Twitterユーザの関心の表現に適したデータ構造の検証

松岡　紀行† 　中村　達哉† 　白川　真澄† 　原　隆浩† 　西尾章治郎††

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1 番 5 号
†† 大阪大学　〒 565–0871 大阪府吹田市山田丘 1 番 1 号
E-mail: †{matsuoka.noriyuki,nakamura.tatsuya,shirakawa.masumi,hara,nishio}@ist.osaka-u.ac.jp

あらまし　近年，Twitter を情報源とした関心抽出の研究が盛んに行われている．既存研究では，ツイートに含まれる単語をユーザの関心の構成要素であると仮定し，ツイートから抽出した単語を特定のデータ構造に当てはめてユーザの関心を表現しているが，ユーザの関心の表現方法自体について詳細な考察や分析は行われてこなかった．そこで本稿では，Twitter ユーザの関心の表現に適したデータ構造の検証を目的として行った実験について述べる．実験では，データ構造間でできるだけ公平な評価を行うために，機械的に関心を抽出するのではなく，被験者が自身の投稿したツイート上の単語を用いて関心を表現し，それぞれのデータ構造で自身の関心をどれだけ正確かつ詳細に表現できたかを評価した．実験によって得られたデータをもとに，それぞれのデータ構造が表現できる関心の性質を分析した．

キーワード　関心抽出，Twitter，データ構造，ユーザプロファイル

1. 序　　論

近年，情報推薦や検索結果のパーソナライズなどを目的として，関心抽出に関する研究が注目されている．ユーザがどのような関心をもっているかを把握することにより，その関心に合わせてユーザに提示する情報を絞り込むことが可能となる．最近では Twitter ユーザの関心抽出に関する研究が注目を浴びている [6], [8], [9], [10]．特に Twitter は，全世界で 3 億を超えるユーザが自身の関心事をツイートと呼ばれる最大 140 文字のテキストとしてリアルタイムに投稿しており，ユーザの日常的でリアルタイムな関心を抽出できる情報源として期待されている．

Twitter ユーザの関心抽出において考慮すべき点がユーザの関心の表現方法である．関心の表現方法によって，表現可能な関心の性質が異なる．例えば，テニスに関心をもつユーザに対して，テニスラケットに関する情報を推薦することを考える．その際，そのユーザがテニスの観戦にのみ関心があったならば，その推薦はユーザにとって適切でないと考えられる．また，ユーザは常に同じ情報を推薦されることを好むわけではない．自動車に対して強い関心をもつユーザであっても，ある時点においてサッカーの試合を観戦している場合，その時点では自動車に関する情報よりもサッカーに関する情報を推薦する方がそのユーザにとって有益であると考えられる．このように，関心抽出では，ユーザの心理的な性質も考慮しながら，関心を適切に表現することが重要である．

Twitter ユーザを対象とした関心抽出に関する既存研究では，機械的な処理の容易さの観点から，関心を表現するための基本要素として，ユーザが投稿したツイート中の単語を用いることが一般的である．例えば，既存研究の多くは，ツイートに含まれる単語を抽出し，TF-IDF（Term Frequency-Inverse Document Frequency）[7] などの単語の重み付け手法を用いて特徴語としての度合を算出することで，Twitter ユーザの関心を単語のベクトルとして表現している [8], [9]．また，LDA（Latent Dirichlet Allocation）[3] を用いて，類似した関心を表す単語集合（グループ）を発見し，ユーザの関心を複数のグループとして抽出する研究も行われている [6], [10]．

この二つの表現方法の大きな違いは，単語間の関係性に関する情報も含めてユーザの関心を表現しているデータ構造であるか否かという点である．また，他に単語間の関係性を表現するデータ構造として，任意の単語ペアの関係の有無を表すグラフ構造などが考えられる．これらのデータ構造には，それぞれ関心を表現する上で長所や短所などの特性があると考えられ，データ構造の特性が関心を利用するアプリケーションの性能に影響する．そのため，データ構造の特性を把握し，表現したい関心に合わせて適切なデータ構造を選択することが重要である．しかし，各データ構造が表現可能なユーザの関心の性質の調査やデータ構造間での比較に関する研究はこれまで行われてこなかった．

