Vol.2020-ICS-198 No.8
2020/3/8

Twitter上での知識獲得をねらってつながる最適コミュニティの検討

伊藤　直也[1,a)]　岡田 佑一[2]　米澤 朋子[3,b)]

概要：本稿では，ユーザがSNSのつながりを通じ特定のジャンルの知識を獲得することを目指し，適切なレベルや範囲の知識を有した人たちを集めたコミュニティを形成する他ユーザとのつながりの最適化手法を提案する．Twitterユーザのアカウントを大量に集め，各アカウントの知識量を算出し，アカウント数やツイート数を制約として与えながら知識量を最大化する最適化問題を解くことでコミュニティ構成を決定する．被験者実験の結果から，提案する最適化手法により得られたユーザ群はランダムに選ばれたユーザ群と比較して有意に当該ジャンルの知識が表れていると解釈されたことが示された．

キーワード：最適化，Twitter，データマイニング

1. はじめに

近年，SNS(social network service)の普及によって，場所・時間を選ばずに，さらに多人数で，容易にコミュニケーションすることが可能になってきた．Twitter[*1]では2017年10月に国内のアクティブユーザ数が4500万であったと報告されている．総務省 [1] によると，SNSを利用することで利用者は様々なメリットを感じており，「新しいつながりの創出」，「既存のつながりの強化」，「情報の収集」，「暇つぶし」が主として挙げられている．特にTwitterは140文字以内の文字制限により気軽に投稿することが出来ることで読み手の負荷が軽いことや情報量が多様である点，また，リアルタイムな情報の収集に適している点にユーザがメリットを感じやすいと考えられる．

このようなメリットの一方で，SNS上でのコミュニケーションは不特定多数との交流になりがちであるため，ユーザは情報過多や返信の義務感に追われることになり，SNS疲れ [2] [3] に陥ってしまう可能性がある．これは情報拡散力が高いTwitterに特にみられる．本稿では情報過多や不特定多数のつながりによる煩わしさからユーザが疲弊することをSNS疲れと定義する．

我々はある特定の目標で情報を得たいユーザに着目し，SNS疲れを軽減しつつ，あるトピック（例えばサッカー等）に関する知識を持っている人とつながることが出来るシステム [4] を提案した．提案システムではTwitter上にいる人のトピックに対する知識量の数値化を行う．SNS疲れの原因を考慮するため投稿頻度とつながり数に制約を与えつつ，ユーザとつながる人の持つ知識量が最大となるような最適化問題を解き，つながる人を決定する．本稿では，提案システムの概略を示すとともに，提案システムを用いた被験者実験により，ユーザが知識を持つユーザ群をつながり候補として提示されたかを検証した．

2. 関連研究

本研究のシステムは，ユーザとつながる他ユーザを探す点において，レコメンド機能の一つとも捉えられる．本節ではSNS上でのレコメンドを行っている関連研究を示す．Twitterを用いてユーザに他ユーザを推薦するシステムやツイートを推薦するシステム [5–7] も多い．

2.1 ユーザのレコメンド手法

田沼ら [8] はTwitterを始めたばかりのユーザが興味のある分野のツイートを発見するのは困難であるとし，特定分野に「濃い」アカウントの発見手法を提案している．ある分野に関する語を含む割合が高いアカウントを選び，そのアカウントとコミュニケーションを取っているアカウントとで行われているやり取りに着目している．特定分野の語句の発話傾向を散布図に表し，閾値を用いることで主観評価を行い，濃いアカウントを判断している．主観評価に

[1] 関西大学大学院総合情報研究科，高槻市
[2] トータス株式会社，高槻市
[3] 関西大学総合情報学部，高槻市
[a)] k016171@kansai-u.ac.jp
[b)] yone@kansai-u.ac.jp
[*1] https://twitter.com/TwitterJP

おいて，ある程度の精度で濃いアカウントを発見することが出来たことが示されている．

井上ら [9] は活動時間帯と活動量を考慮したつながり支援を行っている．活動時間帯と活動量がユーザと類似しているものでないと，ユーザ間の交流が行えないため，ストレスを感じる可能性があるとしている．TwitterAPI から得られた情報を基に，リプライを送りあっているユーザ達の活動時間帯や活動量は類似していることを示した．それまで多くの研究で使用されていたユーザの趣味・興味を取り入れずユーザの活動形態を利用したユーザ推薦の結果，日常的なツイートにおいてもユーザは他ユーザに興味，関心を抱くことが示された．日常的なツイートであるため，多くの人が共通して話題に取り上げるものであり，特にリアルタイムに反応するためには，活動時間帯が同じであったほうが良いことが示された．

