# 知識領域クラスタを用いた Twitter 内におけるコミュニティの最適化

Follower Optimization for Twitter Community of Interest Field Using Knowledge Domain Clusters

伊藤 直也 † 米澤 朋子 ‡
Naoya Ito Tomoko Yonezawa

## 1. はじめに

近年，SNS（social networking service）は若年層を中心にコミュニケーション手段の一つとして日常の中に浸透しつつある．これは，対面コミュニケーションのような時間，空間的な制約がないコミュニケーションであるため，気軽に情報の発信や情報を得たりすることが出来ることに由来すると考えられる [1]．本稿で扱う Twitter* は幅広い年代で使われており，国内アクティブユーザが 4500 万人の非常に規模が大きな SNS である．Twitter を使っている人の利用目的は様々であり，他者とのコミュニケーションのための利用や，情報発信や情報収集のためなどの利用がみられる．一般的に SNS では現実世界でもつながりがある人とコミュニケーションを取ることが多いとされている [2] が，全く知らない人とでもコミュニケーションを取ることが報告されていることから，Twitter はコミュニティを広げられるツールでもある．

特に情報収集を目的としているユーザは，自分の興味・関心に合わせて情報を得たり関わる人を決めていると考えられる．Twitter のおすすめユーザ表示機能は著名人や Twitter 上での有名人が多く表示されてしまうことから，自分の興味を考慮したおすすめユーザ推薦の研究が行われている [3]．日々増える情報の中で，自分の期待する速度・量・質でコミュニケーションを行うことが出来る相手を探すことは困難であり，SNS 疲れ [4] を生じさせる可能性がある．ユーザの興味と合致し，コミュニケーションを取りたいフォローユーザを選択する際にツイートを見て確認することは，ユーザの苦労になる．ユーザの負担を軽減させながら，ユーザに合うコミュニケーション速度で，さらに知識が重ならない他ユーザが，最適な組み合わせで自動的に選び出されることが望ましい．

本稿ではつながりを介した情報収集を狙いとし，Twitter 上の複数ユーザをフォローしながらコミュニケーションを通じ情報を得るシステムとした．提案手法では，特定のキーワードに関する知識に関して，他ユーザが持つ知識をもとに word2vec を用い知識領域に分割したクラスタを形成し，クラスタからユーザを選ぶことで，他者から受け取る情報の知識領域の重複を抑えつつ知識情報量の合計を最大化する最適化問題として扱い，自動的にフォロー対象ユーザを決定するアルゴリズムを提案する．上記の SNS 疲れを軽減させるため，一日の平均ツイート数を制約条件として考慮することでコミュニケーションの速度を制限させる．ツイート中の熟語やカタカナ語をツイートの濃度として制約に入れることによって，つながる人のツイートの質を一定以上にする．組み合わせ最適化問題の一つである非線形ナップサック問題として定式化を行う．また本稿では特定の知識をサッカーに関する知識と固定し記述していく．

## 2. 関連研究

これまでも Twitter におけるユーザ推薦に関する研究は活発に行われている．

柏原 [5] は Twitter の利用動機と利用頻度の関連性を報告している．Twitter の利用動機は 5 つの種類があることを因子分析から明らかにし，1 つの因子では [趣味関心を共有する人の会話・交流] が示唆されていることから自分の興味分野の情報共有のために Twitter を利用している人がいることが分かる．

富永ら [6] はフォロー関係を用いることでユーザの潜在的な興味に合うツイートを推薦する手法を提案している．ツイートを興味，意外性，有用性の観点から分類し，ツイートの名詞を対象とした条件付き確率を用いることで興味がある単語を見つけ出すことを試みている．久米ら [7] も興味領域を考慮したユーザ推薦手法を提案している．これまでの研究 [8] ではツイート群に対する TF-IDF で得られた特徴語を用い，推薦するユーザを決定していた．特徴語の TF-IDF 値をそのまま使ってしまうことにより，違う語句でも同じ意味を表すケースに対応できていないなど複雑なノイズが発生していた．これに対し特徴語のカテゴリ分けを行い，カテゴリの割合を TF-IDF 値に加算することで，ノイズ削除を実現し特徴語の精度を向上させている．

