Vol.2015-ICS-179 No.1
2015/3/20

興味領域を考慮した Twitterユーザ推薦手法の提案と評価

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おり，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だ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，ユーザの興味領域を考慮した Twitter ユーザ推薦を提案する．

キーワード：情報推薦システム，Social Networking Service, Twitter

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，世界で約 2 億 7 千万人 (2015 年 1 月時点)，日本では約 2200 万人 (2013 年 10 月時点) 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 において膨大な数のユーザからフォローユーザを探しだす為には，自分で Twitter ユーザを検索し，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 ユーザ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の代表的なものとしては，コンテンツベース方式を用いた Twitter フォローユーザ推薦がある．この手法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) 等の指標を用いてユーザの興味を示すキーワードを抽出し，推薦を行う．しかし，Twitter では各キーワードは略して投稿されることが多い為，同じ意味を持つキーワードでも表現方法が統一されていない場合が多い．その為，ユーザの特徴が分かりにくくなり，特徴を示すキーワードとノイズとなるキーワードが差別化されず，特徴語抽出の精度の低下を引き起こしてしまう．また，それに伴い，フォローユーザ推薦の精度も低下してしまう．

本研究ではこの問題点を解決する為に，各キーワードをカテゴリ (スポーツや音楽等) 毎に管理する．また，従来のコンテンツベース方式で用いられる指標に各カテゴリに対する興味度合を示す「興味領域」という指標を加える．これらを行う事で，特徴語抽出の精度を向上させる事ができ，より嗜好に合ったユーザを推薦する事が可能になる．

2. Twitter

Twitter とはツイートと呼ばれる短文を投稿できる情報サービスであり，Twitter 社によって提供されている．各ユーザが自分のアカウントを作成することでサービスを利用することができる．画面上部のボックスに 140 字以内で文章を入力し，ツイートボタンを押すことで投稿が完了する．また，Twitter での自分専用のページ「ホーム」には，自分の投稿とあらかじめ登録したユーザの投稿が時系列順に表示される．各ユーザが自分の近況等を投稿し，他のユーザがそれに対して話しかける事でコミュニケーションが生まれる．

[1] 名古屋工業大学工学研究科情報工学専攻

2.1 特長

Twitter の大きな特長は，膨大な数のユーザに迅速に情報を発信・共有・拡散できる点がある．Twitter では 140 字以内の文章を記入し，ツイートボタンを押すだけで容易に情報を発信できる．また，各ユーザのホーム画面でフォローボタンを押すだけで，他のユーザのツイートを閲覧できる為，情報の共有が容易である．更に，他のユーザのツイートをリツイートする事で，そのユーザをフォローしていないユーザもそのツイートを閲覧できる為，情報の拡散を迅速に行う事ができる．これらの特徴を活用すれば，容易に多くのユーザとコミュニケーションを取ることが可能になる．

2.2 問題点

Twitter の利用者数は世界で約 2 億 7 千万人，日本では約 2200 万人と言われており，これらの中から興味があるユーザを手動で検索し，それらの中からフォローするユーザを見つけるには複数の作業が必要となり非常に手間がかかる．ユーザがフォローユーザを自発的に検索する方法として，サーチ機能を利用し，フォローするユーザを検索する方法が挙げられる．この方法には，以下のような作業を行う必要がある．

(1) 自分がフォローしたくなるようなユーザを見つけやすい検索キーワードを選択
(2) サーチ機能を利用し，検索ワードを含んだツイートをしているユーザを検索
(3) 検索結果に登場するユーザのツイート等を確認
(4) そのユーザをフォローするか否か判断

これらの作業は特に Twitter の初心者にとっては負担が大きい為，Twitter 上でのコミュニティの形成の大きな障壁となっている．その結果，インターネットを介したコミュニケーションや情報収集を効率的に行えなくなり，Twitter の有用性を発揮できなくなる．この問題点を解決する為に，近年はユーザの嗜好に合ったユーザ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．

3. フォローユーザ推薦サービス

フォローユーザ推薦サービスとは Twitter ユーザのツイートやフォロー履歴等の行動履歴を基に，嗜好に合ったユーザを推薦するシステムである．システムがユーザの嗜好を取得して，推薦を行う為，各ユーザはフォローするユーザを検索せずに発見することができる．これにより，各ユーザがフォローする上での手間を省く事ができる為，コミュニティの形成への障壁が低くなる．