川口ら [10] は利用者の興味に関係があるタイムラインを多く表示するようなユーザ評価推薦手法を提案している．SNS は匿名性が高いことからネガティブな言葉が多くみられる．そのため投稿内容の感情的要素を点数化し，自分と比較した時に許容できる範囲であるかを検証している．自分の投稿内容と相手の投稿内容の感情値の集合がどのように交わるかにより，フォローした際に交流がうまくいくかの点数化を行っている．

2.2 本研究の新規性

本節で示した関連研究と本研究との異なる点は，コミュニティのようなつながりを目的とした複数人の他ユーザを組み合わせによる総合的な観点からユーザに推薦することが出来る点である．これまでの SNS ユーザ推薦の研究では 1 人を推薦する研究が多かった．また複数を推薦する場合でも，その複数人のうちの誰かとのつながりを持つというものであった．しかし本研究ではユーザが望む利用頻度，人数を満たしたコミュニティとなりうる複数のつながりを推薦することが出来る．これにより SNS 疲れを引き起こさずにユーザが希望する人とのつながりを持つことが出来ることが期待される．

3. 提案システム

3.1 提案システム概要

本稿ではあるトピックに対して情報を持った人たち (コミュニティ) とつながることを目的とする．本研究では TwitterAPI から得られた実際のデータをもとに 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解析を行った．

分析対象とするアカウントはそれぞれのスポーツ名を呟いたことがあり，ツイート数が 50 以上ある日本語で書かれている約 6000 のアカウントとした．対象アカウントのツイート内容，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 API を用いて収集した．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．

提案システムではユーザとつながる人の持つトピックに対する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．

以下の図 1 に提案システムを示す．

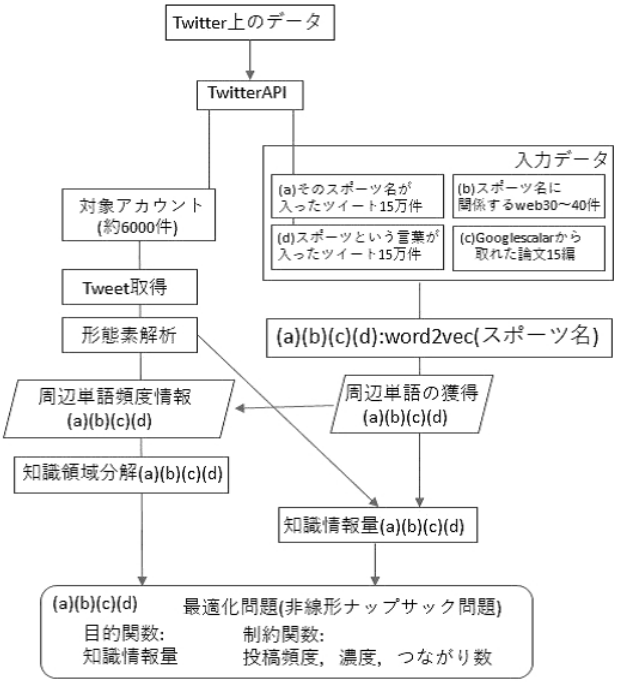

図 1 提案システム図

3.2 word2vec

各スポーツの知識を表している単語を定義するため，word2vec を用いた．word2vec にデータを与えると，検索単語に対して類似値が高い単語が得られる．本研究では word2vec により得られた上位 50 語の単語を周辺単語と定義した．図 1 に示した 4 つの入力データを用い，どの入力データ時に妥当性の高い単語が類似単語として得られたのかを 5.1 節で示す被験者実験にて検証した．入力データ d については 4 節で説明を行う．

3.3 知識情報量

周辺単語の呟き頻度を用いて各アカウントの知識情報量を決定する．知識情報量は以下の式で定義する．

$$\sum_{i=1}^{50} similar * log(count_i + 1)$$

$similar$ は類似値，$count_i$ はその周辺単語の頻度を表している．1 つの周辺単語による知識情報量の過剰な増加を防ぐために対数関数:log を使用した．

3.4 知識領域クラスタの生成

ユーザが複数の他ユーザとコミュニケーションをとる際，他ユーザ間の持っている情報が重複しないほうがユーザが得られる知識の範囲も広いと考えられる．そのため周辺単語の頻度情報を用いてクラスタリングを行うことで，同じような知識を持っているユーザを各クラスタでまとめられると考えた．ユーザと関わる他ユーザを各クラスタから選ぶことで情報の重なりが少ないコミュニティが実現できると考えられる．