そこで本研究では，Twitter ユーザの関心を適切に表現可能なデータ構造について検証するため，被験者を用いた実験を行う．実験では，Twitter ユーザの関心を，そのユーザが投稿したツイート中の単語を構成要素として，代表的なデータ構造を用いてそれぞれ表現し，どのようなデータ構造であれば Twitter ユーザの関心を適切に表現できるかを評価する．実験では，それぞれのデータ構造の表現能力を十分に活かして関心を表現するため，被験者が自身のツイートをもとにそれぞれのデータ構造を用いて自身の関心を表現し，自身の関心がどの程度正確かつ詳細に表現できているかをデータ構造ごとに評価する．そして，得られた結果を分析し，それぞれのデータ構造が表現できる関心の性質を明らかにする．また，認知心理学における研究をもとに定義した関心の性質と分析結果を照らし合わせる．

上記の実験によって得られたデータ構造の特性は，ユーザの関心を理想的に表現できた場合の結果であり，機械的に関心抽出を行った場合は，データ構造ごとの表現のしやすさの影響を受ける．そこで，各データ構造を用いて理想的に表現された関

心に対し，可観測な情報のみからどの程度その関心を再現できるかを調査することにより，機械的に関心抽出を行った場合にどのデータ構造が関心の表現に適しているかを評価する．

2. Twitter ユーザによる実験

以下ではまず，本研究における関心の定義，および本実験で調査の対象とするデータ構造について述べる．その後，Twitter ユーザを被験者として行った実験について詳述する．

2.1 関心の定義

ある対象に関心をもっているかどうかの定義として，その対象について積極的あるいは自発的に行動を起こす（例えば，考え事をしたり，話題に挙げたり，調べ物をしたりするなど）ことがある場合，その対象に関心をもっているとする．また，認知心理学に関する文献 [5] を参考に，本研究で取り扱う関心の性質に関する以下の定義を行った．

2.1.1 関心の局所性

本実験では，関心の対象は概念であるとする．つまり，ある人の関心が「テニス」という単語単体で表現される場合，認知心理学における概念の「テニス」に対する関心を表現していると考える．ここで概念とは，個々の事物，事象に共通した性質を取り出して得られる表象（心の中に表現された情報やその表現形式）である．概念が意味することを明確に言語化することは難しく，概念同士の境界も明確ではない．例えば，ある人が「テニス」と「漫画」のような異なる概念に対して関心をもつ場合であっても，その人が「テニスに関する漫画」に関心をもっていれば，その人にとって「テニス」と「漫画」は明確な境界を設けられる関心ではない．本研究ではこのような関心の性質を「関心の局所性」と定義する．

2.1.2 関心の遷移性

ユーザがある時点である関心の対象（概念）に従って行動していることを，ユーザはその対象に意識を向けていると便宜的に表現する．本研究では，認知心理学におけるプライミング効果を参考に，ユーザがある時点である対象に意識を向けていた場合に，そのユーザは次の時点でその対象に関係のある別の対象に意識を向ける可能性が高いと考える．例えば，ユーザがある時点で「サッカー」の試合を TV で観ていたならば，次の時点でそのユーザが「サッカー」のゲームを始める可能性の方が，「テニス」のゲームを始める可能性よりも高いと考える．なお，どの関心の対象同士の関係が近いかは，人によって異なる．本研究ではこのような関心の性質を「関心の遷移性」と定義する．

2.2 実験において調査するデータ構造

本研究では，認知心理学および関心抽出に関する既存研究を参考に，以下の四種類のデータ構造を調査する．なお，データ構造自体の比較を行うため，単語の重みなどの程度に関する情報は明示的には表現しないが，被験者は自由に程度の大きさを想像で補完できるものとする．

2.2.1 単 語 集 合

単語集合は，ツイート中に出現する単語のうち，ユーザの関心を表す単語を要素とした一つの集合で定義される関心の表現方法である．単語集合は単語間の関係性を表現できないデータ構造であるが，機械的な処理が容易である．関心抽出の既存研究においては，このデータ構造の単語をさらに関心の強さで重み付けしたベクトル表現が良く用いられている [8], [9].