Twitter の特性を考慮した文書クラスタリングも行われている [9]．提案手法はこれまで文書クラスタリングで使われていた leader-follower 法 [10] を拡張したものであり，Twitter 特有のツイートの短さを考慮し，同一アカウントへのツイートやツイート内に記載されている URL を参照しクラスタ分けをすることで精度の向上を示

† 関西大学大学院　総合情報学研究科　知識情報学専攻, Kansai University, Graduate School of Informatics
‡ 関西大学　総合情報学部, Kansai University, Faculty of Informatics
*https://twitter.com

した．本研究でも各ユーザの知識領域を分割するためにツイート内容を解析しクラスタ分けを行う．田村ら [11] はフォローユーザに対するメンションツイートの割合に着目し，新しくフォローユーザになったユーザに対しての会話しやすさを決定する手法を提案している．評価値が高いユーザの方が会話を行っている傾向が明らかに見られ，提案手法の有効性を示している．他にも，ユーザと関わる人のツイート間での類似度から推薦する人を決定するシステム [12] [13] [14] や，ユーザの興味を基にしたグラフ分析によって推薦するシステム [15] が見られる．

ユーザに対して 1 人の他ユーザを推薦するシステムは先行研究にも多くみられるが，コミュニティ単位として複数ユーザを推薦するシステムは少ない．本研究では，あるトピックに対するユーザが得られる知識情報量の合計の最大化を目的としており，ユーザの使い方を考慮した最適なつながりを決定することで，ユーザが関与するユーザ群を決定する点が先行研究とは異なる．これにより，ユーザが目的とする知識領域において，ユーザを中心としたコミュニティが形成されると考えられる．

## 3. 提案手法

### 3.1 概要

本研究では TwitterAPI から得られた実際のデータをもとに R 言語を用いて分析を行った．TwitterAPI を用いてツイートを集め，RMeCab[†]という自然言語処理を行う R 言語パッケージ[‡]を用いて対象データの解析を行った．

本稿で取り扱う知識領域として，サッカーに関する知識への慣れや習得を目指すこととし，サッカーと呟いたことがあり，ツイート数が 50 以上ある日本語で呟かれているアカウントを TwitterAPI から取得し分析対象とした．TwitterAPI から取得した 2249 件のアカウントのうち，アカウント名に bot という文字列を含む自動応答のアカウントと考えられるものは分析の対象外とし，2077 件のアカウントを対象とした (対象アカウント)．対象アカウントからツイート内容，ツイート数，初ツイートから最新のツイートまでの経過日数を収集した．TwitterAPI から取得可能なツイート数は 3200 が上限であるため，ツイート数が上限を超えるアカウントでも分析対象は最新のツイートから遡って 3200 までとした．

ここで，ユーザと関わる人の持つ知識情報量の合計の最大化を行うことを目指し，ツイートを形態素解析し得られたデータをクラスタリングすることで知識領域を分割し最適化処理を行う．例えば，サッカーに関する話題を取り上げる人たちにも，技術的興味，選手への興味，コミュニティへの興味など様々な知識領域が存在すると考えられるため，それらを満遍なくカバーするようフォローユーザを選択することを実現したいと考えた．提案手法の流れを図 1 で示す．