3.5 最適化

3.3 節で定義した知識情報量の最大化を行うため，非線形ナップサック問題として定式化する．

3.5.1 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, . . . , 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．

非線形ナップサック問題を解くためのアルゴリズムは我々が提案したグローバルグリーディ法 [11] を用いた．

3.5.2 問題の作成

目的関数を 3.3 節において定義した知識情報量とし，制約関数を一日の平均ツイート数，濃度，つながる人数とする．平均ツイート数を制約として入れることにより，提案されるコミュニティは全体として投稿頻度が制限されるため，ユーザは自分の期待する速度でコミュニケーションを行い SNS を利用することが出来る．

本稿で実験を行う時に使用したコミュニティは一日の平均ツイート数上限値=80 であり，濃度上限値=16，つながり数=8 と設定し得られたコミュニティである．濃度とは 1 ツイート中に熟語，もしくはカタカナ語がどれくらい入っているのかを示すものである．濃度上限値を変えることにより，コミュニティ全体で Twitter 特有の意味のない文章が減ると考えられる．

4. 上位概念の導入

周辺単語の質を向上させるために単語の上位概念を検討する．本節では，前節で word2vec の入力データ d としたデータの生成方法をサッカーを例として説明する．

4.1 上位概念の定義

本稿における上位概念とは，その単語 a の集合の中に単語 b がすべて含まれている関係を指す．本稿ではサッカーの上位概念をスポーツと定義した．概念についての関係図を図 2 に示す．

4.2 word2vec の性質

word2vec は入力データとして文章を与えられた際に文章中の単語の分散を計算することで分散が近いものを意味が似ているものとして解釈している．そのため，サッカーという単語を含んだツイートを入力データとすると，入力データのいたるところでサッカーという単語が出てくるため単語の分散が正確に計算されなかったのではないかと考えられる．本節で示す上位概念の単語を含んだツイートを入力データとすることで入力データの一部に，類似語を探したい単語が出てくると考えられ，単語の分散が正確に計算され，周辺単語の質の向上が期待される．サッカーという単語を含んだツイートから得られた周辺単語を図 3 に示し，スポーツという単語を含んだツイートから得られた周辺単語を図 4 に示す．

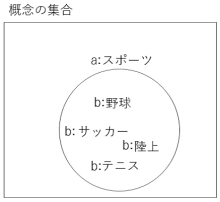

図 2 上位概念の図

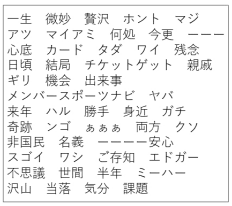

図 3 入力データ (サッカーという単語を含んだツイート)

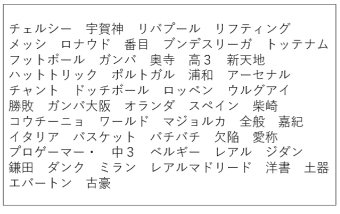

図 4 入力データ (スポーツという単語を含んだツイート)

意味をもたないノイズとなる語が減り，サッカーと関係のありそうな単語が多く得られたと考えられる．

4.3 周辺単語の改良

本節では，図 4 の周辺単語を改良するアルゴリズムを図 5 に示す．

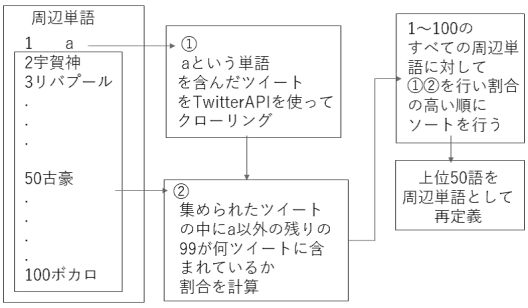

図 5 周辺単語改良アルゴリズム

サッカーに関係のある単語は共起しやすいと考えられるため，サッカーに関係のある単語の割合は高くなり，一般語は割合が低くなると考えられる．

5. 被験者実験提案システムの有効性検証

5.1 実験 1:周辺単語の妥当性検証

目的

本実験では，周辺単語がそのトピックを表しているかを調査する．もし周辺単語が，そのトピックを全く表していなければ周辺単語を呟く回数により求められる情報量を使った最適化計算は意味をなさないことになる．周辺単語 50 語を被験者に見せ，質問項目に答えてもらうことで，周辺単語の妥当性を調査する．