2.2.2 グ ラ フ

グラフは，活性化拡散モデル [4] を参考に，ツイート中のユーザの関心を表す単語をノード，単語間の関係の有無を無向エッジで表現した無向グラフとして定義される関心の表現方法である．グラフは，他のデータ構造と比較して，認知心理学分野における人の認知機能のモデルを最も反映したデータ構造である．

2.2.3 グ ル ー プ

グループは，LDA [3] などのトピックモデルを参考に，一つ以上の単語を要素とした複数の単語集合（以降グループにおける単語集合をセットと呼ぶ）で定義される関心の表現方法である．一つの単語は一つのセットにのみ属することができる (注1)．各セットは，セットに属する単語のみで区別され，セットがもつ意味を表すようなラベルは付与されない．LDA ではそれぞれの単語が何らかのトピックに属しているため，それぞれのトピックは単語を要素とした集合として捉えることができる．

2.2.4 階層グループ

グループではセットが単語のみを要素とするのに対し，階層グループは，セットがセットを要素としてもよいようにグループを拡張した関心の表現方法である．すなわち，セット間で階層関係を表現できる．一つの単語は一つのセットにのみ属することができる (注1)．このデータ構造は階層的なトピックモデル [2] を参考にしている．

2.3 実験の方針

Twitter ユーザのツイートからデータ構造ごとにユーザの関心を抽出し表現する方法として，ツイートの統計情報や既存の関心抽出技術を用いて機械的に抽出する方法と，ユーザ自身がツイート中に出現する語句を用いて関心を表現する方法がある．本章における実験では，各データ構造自体の表現力を公平に比較するために，Twitter ユーザ自身が被験者となり，ツイート中の単語を用いてデータ構造ごとに自身の関心を表現する方法を採用した．これにより，各データ構造を用いて理想的に関心を表現できたときのデータ構造間の比較が可能となる．

被験者のツイート中に出現する全ての単語について，その中から自身の関心を表現しうる単語を選び出し，それぞれのデータ構造に当てはめていく作業は，被験者の負担が非常に大きい．そこで本実験では，被験者が関心を表現するための Web アプリケーションを実装し，実験課題中における被験者の負担が軽減されるように配慮した．図 1 に Web アプリケーションの主な操作画面図を示す．Web アプリケーションでは，単語の重み付け手法である TF-IDF を用いて，被験者のツイートから 150 個の特徴的な単語を抽出し，被験者がその中から自身の関心に含まれる単語だけを使って関心を表現する方式を採用した．なお，単語抽出の対象とするツイートには，実験開始日時を起点として最大で過去 3,000 件のツイートを用いた．

(注1)：実験では，一つの単語が複数のセットに属してもよい場合についても検証したが，紙面の都合上，本稿では結果を割愛する

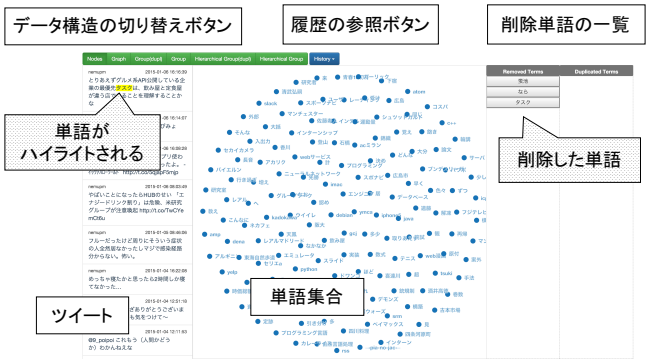

図 1　Web アプリケーションの主な操作画面図

2.4　関心を表現する際の基準

2.4.1　実験課題の終了基準

被験者には各データ構造を用いて最終的に自身の関心が最も正確かつ詳細に表現できるまで実験を行うよう指示した．この基準を設けることで，各データ構造の制約内で被験者が自身の関心を理想的に表現したとみなすことができる．

2.4.2　関心をもっているかどうかの判断基準

2.1 節で述べたように，ある対象について積極的あるいは自発的に行動を起こすとき（例えば，考え事をしたり，話題に挙げたり，調べ物をしたりするなど），その対象に関心をもっていると被験者に判断するよう指示した．

2.4.3　関心の変化の表現基準

関心抽出に関するいくつかの既存研究では，関心が時間に従って変化すると定義している [1]．しかし，関心の時間による変化も含めて表現すると実験の操作が煩雑となるため，本実験では関心の変化を考慮しなかった．つまり，被験者が判断の対象となっている単語を含むツイートを投稿した時点でその単語が表す対象に関心をもっていれば，被験者にはその対象に関心をもっていると判断するよう指示した．