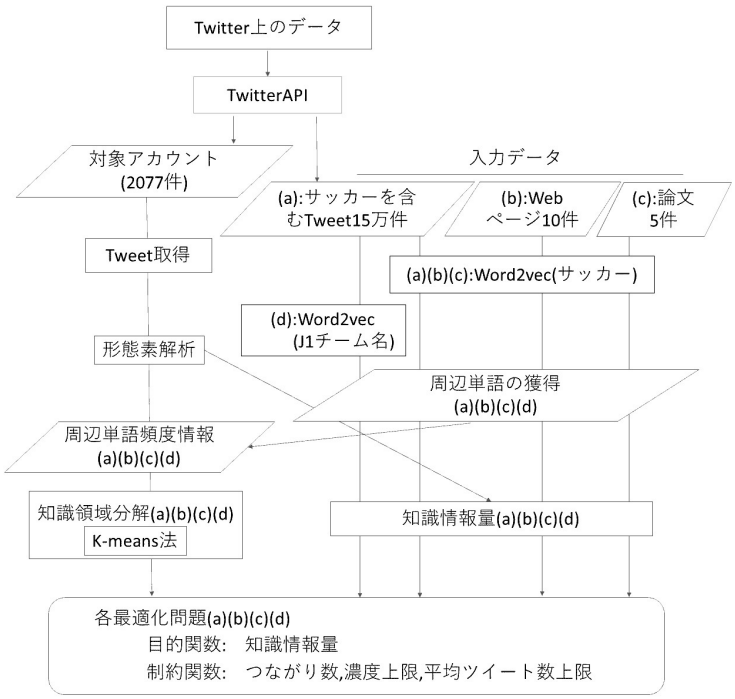


図 1: 提案手法フロー図

### 3.2 形態素解析

自然言語処理の一つである形態素解析を RMeCab を用い各ツイートに適用した．形態素解析 [16] とは，ある文章またはフレーズを意味を持つ最小限の単位に分解する手法である．本稿は品詞が名詞である単語だけに着目することで，関連する知識単語を出現させるユーザを選択できると考えた．また，狙いとする知識に対するコミュニティ形成を目的としているため，顔文字などの特有の表現は分析対象としない．

### 3.3 word2vec

word2vec [17] とは，ある単語の周辺の単語の確率を計算することで単語の意味を捉えることが出来る 2 層からなるニューラルネットワークモデルである．word2vec を用いたトピック抽出の研究 [18] も行われているように，近年の自然言語処理の応用研究に使われている．大規模なテキストデータを解析する際に有効とされている手法であるため，本研究における知識に関する単語の獲得においても使用した．

word2vec は適用する対象データによって出される結果が異なる．特に口語が多く含まれるツイートと，説明的文章を伴う web ページ等とは出力される単語リストに違いがあると考えられる．そのため，本研究ではサッ

[†] http://rmecab.jp/wiki/index.php?plugin=attach&refer=RMeCab&openfile=manual.pdf
[‡] https://www.r-project.org/index.html

表 1: Twitter から得たサッカーの周辺単語

| | | | | |
|---|---|---|---|---|
| 人中 | ー | 人 | 男子 | 箱 |
| 草 | 全部 | 笑 | 事 | 決勝 |
| 野球 | バスケ | バレー | 最近 | ーーー |
| 微妙 | 見事 | オリンピック | 日本 | 友達 |
| 不思議 | 観 | 今 | バレーボール | 絶対 |
| ムカ | 他 | 幸運 | 残念 | 私 |
| 普段 | 文句 | ショック | オール | ドッチボール |
| 贅沢 | タダ | 大丈夫 | 結果 | 今日 |
| 今更 | ワイ | ヨット | マジ | 両方 |
| 女子 | 身近 | 準決勝 | 流石 | 文章 |

カーが含まれているツイート 15 万件，サッカーに関する Web ページ 10 件，Google スカラーから検索した本文にサッカーという語が含まれている論文 5 件をそれぞれ入力データとし，サッカーに対する類似語と類似値を算出した．Web ページ，論文からは広告や参考文献を除いた本文のみを対象としている．上記のデータに word2vec を適用した結果，サッカーとの類似度が高い単語群が得られた．

一方，ツイート 15 万件に word2vec を適用し出力された単語群は，サッカーに関連するとは言えない一般的な意味合いの単語が含まれていた．そのため，J リーグの中でも有名なチームが多いと思われる J1 の 10 チームの名前を「サッカー」に代替し Word2Vec を適用した結果，比較的サッカーへの関連度の高い単語群が選ばれた．これは，サッカーに対する一定の知識や興味があるユーザのツイートをターゲットとするためである．それぞれの入力データから Word2Vec で出力された類似語のうち類似値が高い単語 50 個を，サッカーに関係する周辺単語と定義した．表 1–表 4 においてそれぞれを入力データとした周辺単語を示す．

### 3.4 知識領域分割クラスタリング

どの周辺単語がどの程度の頻度で呟かれているかを用いて各ユーザが持つ知識傾向を明らかに出来ると考え，3.3 節において定義した周辺単語の呟き回数を調べた．各入力データにおいて周辺単語を 50 個得たため各アカウントは 50 次元の周辺単語頻度情報を持つといえる．ここで，知識領域を分割するために 2077 件のアカウントに対し周辺単語頻度情報を基にクラスタリングを行った．知識領域個数を 10 個に設定し非階層クラスタリング手法である k-means 法を用いた．知識領域を細分化することで知識の重なりが少なくなると考えられるが，今回のデータでは 10 以上のクラスタ数における出力ではデータ数が 1 のクラスタが複数あることが確認されたため，知識領域の分割数としてクラスタ数を 10 個とした．