実験仮説　実験参加者は

(1)　入力データ (a4:スポーツという単語を含んだツイート) から得られた周辺単語はトピックに対する知識を表している．

(2)　入力データ (a1:スポーツ名を含んだツイート) から得られた周辺単語はノイズが多い．

と感じられる．

実験条件

word2vec の入力データによる違いを見るため、以下の 1 要因 4 条件を設けた，被験者内計画である．word2vec から得られた周辺単語にスポーツ名が入っていた場合，その単語を除いた周辺単語 50 個を被験者に提示した．除いたスポーツ名は `https://www.ssf.or.jp/dictionary/tabid/884/Default.aspx` に載っている約 200 種類のスポーツ名を対象とした．

a1: 入力データ [Twitter] によって出された周辺単語
a2: 入力データ [Web] によって出された周辺単語
a3: 入力データ [論文] によって出された周辺単語
a4: 入力データ [スポーツ] によって出された周辺単語

サッカー，野球，バスケ，テニスの 4 種類のスポーツにおいてそれぞれ a1–a4 の周辺単語を作成した．各条件はカウンタバランスを考慮して提示した．

実験参加者

実験には 19–27 歳の情報学部の大学生 16 名（女性 1 名，男性 15 名，mean=21.93，SD=1.55）が参加した．

実験手順

まず初めに，練習用として作成した 50 個の適当な単語が記されている紙を用いて実験の流れを説明した．実験参加者に質問項目を確認してもらい，以下の手順で実験を行うことを説明した．

実験参加者は以下のような流れで実験を進める．

(1) 同意書へ署名する
(2) 実験の説明を聞く
(3) 指示に従って紙を閲覧する
(4) スポーツ名を回答する
(5) 答えとなるスポーツを確認する
(6) もう一度紙を閲覧する
(7) 質問項目に答える

被験者は手順 (2)(6) の計 2 回周辺単語を閲覧するが，どちらも時間を区切らずに被験者が閲覧できたと感じたら紙を実験者に渡すように教示した．質問がある場合には，実験者へ質問するよう促した．

評価方法

以下の Q1-Q4 の質問項目に対し，1. あてはまらない，2. ややあてはまらない，3. どちらでもない，4. ややあてはまる，5. あてはまる，の 5 段階で回答させ，MOS（平均オピニオン評点，mean opinion score）を得る．Q0 は記述式で回答させた．得られた MOS に対して一要因分散分析（有意水準 α=0.05）を行う．

Q0　何のスポーツに関する単語群だと思いますか
Q1　そのスポーツに関連する単語が多いと感じる
Q2　そのスポーツを連想する単語が多いと感じる
Q3　そのスポーツに関連しない単語が多いと感じる
Q4　そのスポーツを連想しない単語が多いと感じる

また評価方法として被験者が閲覧した 50 語の単語それぞれに対して以下の 4 つのいずれかで評価を行った．

- そのスポーツに関係があると思われる
- そのスポーツに関係がないと思われる
- 単語の意味が分からない
- そのスポーツに単語が関係あるかどうか分からない

表 1 実験 1：1 要因分散分析の結果

	F	p	多重比較
Q1	4.361	0.0089	a2,a4>a1
Q2	2.937	0.0433	–
Q3	1.413	0.2514	–
Q4	1.689	0.1828	–

結果

Q1–Q4 の各質問項目に対する回答の平均とその標準誤

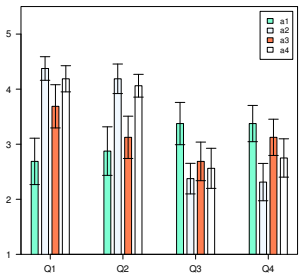

図 6 実験 1：Q1–Q4 結果の平均と標準誤差

表 2 実験 1:周辺単語の単語それぞれの評価 (χ 二乗検定)

	*1	*2	*3	*4
a1	307 ▽	329 ▲	129	35 ▽
a2	433 ▲	211 ▽	128	28 ▽
a3	402 ▲	271	106	21 ▽
a4	342 ▽	224 ▽	128	106 ▲

（ ▲有意に多い, ▽有意に少ない,$p < .05$）

表 3 各条件における被験者の回答データ

	a1	a2	a3	a4
正解数	5	13	12	16
間違えた数	11	3	4	0

表 4 Fisher 検定による p 値

	a1:a2	a1:a3	a1:a4	a2:a3	a2:a4	a3:a4
正解数:間違えた数	0.056	0.127	0.000	1	0.451	0.303

差を図 6 に示す．また，分散分析と多重比較の結果を表 1 に示す．周辺単語の各単語の評価は各条件で回答されたものを集計したデータに χ 二乗検定を行った結果を表 2 に示す．Q0 のスポーツ名の正解数を比較するため，コクラン・ルールに基づいて Fisher 正確確率検定を行った．Q0 における正解，不正解の結果を表 3 に示し，Fisher 検定の結果を表 4 に示す．