3.1 コンテンツベース方式を用いたフォローユーザ推薦

コミュニティ形成を目的としたフォローユーザ推薦手法の代表的なものとして，コンテンツベース方式のフォローユーザ推薦手法がある．この手法は tf・idf 法等を用いてユーザのツイート等の行動履歴からユーザの嗜好を取り出す手法である．tf・idf 法と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．

tf・idf 法によってキーワード i に与えられる重み付け値 (以下 tf・idf 値)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する．

$$w_i = tf_i \cdot log\frac{|U|}{df_i} \tag{1}$$

コンテンツベース方式の Twitter ユーザ推薦は tf・idf 法をツイートに適応し，キーワード毎に重み付け値を算出する．そして，それらの重み付け値を要素とする嗜好ベクトルを作成し，それらを他のユーザと比較し，類似度を算出する．最終的に類似度の値が上位のものを推薦する (図 1)．類似度計算にはコサイン類似度を用いる事が多い．コサイン類似度とは各ユーザの嗜好ベクトルを比較し，その内積を算出し，類似度として扱う．この値が大きい程，ユーザ間の嗜好が類似している可能性が高い．コサイン類似度の計算式は，ユーザ A の嗜好ベクトル v_A・ユーザ B の嗜好ベクトル v_B を用いると以下の式で表せる．

$$v_A = (w_{1A}, w_{2A}, w_{3A}, \cdots w_{NA}) \tag{2}$$

$$v_B = (w_{1B}, w_{2B}, w_{3B}, \cdots w_{NB}) \tag{3}$$

$$cos(v_A, v_B) = v_A \cdot v_B = \frac{\sum_{i=1}^{N} w_{iA} \cdot w_{iB}}{|v_A||v_B|} \tag{4}$$

この手法は，ユーザのツイート解析する事で嗜好情報を直接扱う事ができる為，一般的に他の手法よりも精度の高い推薦を行うことができる．また，解析するツイートを最新のツイートに限定する事で，そのユーザの嗜好の変化にも対応する事ができる．

コンテンツベース方式の長所・短所は以下の様な点が挙げられる．コンテンツベース方式はツイート等のコンテンツを解析して推薦を行う為，ユーザの嗜好を反映した推薦が可能である．その為，商品の PR 等には適さないが，ユーザ間のコミュニーケーションの拡散という目的に適した推薦方式であるといえる．

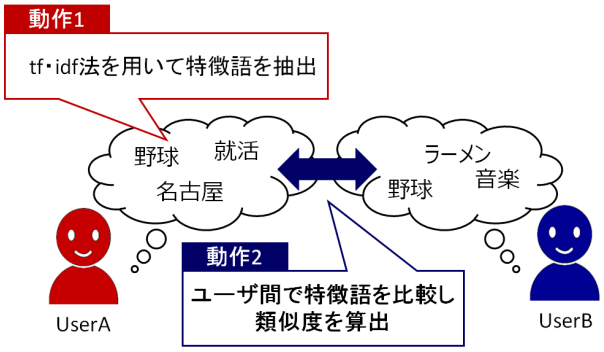

図 1　コンテンツベース方式

コンテンツベース方式のフォローユーザ推薦の利点・欠点

利点

- ユーザの過去の行動履歴を基に推薦を行う為，ユーザの嗜好を配慮した推薦が可能

欠点

- 類似するアイテムばかり推薦され，ユーザの飽きを招く可能性がある
- ユーザ数が膨大になるとコンテンツモデルの生成に多大な負担がかかる

以上を考慮し，本研究ではユーザ間のコミュニケーションの活性化を目的として，コンテンツベース方式のフォローユーザ推薦に着目する．

3.2　コンテンツベース方式の問題点

上記で述べたコンテンツベース方式のフォローユーザ推薦手法はニュース記事等の推薦システムで利用されている場合が多い．しかし，Twitter のフォローユーザ推薦システムには余り利用されていない．この原因として，キーワードの省略等が挙げられる．Twitter では略語が用いられる事が多い為，同じ意味を持つキーワードでも表現が異なる．例えば，「名古屋工業大学」というキーワードでも「名工大」･「名工」とのように多くの略語が存在する．しかし，各ユーザはこれらの略語の統一性を考慮して，ツイートをしている訳では無い．その為，1 人のユーザのツイート内でも多くの略語が存在する事に加え，ユーザ毎でも表現方法が異なる．