一般的に人の知識領域は様々な領域にわたると考えられるため，k-means 法などのハードクラスタリングで 1 つのクラスタに決定することは難しいと考えられるが，本研究では知識領域を分割することでユーザと関わる人たちの持っている知識の重なりを少なくすることが目的であるため k-means 法で決定されたクラスタはそのアカウントの持つ最も特徴的な知識領域であると捉えている．

k-means 法はクラスタ分けが初期値に依存することが知られている [19] が，本研究で行いたい知識領域分割には正解データがないため，k-means 法を一度だけ試行し，得られたクラスタを各アカウントの知識領域としている．各データによる周辺単語を用いた結果，それぞれクラスタがどのように分かれたのかを表 5 に示す．

### 3.5 知識情報量の設定

本稿では，各アカウントが持つ知識情報量を以下の手順で定義する．周辺単語 $i$ が各アカウントで呟かれている回数を $count_i$ とし，word2vec によって出力された周辺単語の類似値を $similar$ とした時に，周辺単語は入力データごとに 50 個あるため，以下の式によってアカウ

表 2: Twitter から得た J1 チーム名の周辺単語

| | | |
|---|---|---|
| 川崎 | 神戸 | 清水 |
| サンフレッチェ広島 | フォルツァ | 名古屋 |
| ジュビロ磐田 | 高槻 | 光紀 |
| 藤口 | アルディージャ | ガールズ |
| バルセロナチェルシー | 広島東洋カープ | 広島 |
| 椛 | フットボールアディクト | 大分 |
| フロンターレサポーター | スペランツァ | ヴァンフォーレ |
| 磐田 | 生野 | 長居 |
| ヤンマー | 不在 | 麻生 |
| ガンバ大阪 | 前座 | アディクト |
| 千里丘 | 北浜 | ヴェルディ川崎 |
| 吹田 | 大宮 | サンフレ |
| トリニータ | パナソニック | 英正 |
| 泉澤 | チアゴ | 択年 |
| フリューゲルス | 初瀬 | 清水 |
| マイナビベガルタ | フリューゲルス | 本拠地 |
| アルディージャ | 西澤 | |

表 3: 論文から得たサッカーの周辺単語

| | | | | |
|---|---|---|---|---|
| パフォーマンス | 化 | 手法 | ゲーム | 質的 |
| コンピュータ | 戦術 | 方法 | データ | 力 |
| 選手 | 部分 | チーム | 等 | 年代 |
| ビデオ | 数量 | パターン | 問題 | 画像 |
| 因果 | 正確 | 動画 | 法 | 背後 |
| 距離 | 時間 | 像 | 技術 | 中 |
| 有効 | 家 | 沖 | 専門 | システム |
| 有用 | 的 | 数値 | 情報処理 | 情報 |
| 複雑 | 集団 | 因子 | 今後 | 性 |
| 要因 | 点 | 領域 | 従来 | モデル |

表 4: Web から得たサッカーの周辺単語

| | | |
|---|---|---|
| 戦法 | 国際 | スコットランド |
| イングランド | カップ | 非公式 |
| ヨーロッパ | 世界中 | イギリス |
| 大会 | 近代 | 現代 |
| 引き分け | ロンドン | 選手 |
| 地域 | 最初 | その後 |
| リーグ | 世界 | 各国 |
| 多く | 公式 | 可能 |
| 後半 | 年代 | 用具 |
| アマチュア | ドイツ | アジア |
| 年間 | ロング | プレイ |
| パブリック | 非常 | 主審 |
| 故意 | チーム | 大学 |
| 項目 | 国内 | 当時 |
| クラブ | プレイヤー | これ |
| 世紀 | その他 | ボール |
| 戦術 | プロフェッショナル | |

ントが持つ知識情報量を決定する．

$$\text{知識情報量} = \sum_{i=1}^{50} similar * log(count_i + 1)$$

$count_i$ を対数関数の変数とし，頻度が上がるごとにアカウントの知識情報量が極端に増加し過ぎるのを防ぐ．真数部分を $count_i + 1$ としているのは $count_i$ が 0 の時に知識情報量が 0 となるようにしているためである．