表 2 の*1–*4 はそれぞれ*1:関係があると思われる，*2:関係がないと思われる，*3:関係があるかどうか分からない，*4:単語の意味が分からない，を示している．

Q1 において $a2 > a1$ と $a4 > a1$ において平均値が有意に高かった．このことから Web を入力データとした時と，スポーツという単語を含んだツイートを入力データとした時は，そのスポーツ名を含んだツイートを入力データ時よりも，よりそのスポーツに関係している単語が多いということが示された．Q2–Q4 間では有意差は見られなかった．各単語の評価に対する χ 二乗検定において，’スポーツに関係がある’ では a2 と a3 が有意に高く，’スポーツに関係がない’ では a1 が有意に高かった．a4 は’単語が関係あるかどうか分からない’ が有意に高いことが示された．

考察

質問項目 Q1 では $a2, a4 > a1$ であることから a2:web と a4:上位概念から得られた周辺単語が a1 よりスポーツに関係しているという結果が得られた．$a4 > a1$ という結果が得られたため，上位概念を導入することで a1:スポーツ名を含んだツイートを入力データにするよりも周辺単語の質が，良くなることが示された．また各単語に関する χ 二乗検定の結果から a1 のそのスポーツに関係ないと判断される単語が有意に高かったため，ノイズが多いことが示唆される．また a4:上位概念の’単語が関係あるかどうか分からない’ が有意に高いことが分かる．これは一般的に知られていないサッカーに関する単語が多かったことが示唆される．マジョルカなどサッカーに関連しているが，一般の人では分からない単語があることが原因だと考えられる．a3 から得られた周辺単語は熟語が多くあったため，単語の意味が分からないと回答する人は少なかったが，周辺単語は，スポーツに関する一般語が多くあったため，あるスポーツ限定の知識を表しているとは言えないと考える．Fisher 検定の結果から，a4>a1 において有意な差が見られることから，a1 よりも a4 は，スポーツを想起させる単語が多かったといえる．

結果から a2:Web と a4:上位概念を考慮したデータから，良い周辺単語が得られると考えられる．web は略称が少ない傾向があるため得られた周辺単語と，省略された語句が多い Twitter 上で一致する語句を探すことは困難であると考えられる．そのため実験 2–3 では上位概念考慮した入力データ:a4 を基に 4.3 節の改良アルゴリズムにより得られた周辺単語を用いてコミュニティの最適化を行った．

5.2 実験 2:周辺単語を用いた最適化コミュニティの有効性の検証

目的

本実験では提案システムで示した知識情報量最大化を目的とする最適化を行い，得られたコミュニティからトピックの知識を感じられるかを検証する．実験 2 では対象スポーツをサッカーとした．

実験仮説

実験参加者は

(1) 最適化により得られたコミュニティはサッカーに対する知識・情報量が他の条件よりも感じられる．

(2) 周辺単語を呟いているコミュニティは，ランダムで選ばれたコミュニティよりもサッカーの知識・情報量を感じられる．

と感じられる．

実験条件

周辺単語の有無，最適化の有無によって，被験者がコミュニティの情報をどのように感じるかを調べるため実験条件は以下の 1 要因 3 条件の被験者内計画とした．b1 では知識情報量が最大になるコミュニティを最適化によって決定する．b2 は乱数で無作為に選んだユーザのコミュニティとした．b3 は無作為に選ぶが b2 とは異なり，周辺単語を呟いたことがあるアカウントを対象とした．

b1:最適化によって選ばれたコミュニティ

b2:ランダムに選ばれたコミュニティ
b3:周辺単語を呟いているコミュニティ
各条件はカウンタバランスを考慮して提示した．

実験素材

b1–b3 すべての条件においてコミュニティ人数を 8 人と設定した．コミュニティメンバーの名前を a さん–h さん，と設定し発言内容とともにコマンドプロンプトに映し出した．コミュニティメンバーはそれぞれ 15 ツイート発言を行うが，発言順序はランダムなものを被験者に提示した．15 ツイート中の周辺単語を含んだ割合は，そのユーザの全ツイートのうち周辺単語を呟いている割合と同じものとした．a さん–h さんの部分だけ色を変えることで，今読んでいる内容は誰が発言しているものかを被験者に分かりやすく提示した．

実験参加者

実験には 19–27 歳の情報学部の大学生 24 名（女性 2 名，男性 22 名，mean=21.70，SD=1.48）が参加した．

実験手順

実験参加者は以下のような流れで実験を進める．
(1) 同意書へ署名する
(2) 実験の説明を聞く
(3) コミュニティメンバーの発言を閲覧する
(4) 質問項目に答える