現在のコンテンツベース方式の推薦手法は各キーワードの意味を考慮していない為，同じ意味を持つキーワードであっても，異なる省略をしていると別のキーワードとして処理してしまう．これによって，キーワードに対する重み付け値が分散してしまい，特徴語にノイズが混入しやすくなってしまう．その結果，推薦の精度の低下を引き起こしてしまう (図 2)．

4.　提案手法

現在のコンテンツベース方式のフォローユーザ推薦手法はキーワードに対する重み付け値が分散してしまい，特徴語にノイズが混入しやすくなるという問題点がある．そこで，本研究ではこの問題点の解消の為に，従来の tf･idf 法を以下のように改良した．

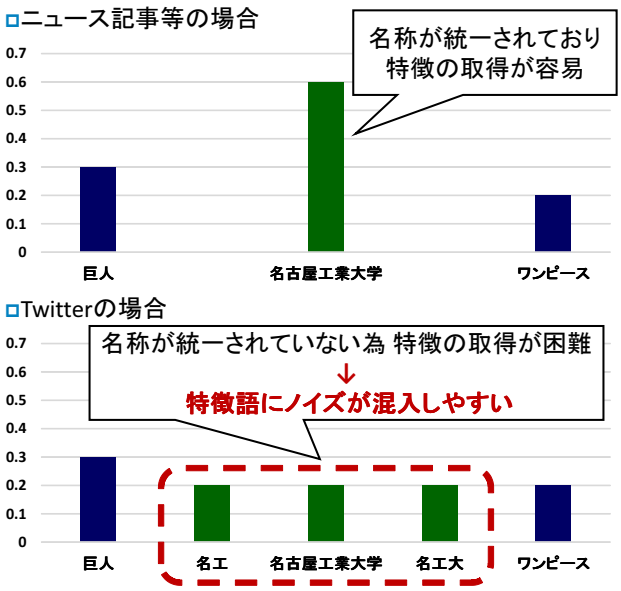

図 2　問題点：重み付け値の分散

キーワードをカテゴリ毎に管理

提案手法では各キーワードをカテゴリ (スポーツや音楽等) 毎に管理する．キーワードに新たにカテゴリという情報を付加する事により，ユーザがどのカテゴリに属すか考慮する事ができる．その為，同じ名前を持つキーワードであっても，属すカテゴリが異なれば別のキーワードとして処理される．

新たな指標として興味領域を追加

従来の tf･idf 法で用いる指標に加え，「興味領域」という指標を加える．興味領域とは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．同じ意味を持つキーワードならば同じカテゴリに属す為，それらのキーワードの重み付け値はカテゴリの興味領域に応じて高くなる．その結果，従来手法で埋もれていたキーワードを取り出す事ができ，ノイズとなるキーワードは特徴語に含まれにくくなる (図 3)．以上により，高精度且つ実用性が高いフォローユーザ推薦が可能になる．

5.　提案手法の構成

提案手法は以下の 3 要素で構成される (図 4)．興味領域取得機能は従来手法に新たに追加した機能であり，特徴語取得機能と推薦フォローユーザ取得機能は従来手法を改良したものである．

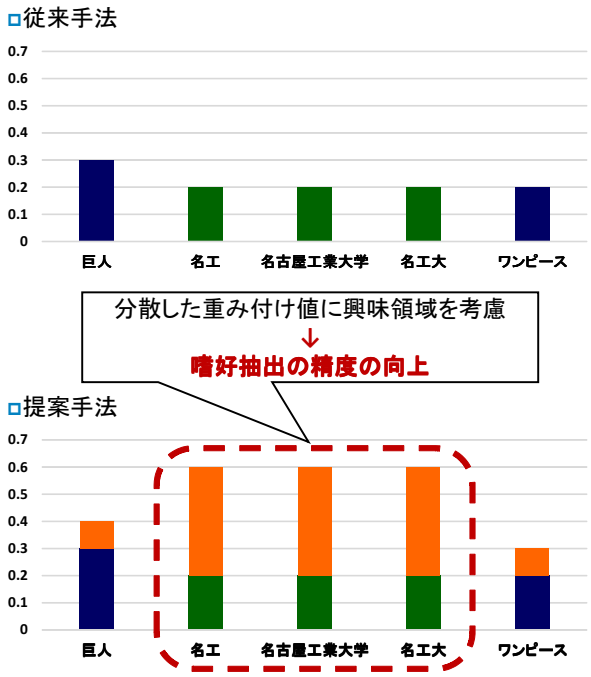

図 3 提案手法による問題点の解消

- 興味領域取得機能
- 特徴語取得機能
- 推薦フォローユーザ取得機能

この手法はユーザのツイートを入力データとして推薦を行う．興味領域取得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．特徴語取得機能は tf・idf 法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のようになる．

