーと呼び、自分のことをフォローしているユーザをフォロワーと呼ぶ。

リプライ (reply)

他のユーザに宛てたツイートを指す。「@ユーザ名 ツイートしたい内容」の形式で投稿すると、そのユーザ宛ての返信として扱われる。自分宛のリプライは一覧画面で確認することができる。

リツイート (retweet)

周知を目的に他のユーザのツイートを再投稿する機能である。

お気に入り (favorites)

気に入ったツイートをお気に入りに登録する機能である。お気に入りに登録されたツイートはブックマークされいつでも閲覧することができる。

2.2 tf・idf 法

tf・idf 法は文章に出現するキーワード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利用されている。この手法では、文章を特徴付けるキーワードはその文章に多く登場し、他の文章にはあまり登場しないようなものであるという考えに基づき、文章を特徴付けるキーワードである程大きな重み付けが行われる。近年は、ブログや Twitter 等の SNS に適応することで、その利用者がどのような嗜好を持っているのかを解析するのに利用されている。

tf・idf 法によってキーワード i に与えられる重み w_i はユーザの投稿における出現頻度 $tf_{i,u}$ 、ユーザの集合 U の総数を $|U|$ 、 U のうちキーワード i を含む投稿を行ったユーザ数を df_i とすると以下の計算式で表現できる。 w_i はユーザの投稿に多く現れ、他のユーザの投稿にあまり現れないキーワードほど高い値を示し、ユーザの特徴を示すキーワードとなる。

$$w_i = tf_i \cdot \log \frac{|U|}{df_i} \quad (1)$$

しかし、Twitter 等には顔文字等のユーザの嗜好情報を

示すものではないキーワードが登場する。このようなキーワードは表現方法がユーザごとに異なり、且つツイートに多く登場する為、高い重み付け値を示す。その為、このようなキーワードは嗜好の抽出においてはノイズとなる為、Twitter において従来の tf・idf 法は実用的ではない。

2.3 関連研究

関連研究として、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 [2] がある。この研究は従来の tf・idf 法によるフォローユーザ推薦手法の改善の為に、ソーシャルブックマークサービス (Social BookMarking Service:SBM) と呼ばれるウェブサービスで利用されている「タグ」に着目している。タグとはユーザが自身のブックマークに付与する短い単語や注釈のことである。この情報には一般的な単語以外に、そのブックマークの内容に関連したものであれば、ブックマークが作成された時点での新語や略語も豊富に含まれている為、従来の類語辞書では対応できなかった単語の検出が可能になる。また、ユーザが用いた単語に対し関連語まで考慮に入れた重み付けを行うことで、より興味や嗜好の度合いを強調して特徴的な語を抽出することができると考察している。

この研究は従来の tf・idf 法における嗜好抽出のノイズに着目し、それに対する解決策を提示している。また、嗜好抽出に基づくフォローユーザ推薦手法を提示している為、本研究と類似している。しかし、この手法の評価実験では、SBM を用いて新語に対応することは実現できているが、上記で示した嗜好抽出のノイズに関する問題の根本的な解決には至っていない。また、SBM を利用しているすべてのユーザが入力したタグを用いている為、意図せずとも被推薦ユーザと直接関わりのないユーザの嗜好情報も用いている。その情報が嗜好情報の抽出においてノイズになる可能性がある。その問題点の改善の為に、被推薦ユーザが入力したタグのみを用いて嗜好抽出を行う必要がある。しかし、提案手法を利用するには被推薦ユーザが SBM を利用していなければならない。その為、システムを利用できるユーザ数に制限があり実用的ではない。

3. 提案手法

本研究では上記で指摘した問題点の解消の為に、従来のフォローユーザ推薦の嗜好抽出に対し、「興味領域」という指標を加える。これにより嗜好抽出においてノイズとなるキーワードを除外し、高精度且つ実用性が高いフォローユーザ推薦を目指す。

興味領域とはスポーツや音楽等のカテゴリに対する興味度合を示す指標である。例えばスポーツに対する興味領域の値が大きい場合はスポーツに対する関心が大きいといえる。また、逆に興味領域の値が低い場合は関心が低いとい