## 4. 最適化

### 4.1 非線形ナップサック問題

多次元制約非線形ナップサック問題は次のように定式化される．

$$\text{maxmize} \sum_{i \in N} f_i(x_i)$$

subject to

$$g_j(x) = \sum_{i \in N} g_j(x_i) \leq \ b_j (j = 1, 2, \cdots, m)$$

$$x_i \in A_i (j = 1, 2, \cdots, m)$$

ここで，N=$\{1, 2, \ldots, n\}$は変数の番号の集合であり，$A_i = \{1, 2, \cdots, a_i\}(i \in N)$ は各変数の項目集合である．多次元制約とは，制約関数が 2 つ以上の複数制約のことを指す．

本稿で扱う問題は制約式が 3 つの複数制約問題であるが，代理制約法 [20] [21] を用いることで単一制約問題へと変形し問題を解いた．代理制約式を満たす解は原問題の制約関数の上限値を越えていることがある．これを代理ギャップと呼び，本研究で解いた問題にもいくつか代理ギャップが発生した解が見られたが，制約関数の上限値を下げることで代理乗数を変化させ，解が原問題の実行可能解になるように調整した．

#### 4.1.1 グローバルグリーディ法

本稿で問題のモデルとした非線形ナップサック問題に対するグローバルグリーディ法 [22] のアルゴリズムの概略を説明する．

グローバルグリーディ法は非線形ナップサック問題の近似解を求めるアルゴリズムとして提案したもので，解を選択するときに全体を見ることでより良い代替案を選択する．

優越テスト

$f_i(x_i)$ に対して優越テストを行う．ある項目に対して $f_i(p) \leq f_i(q)$ かつ $g_i(p) \geq g_i(q)$ となる $f_i(p), g_i(p)$ が存在する場合は解の候補から外す．

DGR 型優越テスト

$f_i(k), g_i(k)$ において $k \rightarrow k+1$ に変化したときの制約関数値の変化量に対する目的関数値の変化量の比 $w_i(k)$ を

$$w_i(k) = \frac{f_i(k+1) - f_i(k)}{g_i(k+1) - g_i(k)}$$

とする．このとき $w_i(x_i)(x_i \in A_i)$ が単調減少（DGR 型: Decreasing Gain Ratio）する関数になるようにする．単調減少しない場合は $k+1$ を $A_i$ から削除する．

アルゴリズム　グローバルグリーディ

暫定解を $x_i^G = 1\{i \in N\}$ と初期化する;
問題 P の代替項目集合 $A_i\{i \in N\}$ に対して優越テストを実施し，優越された項目を削除する;

`Loop`
　暫定解 $x_i^G\{i \in N\}$ を用いて全ての代替項目に対して実行可能性テストを行い，明らかに実行不可能な項目を削除する；
　`If`（実行可能な項目がない）`ExitLoop`；
　代替項目に対して `DGR` 型優越テストを実施し，制約関数に対する目的関数の変化量が単調減少になるように代替項目の一部を削除する；
　生き残った代替項目に対してグリーディ法を用いてグリーディ解を求め，暫定解 $x_i^G\{i \in N\}$ を更新する；
　`If`（最大許容資源量 `b` に余裕がない）`ExitLoop`；
`EndLoop`
得られた暫定解 $x_i^G\{i \in N\}$ をグローバルグリーディ解とする．

更新が一度も行われず，$b = 0$ となる場合，得られた解は必ず厳密解となることが保証されることがこれまでの確率的要素を用いたヒューリスティックな解法 [23] とは異なる点である．

更新処理が終了した時点で $b > 0$ であれば，次に選ばれるはずであった代替案を $b = 0$ となるように比例配分

表 5: クラスタ内データ個数

| 周辺単語を得た入力データ | クラスタ番号 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Twitter(J1) | 253 | 22 | 10 | 5 | 1 | 1749 | 3 | 4 | 8 | 22 |
| Twitter(サッカー) | 6 | 1347 | 329 | 1 | 17 | 53 | 38 | 148 | 25 | 113 |
| Web(サッカー) | 180 | 26 | 86 | 879 | 12 | 12 | 37 | 11 | 581 | 253 |
| 論文 (サッカー) | 5 | 4 | 247 | 1592 | 11 | 14 | 78 | 4 | 116 | 6 |