質問がある場合には，実験者へ質問するよう促した．実験 1 と同じく，コミュニティが話しているスポーツ名を書いてもらい，正解となるスポーツを提示する．被験者に正解となるスポーツを確認してもらった後，読み直す時間はないということを伝えた．そのため被験者に閲覧する際には，あらゆるスポーツの可能性を考えながら読むように教示した．また被験者には全体の傾向を評価の対象とするため，1 つずつの内容は軽く読み進めるように教示した．このように教示することで，被験者は 1 つの内容に固執せずに全体の印象を評価することが出来ると考えられる．

評価方法

以下の Q1-Q9 の質問項目に対し，実験 1 と同様の評価方法・分析手法を用いた．

Q0 何のスポーツについて話しているコミュニティだと思いましたか
Q1 コミュニティはサッカーに関連することを話していると感じた
Q2 コミュニティ内にサッカーに興味を持っていそうな人を多く感じた
Q3 コミュニティ内でサッカーの知識を感じ取れた
Q4 コミュニティ内の人と関わることで自分はサッカーの知識を得られそうだと感じた
Q5 サッカーに関するコアな内容の会話をしている
Q6 コミュニティに対して親しみを感じる
Q7 サッカーに対して興味を持つことが出来た
Q8 コミュニティ内の各ユーザのサッカーについての情報発信量が多いように感じた
Q9 コミュニティ全体でサッカーについての情報量が多いように感じた

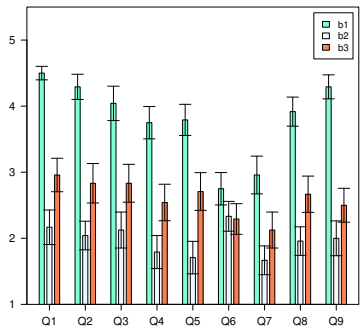

図 7 実験 2：Q1–Q9 結果の平均と標準誤差

表 5 実験 2：1 要因分散分析の結果

	F	p	多重比較
Q1	41.917	0.0000	$b1 > b3 > b2$
Q2	24.456	0.0000	$b1 > b3 > b2$
Q3	18.007	0.0000	$b1 > b2, b3$
Q4	22.864	0.0000	$b1 > b3 > b2$
Q5	19.366	0.0000	$b1 > b3 > b2$
Q6	2.846	0.0683	–
Q7	11.431	0.0001	$b1 > b2, b3$
Q8	19.806	0.0000	$b1 > b3 > b2$
Q9	30.646	0.0000	$b1 > b2, b3$

結果

各質問項目に対する回答の平均とその標準誤差を図 7 に示す．また，分散分析と多重比較の結果を表 5 に示す．Q0 に対しては被験者 24 人中 23 人がすべての条件でサッカーと回答しており，残りの 1 名は b1 でフットサルと回答していた．

分散分析の結果から，Q6 を除いたすべての質問項目において有意差があった．有意差が出たすべての質問項目の多重比較において $b1 > b2, b3$ という関係が見られた．また Q3，Q7，Q9 においては $b3 > b2$ という関係が見られなかった．Q7 では b1 の平均が他条件より有意に高いことが示されているが，b1 の回答平均は 2.958 であることからサッカーに対して興味を持つことが出来たとは言い難い．Q5 において，コアな内容の会話をしている，では b1 の回答平均値が 4 に近く，被験者はコアな内容を話しているコミュニティと感じることが出来たことが示唆される．

考察

提案手法で最適化されたコミュニティ (b1) で，Q6 を除いた全ての質問項目に平均回答値が他の条件よりも有意に高いことが示された．このことから提案手法により集めたユーザ群のツイートはサッカーについての知識を話してい

る／よりコアな内容を話していると評価されたと考えられる．コミュニティ全体としてのサッカーに関する知識量に関わる Q3，Q7，Q9 において $b3$ と $b2$ に差はなかったものの，Q1，Q2，Q4，Q5，Q8 では $b3 > b2$ という結果が得られ，個々のコミュニティメンバーに着目すると b3 では情報発信量は多く感じることが示唆される．そのため周辺単語を含むツイートはそうでないものと比べて，サッカーに興味を持つまでには至らないが，知識を得られそうだとは感じられたと捉えられる．