(1) 興味領域取得機能を用いてユーザが興味を持っているカテゴリを取得
(2) 特徴語取得機能を用いてユーザのツイートの特徴語を取得
(3) 推薦フォローユーザ取得機能を用いて推薦するユーザを取得
(4) 推薦結果をユーザに掲示

5.1 興味領域取得機能

興味領域取得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) を取得し，それを基にユーザの興味領域を算出する機能である．この機能の処理の流れは以下のようになる．なお，形態素解析とは辞書を基に文章を意味のある単語に区切り，品詞等を判別する自然言語処理技術であり，漢字変換等に利用

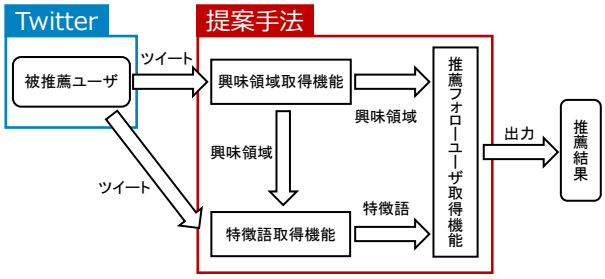

図 4 提案手法の概要図

される．

(1) ユーザのツイートを形態素解析
(2) 形態素解析したキーワードのカテゴリを取得
(3) カテゴリ毎にキーワードの数をカウント
(4) 各カテゴリのキーワード数を基に興味領域を算出

5.1.1 カテゴリの取得

この機能はユーザのツイートを形態素解析を行い，キーワード毎にカテゴリを取得する．

例えば，「ジャイアンツの試合結果をスマホで確認する」というツイートを対象にした場合，この中には野球の球団名を示す「ジャイアンツ」，電化製品をを示す「スマホ」が存在する為，「ジャイアンツ」からカテゴリ「SPORTS」，「スマホ」からカテゴリ「ELECTRONICS」を取得する．

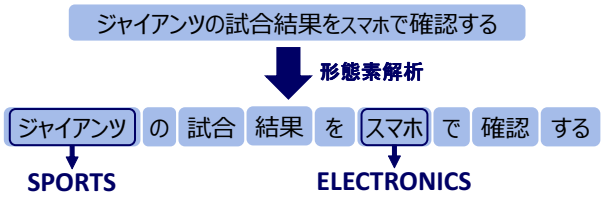

図 5 ツイートからカテゴリの取得

5.1.2 興味領域の取得

興味領域の取得はユーザのツイートから取得したカテゴリを基に行われる．取得したキーワードの数をカテゴリ毎にカウントし，カテゴリを取得したキーワードの総数に対する割合を興味領域とする．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 6)．この計算をすべてのカテゴリに対して行い，ユーザの各カテゴリへの興味度合を反映したデータとして扱う．

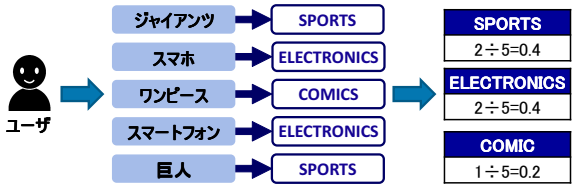

図 6 興味領域の取得

5.2 特徴語取得機能

特徴語取得機能はユーザのツイートの特徴語を取得する機能である．特徴語の取得には，興味領域と tf と idf を用いる．まず，各評価指標を平等に扱う為に，各キーワードの tf・idf 値と興味領域に対し，平均と標準偏差の和に応じて正規化を行う．そして，tf・idf 値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する (図 7)．

提案手法による重み付け値 w' はキーワード i に対する tf・idf 値 w_i と興味領域 int_i を用いると以下のようになる．なお，σ は tf・idf 値の標準偏差 σ' は興味領域の標準偏差である．

$$w_i' = \frac{w_i}{\overline{int} + \sigma} + \frac{int_i}{\overline{w} + \sigma'} \tag{5}$$