して取ることで質のいい上界値を出すことが出来る．上界値を適切に設定できることは分枝限定法などの上界値を目安にして問題を解くアルゴリズムの高速化を実現できる．

### 4.2 問題の作成

本研究ではユーザが受け取る知識情報量の合計の最大化を目的としているため，問題を以下のように設定する．目的関数:3.5 節において定義した知識情報量，制約関数:(1) つながり数 (2) 濃度 (3) 平均ツイート数とする．

また本稿では知識領域分割のために 10 個のクラスタに分けたが，制約につながり数を入れることによって各クラスタから一人ずつ選ぶのではなく本当に重要なクラスタから他ユーザを選択出来るようにした．本稿では一つ目の制約であるつながり数を 8 とした．

二つ目の制約である濃度とは，そのアカウントの 1 ツイートあたりのカタカナ語もしくは漢字熟語の数を表したものである．SNS 特有の意味のない言葉を避けるため設定した．濃度は大きいほど情報があるように思われるが，本研究で定義している知識情報量とはサッカーに関係する情報量であるため，濃度が高くても知識情報量が小さかったり，その逆もある．また濃度が高ければ何らかの意味のあるツイートをしている指標として捉えることが出来るが，下限値を決めると本稿で定義した知識情報量が大きいユーザも選ばれなくなる可能性があるため上限値を決めることで濃度制約によって選ばれないユーザを少なくし，つながるコミュニティの合計濃度が一定以上になると考えたため上限値を設定した．また 1 ツイート中に熟語もしくはカタカナ語を平均して 3 個以上呟いている集団は濃度が濃いコミュニティと捉えることができ，このような情報ばかりを多数受け取ると，ユーザが疲れる可能性もある．これに対し濃度上限値を設定することで，このようなコミュニティ形成を防げると考えた．ユーザによって受け取れる情報の濃度は異なると考えられるため，本稿ではツイートあたりの濃度を 1，2，3 の 3 種類を想定し，つながる 8 ユーザの濃度合計の上限を 8，16，24 の 3 種類とした．

三つ目の制約の平均ツイート数とは，各アカウントユーザが一日に何ツイートしているかを 3.1 節の TwitterAPI から取得した情報を用いて算出した．本稿ではツイート数の平均が比較的少ない人もしくは比較的多い人を集めることを考え，一日に平均 5 ツイートする人もしくは平均 10 ツイートする人を前提に，つながり数が 8 であることより，つながる全ユーザ数 × 一人のツイート想定数として，平均ツイート数上限値を 40，80 と設定した．

以上の設定によって作成された問題 24 問（入力データ 4 種 × 濃度上限 3 種 × 平均ツイート数上限値 2 種）に対して計算を行う．

## 5. 計算機実験

前章で作成した問題をグローバルグリーディ法で解いた実行結果を以下に示す．この実行結果の値は他ユーザからユーザが受け取る知識情報量 3.5 節の合計を示している．

表 6: 平均ツイート数上限値 40 の問題の小数点以下 3 桁の結果

| | 濃度上限 8 | 16 | 24 |
|---|---|---|---|
| Twitter(J1) | 287.337 | 333.904 | 333.904 |
| Twitter | 300.766 | 331.565 | 333.692 |
| Web | 229.471 | 266.640 | 266.640 |
| 論文 | 284.553 | 344.750 | 344.750 |

表 7: 平均ツイート数上限値 80 の問題の小数点以下 3 桁の結果

| | 濃度上限 8 | 16 | 24 |
|---|---|---|---|
| Twitter(J1) | 319.762 | 382.144 | 382.144 |
| Twitter | 304.734 | 333.692 | 333.692 |
| Web | 258.435 | 276.947 | 277.325 |
| 論文 | 336.070 | 372.962 | 372.962 |

## 6. 考察

本稿では平均ツイート数上限値として 40，80 の二種類を設定し問題を作成したが，結果から平均ツイート上限

値 80 の問題の方が全体的にユーザが受け取る知識情報量の合計が大きいことが分かる．このことから，ツイート頻度が高い人の中には周辺単語を呟く回数が多く，かつ本稿で定義した知識情報量が大きい人がいることが分かる．