2.4.4　関心の強さの表現基準

一般的に関心の強さはその対象によって異なる．既存研究では，単語の重み付け手法を用いて，関心の強さをその関心を表す語句の重みとして表現している [8], [9]．しかし，関心の強さも含めて表現すると実験の操作が煩雑となるため，本実験では関心の強さを考慮しなかった．ただし，被験者には自由に関心の強さを想像で補完できることを伝えた．

2.5　実 験 手 順

本実験では，ツイート数 961〜21,443 の 20 代の大学生や大学院生，社会人の Twitter ユーザ 10 名を被験者として実験を行った．実験では操作の慣れによる評価への影響を抑えるために，被験者 A〜E には単語集合，グラフ，グループ，階層グループの順で，被験者 F〜J にはグラフのみ順序を最後に変更して実験を行ってもらった．グラフのみ順序を変更しているのは，グループをベースとしたデータ構造に対し，グラフの操作や表現方法が大きく異なり，操作の慣れによる影響が大きい可能性があると判断したためである．各被験者について，以下の手順で実験を行った．

（1）　被験者に対して実験の目的や手順について説明する．

表 1　データ構造の総合的評価の設問項目

設問 1	あなたが表現した関心の図を見て，あなたの関心が「正確かつ詳細に表現できているか」を評価してください．[1（表現できていない）〜10（表現できている）]
設問 2	あなたが表現した関心の図の中で，データ構造の機能を使って特にうまく表現できた（できなかった）ことは何ですか．[記述]

表 2　関心の二性質についてのデータ構造の評価の設問項目

設問 3	関心を表現する際に関心の局所性を考慮しましたか．[はい/いいえ]
設問 4	関心を表現する際に関心の遷移性を考慮しましたか．[はい/いいえ]
設問 5	（設問 3 ではいと答えた被験者のみ）関心の局所性をどの程度表現できていますか．[1（表現できていない）〜10（表現できている）]
設問 6	（設問 4 ではいと答えた被験者のみ）関心の遷移性をどの程度表現できていますか．[1（表現できていない）〜10（表現できている）]
設問 7	（設問 3 と設問 4 ではいと答えた被験者のみ）関心の局所性と関心の遷移性を総合的に見てどの程度表現できていますか．[1（表現できていない）〜10（表現できている）]

（2）　被験者自身のツイートから抽出された単語集合を構成要素として，被験者が 2.4 節の基準を参考にしながら，自身の関心をそれぞれのデータ構造を用いて表現する．

（3）　四種類のデータ構造を用いて関心を表現した後，被験者が，自身が表現した関心の図を見ながら表 1 に示すアンケートに回答する．

（4）　被験者が，2.1 節の関心の二性質に関する説明を受け，表 2 に示すアンケートに回答する．

3.　Twitter ユーザによる実験の結果と考察

表 3　設問 1「正確かつ詳細に関心を表現できたか」の評価結果

被験者	A	B	C	D	E	F	G	H	I	J	Avg
単語集合	3	3	2	1	3	5	2	2	4	5	3.0
グラフ	4	8	4	2	5	8	7	4	7	6	5.5
グループ	6	8	6	7	7	6	5	7	6	7	6.5
階層グループ	9	8	8	9	9	8	7	8	7	8	8.1

表 3 に設問 1 の評価結果を示す．これより，被験者によってデータ構造の評価が大きく異なることが分かる．設問 1 の評価結果を基に各データ構造の特性を考察するにあたって，各データ構造を表 4 に示す四つの機能（単語集合機能，グラフ機能，グループ機能，階層機能）の観点で整理する．

表 4 の四つの機能はそれぞれ表現できるユーザの関心の性質が異なると考えられる．表現したい関心の性質を表現可能な機能をもつデータ構造を用いることで，ユーザの関心を適切に表現することができる．一方で，各データ構造はより多くの機能をもつほど複雑化し，機械処理が難しくなると考えられる．そのため，各データ構造が表現できるユーザの関心の性質を把握し，表現したい関心の性質に応じて必要最小限の機能をもつデータ構造を選択することが重要となる．以下では，実験結果から各機能がどのような関心を表現できるかを考察する．

表 4　各データ構造がもつ機能

データ構造	単語集合機能	グラフ機能	グループ機能	階層機能
単語集合	○			
グラフ	○	○		
グループ	○		○	
階層グループ	○		○	○

3.1　単語集合機能の評価結果と考察

表 3 より，全ての被験者において，単語集合以外のデータ構造の評価値が単語集合の評価値よりも高くなっている．また，全ての被験者が単語集合に対して 5 以下の低い評価値をつけている．このことから，単語集合機能のみでは関心を適切に表現できない可能性が高い．