える．この指標を従来の $tf \cdot idf$ 法に加えることで，より高精度な嗜好抽出が可能である．

興味領域を計算する為に，各キーワード毎にカテゴリを取得する．その為，顔文字等のユーザの嗜好情報を示すものでないキーワードをどのカテゴリにも属さない単語として扱うことで上記で示したノイズの除去に有効であるといえる．カテゴリを取得する際には，はてなキーワード自動リンク API[3] を用いる．この API はサービスを利用していないユーザでも利用できる為，従来手法と比べ，提案手法は実用性の高い推薦手法であるといえる．

3.1 提案手法の概要

提案システムは以下の 3 要素で構成される (図 1)．

- 興味領域抽出機能
- 特徴語抽出機能
- 推薦フォローユーザ取得機能

興味領域抽出機能ではユーザのツイートを解析し，興味領域を取得する．特徴語抽出機能は tf と興味領域を用いてユーザの特徴を示すキーワード (以下特徴語) を取得する．推薦フォローユーザ取得機能では，興味領域と特徴語を基にユーザに提示するフォローユーザを選別する．提案手法の処理の流れは以下になる．

- (1) 興味領域抽出機能を用いてユーザが興味を持っているカテゴリを抽出
- (2) 特徴語抽出機能を用いてユーザのツイートの特徴語を抽出
- (3) 抽出した特徴語をツイートしているユーザを検索
- (4) 推薦フォローユーザ取得機能を用いて推薦アカウントを選抜
- (5) 推薦結果をユーザに掲示

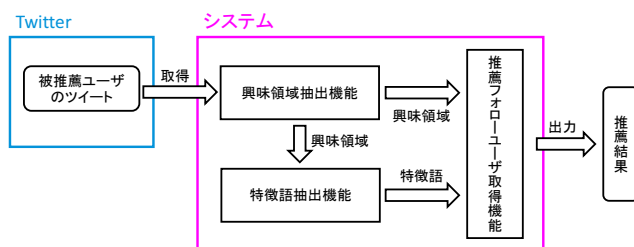


図 1 提案手法の概要図

3.2 興味領域抽出機能

興味領域抽出機能はユーザのツイートを解析し，ユーザが興味を持っているカテゴリ (スポーツ・音楽等) を抽出する．これにより，ユーザの大まかな嗜好を取り出すことができる．興味領域は興味のあるカテゴリならばユーザのツ

イートに表示される回数が多いという仮説に基づき計算を行っている．例えば，スポーツに関するキーワードを多く含むツイートをしているユーザならばスポーツに対する関心が大きいと考察する．

この機能はユーザのツイートを形態素解析し，キーワードごとにカテゴリを取得する．そして，カテゴリごとにキーワードの数をカウントし，カテゴリを取得したキーワードの総数に対する割合として算出する．例えば，カテゴリを抽出したキーワードが 5 つあり，そのうちカテゴリが SPORTS のキーワードが 2 つあった場合，ユーザの SPORTS に対する興味領域は $2/5 = 0.4$ となる (図 2)．

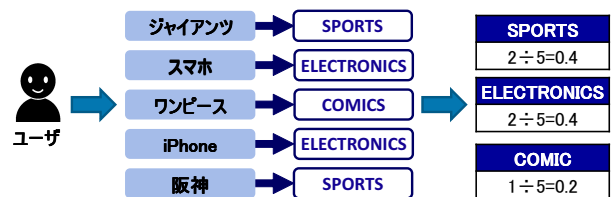


図 2 興味領域抽出機能の動作例

3.3 特徴語抽出機能

特徴語抽出機能はユーザのツイートの特徴語を算出する機能である．特徴語抽出には，興味領域と tf を用いる．本手法の被推薦ユーザのツイート解析の段階ではまだ推薦候補のユーザの検索を行っていない為，他のユーザのツイートを取得できない．その為，他のユーザのツイートを解析必要がある idf は本研究では利用しない．

キーワード i と文書 j についての単語頻度 $tf_{i,j}$ は文書 j におけるキーワード i の出現数 $n_{i,j}$ を用いると以下のよう

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2)$$

そして，各キーワードの tf と興味領域に対し，平均と分散の和に応じて正規化を行う．そして，最終的に tf に興味領域の値に応じて重み付けを行ったものをキーワードに対する評価値とし，その値の大きさが上位のものをユーザの特徴語とする (図 3)．