入力データの違いとして，周辺単語 (Twitter-サッカー) を用いた結果は，平均ツイート数上限値の変化による目的関数値の変化が他のデータに比べ少ない．周辺単語 (Twitter-サッカー) に一般単語と考えられるものが多く含まれたことから，知識情報量に関わる単語を呟いている回数は頻繁にツイートする人とそうでない人の間では差が見られないということが示唆される．またほとんどの問題において濃度上限が 16，24 のときに解の変化が見られないことが分かる．これは 1 ツイート中に平均して 2 個以上の熟語またはカタカナ語を呟いている人が少ないために濃度上限を増やしても計算結果が変わらなかったと考えられる．濃度上限を 24 とすることは制約として意味をなさないことが示唆される．

論文と Web を入力データとした結果を見ると，論文の目的関数値は平均ツイート数上限値に関わらず Web と比べると高い値が得られており，知識情報量が多く得られると考えられる．論文の周辺単語を見ると周辺単語の前半はサッカーに関係すると考えられる言葉であるが，後半はジャンルを問わず一般的な言葉が目立つことから，周辺単語が発言に含まれやすくなり知識情報量が多くなったと考えられる．しかし同制約の問題において，論文と Web を比較した時に論文を入力データとして得られた目的関数値の方が平均して 70 以上大きいことから，論文を入力データとした場合には，サッカーについての情報を比較的多く得ることが期待できる．Web を入力データとした周辺単語はサッカーの特徴を表しているといえるが，140 文字以内で気軽に呟くことが出来る Twitter では略語が多いことが考えられるため，周辺単語 (Web) には略語は少なく，それを基準にした知識情報量の値が低くなった可能性が示唆される．

平均ツイート上限値 40 の問題では周辺単語 (Twitter-サッカー) の目的関数値が高くユーザが受け取る知識情報量は大きいといえるが，この周辺単語の多くが日常単語に近いため，これを用い設定した知識情報量は大きくなっているが，実際にサッカーに関係がある単語（現段階では真値は不明）に基づき知識情報量を算出した場合は小さくなると考えられる．周辺単語 (Twitter-J1) では周辺単語 (Twitter-サッカー) よりも一般性のない，サッカーに関連のある単語が多い中で目的関数値が大きい結果が多く，よりサッカーの知識を有するユーザとつながることが出来ていると考えられる．つまり，周辺単語 (Twitter-サッカー) よりも周辺単語 (Twitter-J1) からクラスタを作成しユーザ選定を最適化することで，知識を獲得する上でのつながりの最適化が適切に行われたと考えられる．

以上のことから，周辺単語 (Twitter-J1) および周辺単語 (論文) を用い得られた結果では，ユーザが受け取る知識情報量の合計を最大化しつつ，つながる意味のあるユーザを決定することが比較的可能だと考えられる．

## 7. おわりに

本稿では Twitter において，あるトピックに対して情報収集する際にユーザの得られる知識情報量の重なりを考慮し最大化できるように最適化問題を設計し結果を示した．各ユーザが呟いている単語をもとに知識領域を分けるクラスタリングを行うことで情報の重なりを考慮した．様々な入力データを基に word2vec から出された周辺単語を用い，各ユーザの持っているトピックに対する知識情報量を定義した．本稿ではトピックをサッカーに限定して設定し，限られたつながり数の中での知識情報量の最大化を行った．最適化の結果から，ツイート内容を基に J1 のチーム名から得られた周辺単語をもとにした結果と，論文のデータ中の単語 (サッカー) から得られた周辺単語をもとにした結果において，Twitter 中の単語 (サッカー) から得られた周辺単語や Web 中の単語 (サッカー) から得られた周辺単語をもとにした結果と比較した際に最適化によって得られた知識情報量の合計が大きくなり，つながる他ユーザが適切に選ばれたと考えられる．選ばれたユーザたちの持っている情報の重複度合いに対するユーザ評価は今後の課題としたい．

## 謝辞

本研究は科研費 19H04154，18K11383，25700021 の助成の一部を受け実施した．

## 参考文献

[1] 宮木由貴子. 多様化する SNS の利用目的. *Life design report*, No. 202, pp. 42–44, 2012.

[2] 董逸斐. 大学生における SNS の利用と満足. コミュニケーション科学, No. 34, pp. 65–83, 2011.

[3] 大村涼, 赤石美奈, 佐藤健. 語彙連鎖構造を用いた twitter ユーザー推薦手法の提案. 第 75 回全国大会講演論文集, 第 2013 巻, pp. 609–610, mar 2013.