一方で，10 人中 9 人の被験者が単語集合の評価値に 2 以上の回答をしていることから，単語集合機能のみで表現できる関心があると言える．一つの例として，単語集合を一つのグループとみなすことで，ある分野に対する関心を単語集合全体として表現できると考えられる．実際に，被験者 B は自身の最大の関心である「麻雀」の関連語が多く抽出されており，単語集合でも麻雀に特に関心があることがある程度表現できたと回答した．

また，もう一つの例として，単語単体が指しうる概念が比較的明確である場合があると考えられる．例えば，被験者 J が作成した単語集合には漫画の作品名が多く含まれていた．ほとんどの人にとって漫画は読んで楽しむ娯楽作品であるため，漫画の作品名に関する単語単体が，その漫画を読むことに関して関心があることを十分に表現できていたと考えられる．一方で，多くの単語は，単語単体では何に関心をもっているのかを十分に表現できない．例えば，被験者 G の単語には「パナソニック」という電機メーカーを表す単語が含まれていた．会社名によって表現される関心として，その会社の製品に関心があることが考えられるが，実際に被験者 G に聞いたところ，就職先の候補として関心をもっていると答えた．

以上のことから，単語集合機能についてまとめると，単語集合全体としてある分野に関心をもっていることや，単語単体として関心を表現できる場合がある一方，関心を正確かつ詳細に表現するという観点では，単語集合機能のみでは十分に表現できない点が多いと考えられる．

3.2　グラフ機能の評価結果と考察

表 3 より，全ての被験者で単語集合よりグラフの評価値が高かった．このことから，グラフ機能はユーザの関心を正確かつ詳細に表現する上で有効であると考えられる．グラフ機能は，任意の単語ペアの関係の有無を明確に表現できるため，複数の単語を用いることではじめて対象が限定される粒度の細かい関心を表現できると考えられる．例えば，被験者 B は「リンク」というゲームのキャラクターが「スマブラ」というゲームで「崖」に掴まっている状況に関心をもっていることを，それら三つの単語を互いにエッジで接続することで表現できたと回答した．また，ある単語に他の多くの単語とのエッジを接続することで，その単語が自身の関心の中で特に重要な関心を表すことが表現できると考えられる．実際に被験者 E や被験者 F，被験者 I は，被エッジ数によってその単語の重要度を表すことができたと回答した．

グラフ機能の注目すべき特性として，エッジが特定の単語に偏って接続されることで，グラフ全体で見たときに密な部分と疎な部分が生じることが挙げられる．そのため，疎な部分を境にグラフを分割することで，分割された各部分が被験者の関心の対象である概念を表現できると考えられる．このとき，分割された各部分は疎に接続されているため，関心の対象を単語のまとまりとして表現しつつお互いの関係性も同時に表現できる．しかし，それぞれの関心の対象の境界が曖昧に表現されることに関して，本実験では被験者の意見が二つに分かれた．例えば，被験者 B は「スマブラ」というゲームと「麻雀」が「リスク」を介して繋がっていることを挙げ，自身の中で全く異なる関心の対象である「スマブラ」と「麻雀」の境界が曖昧に表現されることに対して違和感があると回答した．それに対して被験者 E は，自身の異なる関心の対象の間で関係性が表現されているのは，自身の関心の具体的な意味を良く表現できていると肯定的に回答した．このように，グラフは関心の対象の境界の曖昧性をエッジによって表現できるが，一方で，そのことが関心を正確かつ詳細に表現する上で有利に働くとは限らない．

また，密な部分のエッジ数が多くなることで，単語間の関係性が複雑になりすぎる場合があることが分かった．被験者 C は，エッジが多くなるとグラフが具体的に自身のどのような関心を表現しているか分からなくなったと回答した．実際に，グラフにおける単語の被エッジ数の平均とグラフの評価には負の相関傾向が見られ，相関係数を算出すると -0.70 であった．このことから，グラフは，エッジが増えると被験者にとってそれが自身のどのような関心を表しているのか判断するのが困難になったと考えられる．