平均と分散の和に応じて tf と興味領域を正規化			tf と興味領域の合計値が上位のキーワードを抽出	
キーワード	tf	興味領域	キーワード	評価値
ジャイアンツ	0.2 → 0.50	0.4 → 1.11	ジャイアンツ	$0.50 + 1.11 = 1.61$
スマホ	0.3 → 0.75	0.4 → 1.11	スマホ	$0.75 + 1.11 = 1.86$
ワンピース	0.5 → 1.26	0.2 → 0.55	ワンピース	$1.26 + 0.55 = 1.78$
iPhone	0.1 → 0.25	0.4 → 1.11	iPhone	$0.25 + 1.11 = 1.36$
阪神	0.1 → 0.25	0.4 → 1.11	阪神	$0.25 + 1.11 = 1.36$

図 3 特徴語抽出機能の動作例

3.4 推薦フォローユーザ取得機能

推薦フォローユーザ取得機能は抽出した特徴語を用いて、ユーザの嗜好に合ったアカウントを推薦する機能である。この機能では、被推薦ユーザと同じカテゴリに対する興味領域が高いユーザならば嗜好が一致する可能性が高いという仮説に基づき推薦を行っている。抽出した特徴語を用いて、今現在そのキーワードをツイートしているアカウントを推薦アカウントの候補とする。各推薦候補アカウントに対し、興味領域抽出機能を用いて興味領域を算出し、その興味領域とそのアカウントを検索した特徴語に対する評価値の積をアカウントの評価値として扱う。その評価値の大きさが上位のものをユーザに対し推薦する(図4)。

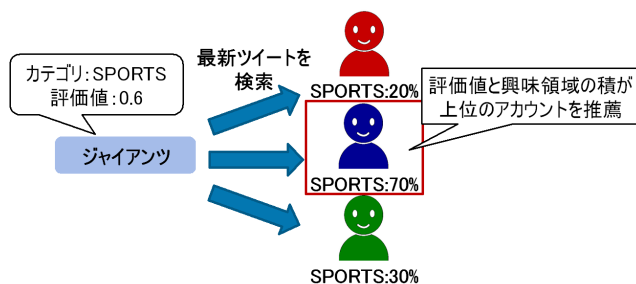


図4 推薦フォローユーザ取得機能

4. 提案手法の実装

提案手法を Twitter API の Java ラッパである Twitter4j[4]、形態素解析を行う Mecab[5] の機能を利用できる SlotLib[6]、及び、キーワードごとにカテゴリを取得するのははてなキーワード自動リンク API、計3つのライブラリを用いて実装を行った。

4.1 はてなキーワード自動リンク API

はてなキーワード自動リンク API は任意のテキストを入力すると、そのテキストからはてなキーワードに登録されているキーワードを抽出し、それに関する情報を返信する API である。この情報の中に各キーワードのカテゴリが含まれており、それを取り出し興味領域の計算に利用している。はてなキーワードのカテゴリは図5に示す20種類があり、提案手法では「一般」・「はてな」・「はてなダイアリークラブ」の3つのカテゴリは除外し、残りの17種類のカテゴリを用いている。この API は会員登録等が必要が無い事に加え、SBM のように行動履歴を必要としない為、実用性の高いフォローユーザ推薦を実現という本研究の目的に適しているといえる。

4.2 Mecab

Mecab とは京都大学の研究チームで開発されたオープンソース形態素解析エンジンである。Mecab は言語・辞書・

キーワードカテゴリ			
一般	読書	音楽	映画
ウェブ	コンピュータ	動物	テレビ
アニメ	食	スポーツ	ゲーム
マンガ	アイドル	社会	地理
アート	サイエンス	はてな	はてなダイアリークラブ

図5 はてなキーワード自動リンク API のカテゴリ

コーパスに依存しない汎用的な設計がなされており、利用者が辞書・コーパス・品詞体系等を用意することで新規語等の現代日本語以外の言語でもサポートが可能な構造を有している。以上の点を考慮し、提案手法の実装に適した形態素解析エンジンであると考え、利用した。