重み付けを行ったキーワードは各カテゴリごとに集約し，嗜好ベクトルとして管理しておく．これにより，ユーザが興味を持たないカテゴリのキーワードを排除することができ，キーワードのノイズの排除をすることが可能になる (図 8)．

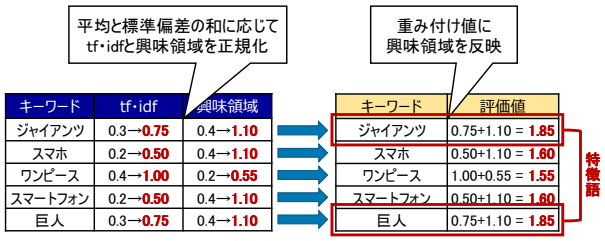

図 7 特徴語取得機能の動作例

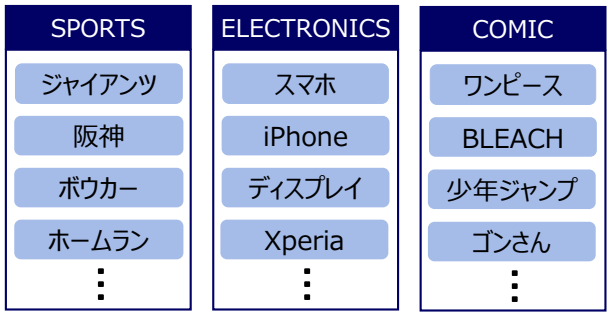

図 8 嗜好ベクトルの管理

5.3 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ユーザ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．推薦フォローユーザ取得機能は以下のように動作する．これらの動作を全てのユーザに対し行い，類似度の値が上位のユーザを推薦ユーザとして提示する．推薦ユーザ数は利用形態によって異なるが，本論文の評価実験では，被推薦ユーザのフォローユーザ数と同数のユーザを推薦する．

(1) 各ユーザに対して興味領域の値が上位のカテゴリを比較
(2) 上位のカテゴリ内に一致するカテゴリがあればそれに対応する嗜好ベクトルを比較
(3) 参照した嗜好ベクトルを用いて類似度を算出

興味領域の値が大きいカテゴリである程ユーザが興味を持っているものである．その為，興味領域の値が上位のカテゴリの嗜好ベクトルに絞る事で，従来の手法と比べユーザが興味を持たないキーワードが含まれる嗜好ベクトルを類似度計算に用いられる可能性を低下させる．これにより，よりユーザの嗜好を反映した嗜好ベクトルを類似度計算に用いる事ができ，推薦精度を向上させる事が可能になる．

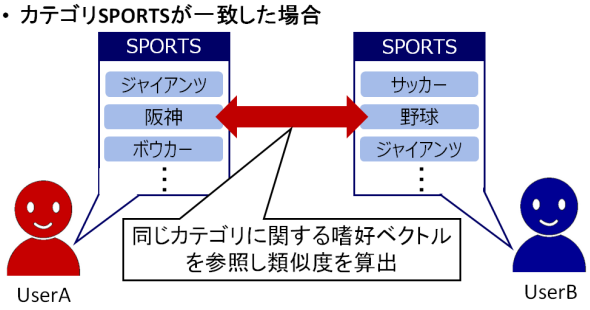

図 9 推薦フォローユーザ取得機能

6. 提案手法の実装

評価実験を行う為に，提案手法のプロトタイプを実装した．提案手法の実装には，Twitter API[3] 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 SlothLib[6] ，及び，キーワードごとにカテゴリを取得する為のはてなキーワード自動リンク API[7] ，計 3 つのライブラリを用いた．

6.1 TwitterAPI

TwitterAPI とは Twitter 社が提供しているサービスで，Web サイトやスマートフォンアプリなどを介して Twitter の機能を呼び出す事ができる．この API は Twitter のアカウント情報とアプリケーションを登録する事で利用できる．

今回は TwitterAPI の機能の内，Twitter ユーザのアカウント情報とツイート，フォローユーザを取得するのに利用した．

6.2 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．今回は興味領域算出機能での各キーワードのカテゴリを取得するのに利用した．