以上のことから，グラフ機能では，複数の単語によって粒度の細かい関心を表現したり，被エッジ数によって関心の重要度を表現したりできることが，単語集合機能のみの場合よりも高い評価につながったと考えられる．また，互いに境界が曖昧な複数の関心の対象を表現できるが，このことは場合によっては関心を正確かつ詳細に表現する上で不利に働く可能性があることが分かった．エッジの数が多すぎると，どのような関心を表現しているかが曖昧になるという欠点も見受けられた．

3.3　グループ機能の評価結果と考察

表 3 より，10 人中 5 人の被験者のグループの評価値が 7 以上であった．また，全ての被験者においてグループの評価値が単語集合の評価値よりも高かった．このことから，グループ機能はユーザの関心を正確かつ詳細に表現する上で有効であると考えられる．

グループ機能では，セット（一つ以上の単語集合）によって様々な粒度の関心を表現することができると考えられる．例えば，被験者 G は研究室生活に関する様々な単語を一つのセットに含めることで,「研究室生活」という大まかな関心を表現できたと回答した．セットに含まれる具体的な単語群によって関心を表現することで,「研究室生活」という語句単体で表現した場合と比べて，より具体的に被験者自身がもっている「研究室生

活」に関する関心を表現することができたと考えられる．

また，グループ機能はセット内の単語が互いに関係があることを表現すると同時に，セット内の単語がセット外の単語と関係が無いことも表現しているため，関心の対象となる概念を明確に区別して表現できると考えられる．実際に，被験者 E は概念の境界を明瞭に表現することで，自身がどのような関心をもっているかをより明確に表現できたと回答した．

一方で，被験者 C は，グループでうまく表現できなかった点として，セット内の単語が多くなると，そのセットが具体的に自身のどのような関心を表現しているかが分からなくなったと回答した．実際に被験者 C は,「ごはん」というラベル(注2)のセットに 22 個の単語を加えており，中には様々なジャンルの食べ物や複数の地名が含まれていたため，そのセットが表現している関心が曖昧になったと考えられる．

以上のことから，グループ機能では，様々な粒度の関心を具体的な構成要素（単語）により表現できることや，異なる関心の対象を明確に区別して表現できることが，関心を正確かつ詳細に表現する上で有効に機能していたといえる．一方で，セット内の単語が多くなると，そのセットがどのような関心を表現しているかが曖昧になる場合があった．

3.4 階層機能の評価結果と考察

表 3 より，10 人中 9 人の被験者において，グループの評価値よりも階層グループの評価値が高かった．このことから，階層機能を用いることで関心をより正確かつ詳細に表現できることが分かった．階層機能を用いることで，関心を複数の粒度で表現できると考えられる．被験者 C は，大量の単語を含む「ごはん」というセットの下に「麺類」「洋食」「場所」などのセットを作成し，階層化することで自身の関心がより正確に表現できたと回答した．この被験者は，グループにおいてセットに単語を加えすぎることでセットが表現する関心が曖昧になっていたと回答していたが，階層化機能によってそのような問題が解決できたことを示している．

また，被験者 D や被験者 E は，グループで表現していたセットを被験者 C と同様に分割したと同時に，それらのセットを新しく生成した上位階層のセットの要素に加えていた．例えば，被験者 D は「生放送ラジオ」というセットを「プラモデル」というセットの要素に加えていた．これにより，被験者 D はプラモデルや生放送ラジオに対する関心に加えて，プラモデルに関係のある生放送ラジオにも関心をもっていることを表現できたと考えられる．このように，階層機能を用いて，セットで表現された関心をさらに組み合わせることで，様々な粒度の重複する関心を表現できると考えられる．

以上より，階層機能では，様々な粒度の重複する関心を正確かつ詳細に表現できたことが高い評価値につながったと考えられる．

3.5 関心の局所性および遷移性に関する考察

設問 3〜7 により，2.1 節で仮説を立てた関心の二性質「関心の局所性」,「関心の遷移性」について，被験者が関心を表現する際にこれらの性質を考慮しているのか，およびそれぞれのデータ構造がこれらの性質をどの程度適切に表現可能なのかを評価した．表 5 に設問 5〜7 の回答結果を示す．なお，本実験を行った際，設問 3「関心の局所性を考慮して自身の関心を表現したか」について全ての被験者が「はい」と回答したのに対し，設問 4「関心の遷移性を考慮して自身の関心を表現したか」に対して「はい」と答えた被験者は G,I,J の三人のみであった．そこで，設問 6,7 の考察材料を増やすために，被験者 B と被験者 D には，本実験終了後に，関心の遷移性を考慮しながら再度図の修正と評価を行う追加実験を行った．以下では，関心の局所性および関心の遷移性のそれぞれの観点から各データ構造の特性について再考し，上記の二性質の妥当性を検証する．