本研究ではコミュニティを形成する他ユーザ間でのコミュニケーションは考慮していないため，親しみを持たれなかったと考えられる．しかし Q6 の有意傾向から，知識量が多いコミュニティでは親しみが発生があることが示唆される．以上のことから，最適化で他ユーザを選ぶ提案手法はよりサッカーに対する知識が多く表れていると捉えられ，周辺単語を呟くコミュニティはランダムなユーザ選択コミュニティに比べ，個々の情報発信量が多いと捉えられる傾向から，仮説 1，2 はそれぞれ支持されたと考える．

5.3 実験 3:知識領域クラスタを用いた知識の広がりの検証

目的

本実験では，周辺単語に基づいた知識領域クラスタ分けを行うことで，ユーザは知識の広がりを感じられるかどうかの検証を行った．本実験で，対象とするスポーツは野球とした．50 語の周辺単語を呟いている頻度をもとにユーザをクラスタに分けた時に，クラスタ間からユーザが選ばれることによって使われる周辺単語の傾向が変わる．そのため，ユーザはコミュニティメンバーの知識の重なる領域が少なく，知識の広がりを感じることが出来ると考えた．k-means 法を使ってクラスタ分けを行ってから最適化したものと，得られたユーザデータをランダムに並べ 10 分割してクラスタを形成し，最適化を行ったものの 2 つで比較を行った．

実験仮説

実験参加者は

(1) k-means を使ってクラスタを形成し，最適化されたコミュニティの方が知識の広がりを感じる．

と感じられる．

実験条件

周辺単語によるユーザのクラスタ分けの有無が知識の広がりを感じるかを調べるため，実験条件は以下の 1 要因 2 条件の被験者内計画とした．

c1:k-means 法を用いて周辺単語頻度情報を基にクラスタを作成し，最適化することで選ばれたコミュニティ

c2:データ総数を 10 分割することによって得られたクラスタを最適化し，選ばれたコミュニティ

c1 では k-means 法を用いて周辺単語の頻度に応じてユーザを 10 のクラスタに分け，クラスタから 8 人選び出すことでコミュニティとする．c2 ではユーザデータをランダムに並べ，総数を 10 分割してクラスタを 10 個形成した．8 人選び出すため必要最低限のクラスタ数は 8 であるが，クラスタを 10 個に分けたほうがより各クラスタが有する情報の重なりが少なくなると考えられたため，クラスタ数を 10 に決定した．クラスタ数が多ければ多いほど各クラスタで持つ情報の重なりは少なくなるが，10 以上になると 1 人しかいないクラスタが複数個形成されたため 10 とした．各条件はカウンタバランスを考慮して提示した．

実験参加者

実験には 19–27 歳の情報学部の大学生 24 名（女性 2 名，男性 22 名，mean=21.70，SD=1.48）が参加した．

実験手順

実験参加者は以下のような流れで実験を進める．

(1) 実験の説明を聞く
(2) コミュニティメンバーの発言を閲覧する
(3) 質問項目に答える

被験者に対する教示内容は，実験 2 と同じである．

評価方法

以下の Q1-Q6 の質問項目に対し，実験 1 と同様の評価方法・分析手法を用いた．Q0 も同様に記述式で回答させた．

Q0 何のスポーツについて話しているコミュニティだと思いましたか
Q1 コミュニティ内の人々が共通した知識を持っていると感じた
Q2 コミュニティ内の人々の発言は野球についてのものだと感じた
Q3 コミュニティ内の情報にバリエーションを感じた
Q4 コミュニティ内の人々の会話がかみ合っていると感じた
Q5 コミュニティの人々が情報交換をしているように感じた
Q6 発言者それぞれが異なる得意分野があるように感じた

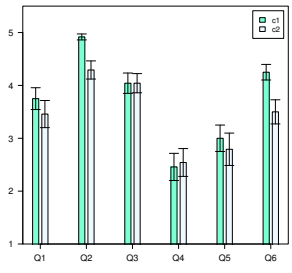

図 8 実験 3：Q1–Q6 結果の平均と標準誤差

結果

各質問項目に対する回答の平均とその標準誤差を図 8 に示す．また，t 検定の結果を表 6 に示す．Q0 に対しては被

表 6　実験 3:t 検定の結果

	t	p
Q1	0.86708	0.3904
Q2	3.3874	0.0021
Q3	0	1.0000
Q4	-0.22181	0.8254
Q5	0.51607	0.6083
Q6	2.7029	0.0064