4.3 推薦結果の表示

以下に提案手法によって表示される推薦結果を表示する。Recommend_user は提案手法によって推薦されたユーザを示し、userName はそのアカウントの TwitterID を示している。関連キーワードはそのアカウントを検索した際に利用したキーワードを示している。

推薦結果

```
Recommend_user[userName:okadashu  
関連キーワード:大学院生]  
Recommend_user[userName:ijimemaps  
関連キーワード:大学院生]  
Recommend_user[userName:abfly_love  
関連キーワード:情報工学]  
Recommend_user[userName:kanami_gsh  
関連キーワード:ジャイアンツ]  
Recommend_user[userName:Shinyaa_  
関連キーワード:大学院生]  
Recommend_user[userName:nakagawadesuyoo  
関連キーワード:ツイート]  
Recommend_user[userName:kuroto815  
関連キーワード:ツイート]  
Recommend_user[userName:J_POP_NEWS  
関連キーワード:宇多田ヒカル]  
Recommend_user[userName:fink13ciel  
関連キーワード:Arc]  
Recommend_user[userName:d0c_0d  
関連キーワード:ツイート]
```

5. 評価実験

各機能の抽出結果に対して、アンケート調査を行った。被推薦ユーザの Twitter アカウントから最新 500 件のツイートを取得し、興味領域と特徴語を 10 件抽出した。抽出したキーワードごとに 10 件の最新ツイートを検索し、そのキーワードをツイートしているアカウントを取り出して推薦を行った。アンケートを実施したユーザは 6 人であ

り、抽出結果が自分の嗜好に合ったものか○×で評価してもらい提案手法の有効性を検討した。

5.1 評価指標

各機能の抽出結果を用いし、評価指標として適合率を用いて評価実験を行った。適合率とは抽出結果すべてに対するユーザの嗜好に合ったものの割合を示す指標であり、ユーザ u の嗜好にあったアイテムの集合 $hits_u$ と推薦されたアイテムの集合 $|recset_u|$ を用いると適合率 P_u は以下のような計算式で表現できる。

$$P_u = \frac{|hits_u|}{|recset_u|} \quad (3)$$

5.2 比較対象：Twitter 公式推薦システム「おすすめユーザ」

Twitter にはユーザに対しフォローすべきユーザを推薦する「おすすめユーザ」という機能がある。おすすめユーザは被推薦ユーザの過去の行動履歴をもとに選抜され、図 6 のように表示される。推薦されるおすすめユーザは 2 種類に大別される。このシステムは Twitter ユーザが一番利用する可能性が高いものである為、比較対象とした。

著名人・有名企業のアカウント

Twitter を始めたばかりの初心者ほどのユーザをフォローして良いか分からない場合が多い為、芸能やスポーツなどのジャンルごとに著名人・有名企業のアカウントを列举して推薦を行う。

フォロイーやフォロワーがフォローしているユーザ

Twitter ユーザのフォロー情報などに基づきユーザを推薦する手法である。フォロイーやフォロワーの多くがフォローを行なっているユーザが優先的に推薦されるため、フォロワーが多い有名人が推薦される可能性が高くなる。



図 6 おすすめユーザ

5.3 実験結果

興味領域抽出機能は興味領域の大きさの上位 5 件、特徴語抽出機能は評価値の大きさ上位 5 件と上位 10 件、推薦結果は評価値の大きさの上位 10 件をユーザに 3 回提示しその平均値と標準偏差を評価した。また、推薦結果の推薦結果の評価の比較として、Twitter 公式おすすめユーザによる推薦結果の上位 10 件の評価を平行して行った。興味領域推薦機能と特徴語抽出機能の評価結果を表 1 に示す。また、提案手法による推薦結果と Twitter 公式おすすめユーザの推薦結果の適合率を比較し、高い適合率を記録したアカウントの数を表 2 に示す。

表 1 より興味領域抽出機能では 0.9 という非常に高い適合率を記録している。また、特徴語抽出機能では上位 10 件では 0.66, 上位 5 件では 0.68 を記録しており、概ね良好な結果を得ることができた。また、標準偏差は興味領域抽出機能では 0.15, 特徴語抽出機能では上位 5 件では 0.42, 上位 10 件では 0.38 を記録している。