3.5.1 関心の局所性とデータ構造の関係

設問 5「関心の局所性をどの程度表現できたか」に対する 10 人の被験者の回答を参考に，データ構造の各機能が関心の局所性を表現できる場合とできない場合について考察する．

単語集合については，全ての被験者が今回調査したデータ構造の中で最も関心の局所性を表現できなかったと回答した．一方で，被験者 I と被験者 J は単語集合に対して評価値 6 と回答した．これは単語集合機能が，単語単体あるいは単語集合全体として関心の対象となる概念を表現することがあり，その場合には関心の局所性をある程度表現できたと評価されたためであると考えられる．

グラフについては，10 人中 5 人の被験者が 7 以上の評価値を回答している．概念間の境界が曖昧であるという関心の局所性の性質を，グラフ機能がエッジによって表現できたためであると考えられる．一方で，被験者 A と被験者 D はグラフに対して評価値 2 と回答している．これは，概念の境界を曖昧なまま表現することにより，関心の対象自体が局所的にまとまっていることを適切に表現できなかったためであると考えられる．

グループについては，全ての被験者が 6 以上の評価値を回答した．これは，本来境界が曖昧な関心の対象に明確な境界を設けることで，関心の対象となる概念が局所的にまとまっていることを表現できたためであると考えられる．一方で，セット内の単語が多くなると，そのセットがどのような関心を表現しているかが曖昧になる場合があった．これは，関心の対象に境界を設定することが難しい場合があり，うまく境界を設定できない場合には関心の局所性が表現できないことを示している．

ほとんどの被験者が，階層グループに最も高い評価値を与えていた．これは，グループにおいて関心の対象に境界を設定することが難しい場合に，階層機能を用いてセットを階層化することで解決できることがあり，その場合には関心の局所性をより適切に表現できたためであると考えられる．

3.5.2 関心の遷移性とデータ構造の関係

設問 6「関心の遷移性をどの程度表現できたか」に対する 5 人の被験者の回答を参考に，データ構造の各機能が関心の遷移性を表現できる場合とできない場合について考察する．

単語集合については，5 人中 3 人の被験者が評価値 1 と回答した．これは，単語集合機能のみの場合は単語間の関係性の情

(注2)：実験中は便宜上，被験者がセットを構築しやすくするためにラベルを付与できる．

報をもたないために，関心の遷移性を全く表現できなかったことを意味していると考えられる．一方で，被験者 I と被験者 J は単語集合に対してそれぞれ評価値 4 と 5 を回答した．これは，全単語間で遷移があるとみなすことで関心の遷移性をある程度表現できていると評価されたためであると考えられる．

グラフについては，被験者 G と被験者 J がそれぞれ 9，8 と高い評価値を回答した．これは，エッジによって概念間の遷移関係を表現できたためであると考えられる．一方で，被験者 D は評価値 1 を回答した．その理由として被験者 D は，今回定義したグラフが無向グラフであるために遷移の方向を表現できず，具体的に概念間をどのように遷移していくかを全く表現できなかったと感じたためと答えた．また，被験者 D は関心の局所性についても低い評価値をつけている．関心の局所性が表現できていない，すなわち，遷移関係を表すべき関心の対象自体が表現できていない場合，関心の遷移性についても表現できないと捉えることができる．

グループについては，5 人中 4 人の被験者が 5 または 6 の評価値を回答した．これは，グループ機能ではセット間の遷移は表現できないが，セット内の単語で表される複数の概念間の遷移は表現できたためであると考えられる．一方で，被験者 D は評価値 1 を回答した．これは，被験者 D がそれぞれのセットを一つの概念と捉えており，概念間の遷移を表現できなかったためであると考えられる．

階層グループに対しては 5 人中 3 人の被験者が 8 または 9 の高い評価値を回答した．これは，セット内で表現される遷移に加えて，階層関係にあるセット間でも遷移を表現できたためであると考えられる．