[4] 來迎直裕, 小笠原直人, 佐藤究, 布川博士. 消えるメッセージによる義務感を軽減するコミュニケーションツール. Technical Report 1, 岩手県立大学大学院ソフトウェア情報学研究科ソフトウェア情報学専攻, 岩手県立大学大学院ソフトウェア情報学研究科ソフトウェア情報学専攻, 岩手県立大学大学院ソ

フトウェア情報学研究科ソフトウェア情報学専攻, 岩手県立大学大学院ソフトウェア情報学研究科ソフトウェア情報学専攻, mar 2014.

[5] 柏原勤. Twitter の利用動機と利用頻度の関連性:「利用と満足」研究アプローチからの検討. 慶應義塾大学大学院社会学研究科紀要 : 社会学・心理学・教育学 : 人間と社会の探究, No. 72, pp. 89–107, 2011.

[6] 富永一成, 牛尼剛総. フォローネットワークを利用したユーザの新しい興味の発見につながる tweet 推薦手法. *DEIM Forum*, Vol. F7-3, pp. 1–5, 2012.

[7] 久米雄介, 打矢隆弘, 内巧逸. 興味領域を考慮した twitter ユーザ推薦手法の提案と評価. 情報処理学会研究報告, Vol. F7-3, pp. 1–5, 2015.

[8] 貴志将考, 大沢英一. twitter における共通の関心を持つユーザのレコメンド. 全国大会講演論文集, Vol. 2013, No. 1, pp. 103–105, mar 2013.

[9] 辻井由佳, 西山裕之. マイクロブログの特徴を考慮した文書クラスタリング手法の提案と実装. 人工知能学会全国大会論文集, Vol. 2012, pp. 4I1R92–4I1R92, 2012.

[10] 岸田和明. 文書クラスタリングの技法:文献レビュー. *Library and information science*, No. 49, pp. 33–75, 2003.

[11] 田村政人, 小林亜樹. Twitter における会話しやすいユーザの推薦手法. 第 75 回全国大会講演論文集, 第 2013 巻, pp. 605–606, mar 2013.

[12] 森戸健太, 小林亜樹. フォロー関係を用いた異なるコミュニティのアカウント推薦手法. 第 79 回全国大会講演論文集, 第 2017 巻, pp. 469–470, mar 2017.

[13] 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. Technical Report 18, 株式会社きざしカンパニー, 株式会社きざしカンパニー, 株式会社きざしカンパニー, 京都産業大学, 京都産業大学, nov 2009.

[14] 藤野巖, 星野祐子. Twitter における「友達レコメンド」の実現方法について (データ工学). 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 113, No. 105, pp. 59–64, jun 2013.

[15] Marco Pennacchiotti and Siva Gurumurthy. Investigating topic models for social media user recommendation. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pp. 101–102. ACM, 2011.

[16] 中野洋, 野村雅昭. 日本語情報処理 : 日本語の形態素分析. 情報処理, Vol. 20, No. 10, pp. p857–864, oct 1979.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[18] 尾崎花奈, 小林一郎. 分散表現を用いたトピック抽出における確率的変分推論法適用への一考察. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 34, pp. 209–213, 2018.

[19] 小野田崇, 坂井美帆, 山田誠二. k-means 法の様々な初期値設定によるクラスタリング結果の実験的比較. 人工知能学会全国大会論文集, Vol. 25, pp. 1–4, 2011.

[20] Fred Glover. Surrogate constraints. *Operations Research*, Vol. 16, No. 4, pp. 741–749, 1968.

[21] 仲川勇二, 疋田光伯, 鎌田弘. 代理双対問題を解くためのアルゴリズム. 電子通信学会論文誌 A, Vol. 67, No. 1, pp. p53–59, jan 1984.

[22] 伊藤直也, 米澤朋子, 岡田佑一, 仲川勇二. 非線形ナップサック問題に対するグローバルグリーディ法. 2018 年度 情報処理学会関西支部 支部大会 講演論文集, 第 2018 巻, sep 2018.

[23] 井田憲一, 菅良平, 玄光男. ナップサック問題のための探索範囲調節型 ga の提案. 電気学会論文誌C（電子・情報・システム部門誌）, Vol. 124, No. 9, pp. 1861–1867, 2004.