推薦結果の適合率の比較結果は Twitter 公式おすすめユーザでは 5, 提案手法では 1 となり、提案手法の方が低い適合率を示したアカウントの数が多いことが分かる。

表 1 興味領域・特徴語の適合率

	適合率 (平均)	適合率 (標準偏差)
興味領域	0.9	0.15
特徴語 (上位 5 件)	0.68	0.42
特徴語 (上位 10 件)	0.66	0.38

表 2 実験結果 (推薦結果)

	Twitter 公式	提案手法
比較結果	5	1

5.4 被験者の意見

以下に提案手法を利用した被験者の感想を示す。興味領域や特徴語に関しては比較的良好な感想を得られたが、推薦結果に関しては不満点が多かった。特に、Twitter ユーザの推薦において嗜好の一致だけでなく、ツイート数やフォロワー数も重視している利用者からの不満が多かった。

- 特徴語は適切に抽出できているがそれを活かしてきいていない
- ツイート・フォロワー数が少ないものや商品の宣伝等としているだけのアカウントはフォローしたくない
- 似たようなアカウントばかり推薦結果に含まれている

5.5 考察

実験結果より、興味領域抽出機能と特徴語抽出機能の適合率は比較的良好な結果が得られたといえる。しかし、推薦フォローユーザ取得機能では Twitter 公式おすすめユー

ザと比べ、低い適合率を記録した。これは推薦フォローユーザ取得機能に問題があると考察する。アンケート調査より、興味領域と特徴語に対する評価が高いことを考慮すると、ツイートから取得した嗜好情報をフォローユーザ推薦に活かしてきれていないといえる。

推薦フォローユーザ推薦機能の改善点としては以下のようなものがある。

推薦フォローユーザの偏りの防止

被験者より、自分と嗜好が類似しているユーザは推薦結果に含まれているという意見があった。しかし、似たようなユーザばかり推薦されてしまい、推薦結果としての魅力に欠けるという意見があった。これは特徴語抽出機能の中で興味領域に値に応じて重み付けを行っている為、興味領域の値が大きいカテゴリの属するキーワードが特徴語として抽出されやすくなっている点が原因であるといえる。この問題点は特徴語と推薦フォローユーザをカテゴリごとに管理し、推薦結果に偏りが無いように調整することで解消することができると考察する。

ツイート数・フォロワー数を考慮した推薦

被験者より、Twitter の初心者等のツイート数・フォロワー数が少ないユーザにあまり魅力を感じないという意見があった。興味領域はカテゴリごとにキーワードの数をカウントし、カテゴリを取得したキーワードの総数に対する割合である為、キーワードの総数が少ない初心者等が推薦結果に含まれやすい傾向がある。キーワードの総数はユーザごとに異なる為、その総数を興味領域に反映する必要があると考察する。

企業アカウント・スパムアカウントの検出

Twitter で商品の宣伝等を行っているユーザは同じツイートを繰り返している為、一つのカテゴリに対する興味領域の値が極端に大きくなる傾向があり、推薦結果に含まれやすくなっている。この問題点は複数のカテゴリの興味領域を用いて推薦を行う手法や極端に興味領域が大きいユーザを推薦結果に含まない手法を用いれば対処できると考察する。

6. まとめ

従来のフォローユーザ推薦システムの改善として興味領域を考慮した Twitter フォローユーザ推薦を提案した。今後は推薦フォローユーザ機能の改善として、推薦フォローユーザの偏りの防止や企業アカウント・スパムアカウントの検出に重点を置き、ユーザにとって有効なフォローユーザ推薦手法の確立を目指す。

参考文献

- [1] “Twitter”, <https://twitter.com/>.
- [2] 斎藤準樹 他, “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, 2011-DD-80(2), 1-8, 2011-03-21.

- [3] “はてなキーワード自動リンク API”, <http://developer.hatena.ne.jp/ja/documents/keyword/apis/rest>.
- [4] “Twitter4j”, <http://twitter4j.org/ja/index.html>.
- [5] “Mecab”, <http://sourceforge.jp/projects/mecab/>.
- [6] “SlothLib Wiki”, <http://www.dl.kuis.kyoto-u.ac.jp/slothlib/?FrontPage>.