はてなキーワードのカテゴリは図 10 に示す 20 種類がある．評価実験ではカテゴリとしての意味を持たない「一般」・「はてな」・「はてなダイヤリークラブ」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．

6.3 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・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，興味領域取得機能でのツイートの形態素解析を行うツールとして利用した．

図 10 はてなキーワード自動リンク API のカテゴリ

7. 評価実験：提案手法の有効性

提案手法がユーザの嗜好にあった Twitter アカウントを推薦できているか評価を行い，従来手法と比較を行った．従来手法は tf・idf 法とコサイン類似度を用いたフォローユーザ推薦手法，及び，ランダム推薦の 2 種類用意した．

7.1 評価対象

評価対象として，以下の条件をすべて満たす Twitter ユーザ 300 人をランダムに取得して，実験対象ユーザとした．

条件 1:累計ツイート数が 1000 以上

評価実験を行う為には一定以上のツイート数が必要である為，提案手法を実行する上で十分な累計ツイート数を有しているユーザを対象にした．

条件 2:フォローユーザ数が 50〜59 人

より一般的なユーザを実験対象にする為に，Twitter ユーザのフォローユーザ数の平均値付近である 50〜59 人のフォローユーザ数を有しているユーザを対象にした．

7.2 評価用データセット

本実験では提案手法が実用的な手法であるか評価する為に，評価用データセットを作成した．評価用データセットは以下の 2 つの評価用ユーザで構成され，これらのデータセットを実験対象ユーザ 300 人に対して作成した．作成した評価用データセットは 100 人ずつに分け，それぞれ実験対象ユーザ群として扱う．

評価用ユーザ 1:実験対象ユーザのフォローユーザ

実験対象ユーザが既にフォローしているユーザである．なおツイートを一般公開していないユーザは対象外とした．このユーザは既にフォローされているユーザである為，実験対象ユーザの嗜好に合ったユーザであると言える．

評価用ユーザ 2:実験対象ユーザと関わりの無いユーザ

実験対象ユーザがフォロー・リプライ・リツイートを一度もした事のないユーザである．なおツイートを一般公開していないユーザは対象外とした．このユーザは評価用ユーザ 1 と比べ，実験対象ユーザの嗜好に合っていないユーザであると言える．

評価用ユーザ 2 のサイズは評価用ユーザ 1 のサイズに応じて可変とし，このサイズを変化させる事で，評価用データセット全体のサイズに対する評価用ユーザ 1 のサイズの割合 α を変化させる．例えば，評価用ユーザ 1 のサイズが 50 の場合に評価用ユーザに対し，$\alpha = 1/4$ の評価用データセットを作成する場合，評価用ユーザ 2 のサイズを 150 にする事により $\alpha = 1/4$ に設定する．α はユーザ u の評価用ユーザ 1 の集合 $follow_u$ と評価用ユーザ 2 の集合 $unfollow_u$ を用いると以下のように表せる．

$$\alpha = \frac{|follow_u|}{|follow_u| + |unfollow_u|} \tag{6}$$

7.3 実験方法

上記の方法で作成した評価用データセットに対し，推薦手法を用いて被推薦ユーザのフォローユーザ数と同数のユーザを推薦した．この中からどの程度評価用ユーザ 1 を推薦できたかを再現率を用いて数値化し，各手法の評価とする．

推薦ユーザ数はフォローユーザ数と同数にした．これは，各被推薦ユーザの評価用データセットの大きさはフォローユーザ数によって異なる為，このように設定した．これにより，各ユーザのフォローユーザ数に関わらず平等な評価を行う事が可能になる．

再現率はユーザのフォローユーザの総数に対する推薦結果に含まれるフォローユーザの数の割合で表される為，本研究の趣旨に合ったものであると考え利用した．本実験における再現率は，ユーザ u の評価用ユーザ 1 の集合 $follow_u$ と推薦された評価用ユーザ 1 の集合 $|hits_u|$ を用い