験者 24 人中 24 人がすべての条件で野球と回答していた．表 6 に示す通り Q2 と Q6 において有意差が示された．Q1 においては c1，c2 ともに野球のことについて話しているコミュニティであるため，有意差は見られなかった．Q2 では有意差が出たが，平均回答値を見るとどちらも 4 を上回っているため最適化の結果としては実験 2 と同様の効果が得られたということが示唆される．Q3 ではそれぞれの条件の回答平均値が同じであり，4 を超えており，どちらの条件もバリエーションを感じられるようであった．Q4–Q5 では，有意差が出ずにどちらの条件も回答平均値が 3 より低いことから，各ユーザの投稿内容をそれぞれ提示するだけでは会話や情報交換をしているように被験者には見えないことが示唆される．

考察

Q6 では c1 の回答平均値が有意に高いことが示された．このことから仮説 1 が支持されたと考えられる．Q3 では，コミュニティ全体としての情報のバリエーションはどちらも，同程度感じることが出来たことが示唆される．これはコミュニティメンバーの発信した情報の深さをバリエーションとして捉えた時に，情報量が各条件ともに最適化されているため条件間で深さの違いを感じ取れなかったが可能性がある．Q6 で c1 の回答平均値が高いことから，わずか数個の関連のある投稿内容からコミュニティ内の発言者の得意分野の違いがあるように感じられたことが示され，k-means 法による周辺単語の頻度情報を用いたユーザのクラスタ分けは効果があったことが示唆される．

50 語の周辺単語自体に場所・選手名・チーム名などの各単語が何を表しているかがはっきりとし，ノイズとなる単語がもっと少なければコミュニティ内の情報にもバリエーションを感じられたのではないかと考えられる．特定の知識に対する周辺単語の分野分けは今後の課題である．

6. おわりに

本稿では Twitter において知識を得るのに最適なつながりを提案するシステム [4] の検証を被験者実験を通して行った．このシステムでは，word2vec によって得られた単語を周辺単語と定義し，周辺単語を呟いている量により各ユーザの知識量を算出した．

入力データによって結果が変わる word2vec に対し，適切な入力データを比較検討した結果，a2:Web と a4:上位概念を考慮した Twitter データを用いた時に良い周辺単語が得られる可能性が示された．次に，提案手法の最適化によって選ばれたコミュニティが比較条件に対して有意にその知識を含んでいるという結果が示された．さらに，周辺単語頻度情報に応じ k-means 法でクラスタリングを適用し，各クラスタから他ユーザを選出するアルゴリズムが知識の広がりを感じさせるという結果も得られた．

今後はシステムに提案されたユーザ群と長期間つながることによるユーザおよび他ユーザの変化の検証や，動的なつながりの変容を含むシステムの改良を課題とする．

謝辞

本研究は一部科研費 19H04154，19K12090，18K11383，2570021 の助成を受け実施したものである．

参考文献

[1] 総務省　ソーシャルメディア利用のメリット .https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/html/nd142220.html.
[2] 植田浩章, 孟暁順, 吉田直人, 米澤朋子. エージェントの伝聞口調による sns 発言伝達手法における相手抽象化と代理効果の検証. Technical Report 3, 関西大学大学院総合情報学研究科, 関西大学大学院総合情報学研究科, 関西大学大学院総合情報学研究科, 関西大学総合情報学部, mar 2016.
[3] 加藤千枝.「sns 疲れ」に繋がるネガティブ経験の実態 : 高校生 15 名への面接結果に基づいて (研究). 社会情報学, Vol. 2, No. 1, pp. 31–43, 2013.
[4] 伊藤直也, 米澤朋子. 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. 2019 年度 情報処理学会関西支部 支部大会 講演論文集, 第 2019 巻, sep 2019,G-114.
[5] 原克彬, 中山泰一. つぶやきに基づいたユーザ推薦システムの提案. 第 74 回全国大会講演論文集, 第 2012 巻, pp. 637–638, mar 2012.
[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] 渡邊恵太, 加藤昇平. トピックモデルと協調フィルタリングに基づくユーザ興味を反映した情報推薦システム. 人工知能学会全国大会論文集, Vol. JSAI2014, pp. 2M34–2M34, 2014.
[8] 田沼勇輝.Twitter における特定分野に「濃い」アカウントの発見手法. *DEIM Forum 2011*, 2011.
[9] 井上翔太, 櫨山淳雄. 活動時間帯と活動量を考慮した twitter でのつながり構築支援手法とつながり構築支援システムの開発とその評価. 情報処理学会研究報告. HCI, ヒューマンコンピュータインタラクション研究会報告, Vol. 2014, No. 56, pp. 1–6, mar 2014.
[10] 川口辰弥, 塚田晃司.Twitter における利用者の興味に即したタイムラインを構成するユーザー評価推薦手法の提案. Technical Report 38, 和歌山大学大学院システム工学研究科, 和歌山大学システム工学部, jan 2019.
[11] 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.