ると再現率 $Recall_u$ は以下のような計算式で表現できる．最終的に 100 人の実験対象ユーザ群に対して評価を行い，各ユーザの再現率の平均値を最終的な評価とする (図 10)．

$$Recall_u = \frac{|hits_u|}{|follow_u|} \tag{7}$$

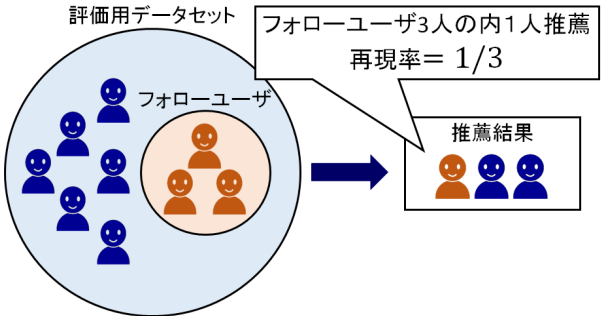

図 11 評価実験の問題設定

また，本実験では以下の 3 つの値を変化させて，各値と再現率の関連性について評価する．なお，変化させるパラメータ以外の値は常に $\alpha = 1/4$，ツイート数 = 500 ツイート，採用カテゴリ数 = 上位 5 位に固定してある．

評価用データセット中の評価用ユーザ 1 の割合 α

評価用データセット中の評価用ユーザ 1 の割合 α を変化させて，再現率との関連性について調査する．実際に提案手法を導入し，推薦を行う際には膨大な数のユーザから自分の嗜好に合った推薦ユーザを選抜する必要がある為，α の値が小さい程，より現実的なシミュレーションを行うことができると考察する．

解析するのに用いた 1 人あたりのツイート数

解析するツイート数と再現率との関連性について調査する．ツイート数が多くなる程，計算量が増加する為，少ないツイート数で高い再現率を実現する必要がある．本実験では，直近のツイートを優先して解析する．

採用するカテゴリ数

提案手法では，作成した嗜好ベクトルの内，興味領域の値が上位のものを採用し，類似度計算を行っている．その為，どの程度上位の嗜好ベクトルを採用する必要があるか調査する必要がある．

これらの実験を 1 つの実験対象ユーザ群に対し 1 回ずつ行った．それぞれの実験結果の違いを確認し，提案手法の有効性と特徴について考察した．

7.4 実験結果

評価実験の結果とその考察を以下に述べる．

評価用データセット中のフォローユーザの割合 α

α に関する実験結果を図 12 に示す．どの手法においても α の値が小さくなる程，すべての手法の再現率が悪化した．これは，ランダム推薦の結果を見る限り，評価用データセット中の評価用ユーザ 1 の割合が小さくなる為，ユーザの嗜好に合ったユーザを推薦できる可能性が低くなったことが原因だと考えられる．

提案手法の再現率は常に他の 2 つの手法よりも高い値を維持している．また，その値は α が小さくなる程，他の手法との値の差が大きくなっている．特に，$\alpha = 32$ に関しては，従来手法の 2.5 倍・ランダム推薦の 12 倍の再現率を記録しており，提案手法がより実用的な手法であるといえる．これは，提案手法が従来手法よりも実験対象ユーザの嗜好に合った推薦を行うことができる事を意味している．

以上より，提案手法が最も有効な推薦手法である事を証明できた．

解析するのに用いた 1 人あたりのツイート数

解析ツイート数に関する実験結果を図 13 に示す．なお，ランダム推薦ではツイートの解析を行っていない為，本実験では除外する．どちらの手法においても一部を除き，解析ツイート数が多くなるほど徐々に再現率が向上している．それに加えて，どの解析ツイート数にでも，提案手法の方が従来手法より高い再現率を算出している．

また，提案手法の 200 ツイートにおける再現率と従来手法 1000 ツイートにおける再現率の値がほぼ一致している．これを考慮すると，提案手法では従来手法の 1/5 のツイート数で同じ性能を実現できる為，提案手法の方が実用的な推薦手法であると言える．

以上より，提案手法が最も有効な推薦手法である事を証明できた．

採用するカテゴリ数

採用カテゴリ数に関する実験結果を図 14 に示す．なお，従来手法・ランダム推薦ではツイートの解析を行っていない為，本実験では除外する．採用カテゴリ数においては上位 4 位～上位 7 位の範囲に再現率のピークがあり，その範囲から離れるほど再現率が悪化した．以上より，提案手法を利用する際には採用するカテゴリ数を適切に設定する必要がある事が確認できた．

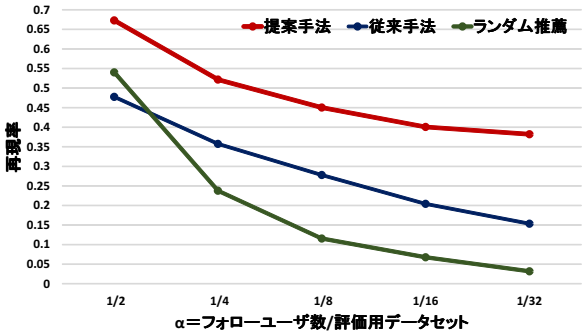

図 12 評価実験:α と再現率との関連性

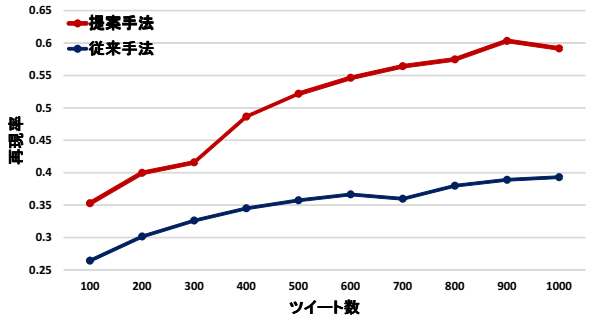

図 13 評価実験:ツイート数と再現率との関連性

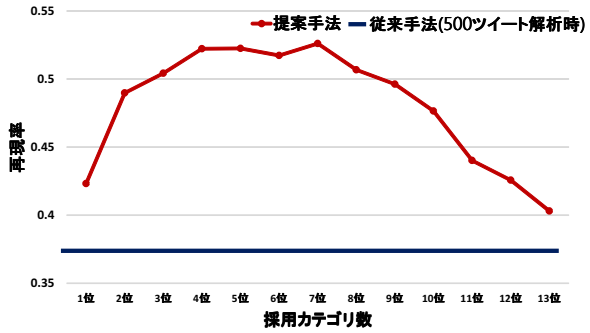

図 14 評価実験:採用するカテゴリ数と再現率との関連性

8. まとめ

本研究の目的は，コンテンツベース方式を利用した Twitter フォローユーザ推薦の問題点を解決し，推薦精度を向上させる事である．

本稿ではまず，Twitter の基本的な知識ととして Twitter の機能と特長についての説明を行った．そして，Twitter によるコミュニケーションの有用性，問題点について述べた．次に，フォローユーザ推薦のアルゴリズムとその問題点について述べた．

そして，Twitter フォローユーザ推薦の問題点を解決する為に，興味領域を考慮した Twitter フォローユーザ推薦を提案した．これは，従来のコンテンツベース方式のフォローユーザ推薦にスポーツや音楽等のカテゴリの情報を付加し，キーワードをカテゴリ毎に管理を行った．さらに特徴語抽出手法である tf・idf 法に新たに「興味領域」という指標を加えた．これを行う事で，各キーワードの意味を考慮した上での推薦を行う事が可能になり，従来の推薦手法よりも精度の高い推薦を行う事ができる．

最後に，提案手法の実装を行い，評価実験として推薦結果の推薦精度・計算時間を比較した．評価実験の結果から，提案手法の有効性を確認する事ができた．

9. 今後の課題

今後の課題を以下に示す．

- はてなキーワード自動リンク API に依存しないキーワードのカテゴリ分類の実現

  現段階の提案手法でのキーワードのカテゴリ分類はすべではてなキーワード自動リンク API の機能を利用して行っている．その為，この API でカテゴリ分類できないキーワードは提案手法では利用する事ができない．この問題点を解決し，提案手法のみでカテゴリ分類を行う事ができれば，有効性が更に増すと考察できる．
- 計算時間の削減もしくはそれの影響を抑える利用形態の提案

  提案手法はコンテンツベース方式を基に設計されている．その為，ルールベース方式等の他の推薦手法と比べ，計算時間が大きくなってしまう．これらを解決する為には，計算時間の削減が必要であるが，それを行うと推薦精度が低下する恐れがある．その為，この欠点の影響を抑える利用形態を提案する必要がある．

参考文献

[1] “Twitter”，https://twitter.com/.
[2] 土方嘉徳, “嗜好抽出と情報推薦技術”, 情報処理学会論文誌, Vol.48, No.9, 2007.
[3] “Twitter Developers”，https://dev.twitter.com/.
[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.
[7] “はてなキーワード自動リンク API”，http://developer.hatena.ne.jp/ja/documents/keyword/apis/rest.