3.5.3 関心の局所性および遷移性の妥当性の検証

各データ構造が関心の局所性と関心の遷移性を表現できることが，被験者の関心を正確かつ詳細に表現できることととの程度関係性があるかを考察する．表 3 と表 5 より，同じデータ構造に対する設問 1 と設問 5，設問 1 と設問 7 の評価値は相関が高いことが分かる．このことから，関心の局所性は，人の関心を正確かつ詳細に表現する上で重要な性質であるといえる．また，関心の局所性と関心の遷移性の二性質を表現することも，関心を正確かつ詳細に表現する上で重要であるが，関心の遷移性は関心の局所性に対して重要度は相対的に低いと考えられる．

4. 再現実験

機械的に関心を抽出し表現した場合に，各データ構造がどの程度ユーザの関心を正確かつ詳細に表現できるのかを調査し，3. で得た結果と比較を行った．

4.1 手法の構築方針

再現実験では，被験者のツイート中に含まれる単語を用いて，グラフ，グループ，階層グループの三種類のデータ構造で被験者の関心を表現する手法を構築する．本実験では，被験者が 2. で述べた実験で自身の関心の表現に用いた単語を入力とし，それらの単語間に関係性を加えて出力する手法を構築した．実際のアプリケーションの処理には，ツイートから単語を抽出する段階も含まれるが，本実験では，結果の考察においてデータ構造の比較を行うため，データ構造の生成以外の処理を理想的に行うことで，実験結果に影響を与える要因を絞っている．

表 5　関心の局所性と関心の遷移性についてどの程度表現できたかに関する設問 5〜7 の評価結果．

	被験者	B	D	G	I	J	A	C	E	F	H
設問 5	単語集合	3	1	2	6	6	1	1	2	4	1
	グラフ	8	2	7	7	6	2	4	8	8	4
	グループ	7	7	6	6	7	7	7	9	6	6
	階層グループ	9	10	7	8	8	10	9	8	7	8
設問 6	単語集合	1	1	1	4	5					
	グラフ	5	1	9	6	8					
	グループ	5	1	6	6	5					
	階層グループ	8	4	8	9	6					
設問 7	単語集合	3	1	2	5	5					
	グラフ	8	2	8	8	7					
	グループ	7	4	6	6	6					
	階層グループ	9	8	7	9	7					

グラフやグループなどを生成する手法を構築するには，単語間の関係の強さを計算する評価尺度が必要である．評価尺度を選択する上で，手法のチューニングをあまり必要としないことや一般的に用いられる評価尺度であることなどを重視し，Pointwise Mutual Information（PMI）に基いて手法を構築した．PMI では，任意の単語ペア x, y について自己相互情報量 $pmi(x; y)$ を以下の式により計算する．

$$pmi(x; y) = \log \frac{p(x, y)}{p(x)p(y)} \tag{1}$$

ここで，$p(x, y)$ は単語 x と単語 y が同一文書内で共起する確率，$p(x)$ と $p(y)$ はそれぞれ文書に単語 x と単語 y が単独で出現する確率であり，文書の単位は手法の目的などによって異なる．なお，本実験では PMI を計算するための情報源として用いるコーパスに，被験者自身のツイートだけでなく，外部知識として Wikipedia や他の Twitter ユーザのツイート（Twitter）の三種類を用いた．Wikipedia のデータとして 2014 年 11 月 22 日時点の日本語版 Wikipedia の約 94 万エントリのテキストデータを用意し，Twitter のデータとして 2015 年 10 月 1 日〜2015 年 10 月 7 日に収集した約 398 万の日本語ツイートを用いた．PMI の計算の際には，Wikipedia では一エントリを一文書として，Twitter や被験者のツイートでは一ツイートを一文書として計算した．また，これらの組合せによって以下の七種類の手法を構築した．

（1） Wikipedia のみを用いる手法（手法 W）
（2） Twitter のみを用いる手法（手法 T）
（3） 被験者のツイートのみを用いる手法（手法 O）
（4） Wikipedia と Twitter を用いる手法（手法 WT）
（5） Twitter と被験者のツイートを用いる手法（手法 TO）
（6） Wikipedia と被験者のツイートを用いる手法（手法 WO）
（7） Wikipedia，Twitter と被験者のツイート全てを用いる手法（手法 WTO）

再現実験では，これら七つの手法それぞれについて，グラフ，グループ，および，階層グループの三つのデータ構造について

手法を構築し，計 21 の手法を評価した．

4.2 手法の構築

4.2.1 グラフ生成手法

入力された単語集合に対して優先度の高い単語ペアを順にエッジで接続し，指定されたエッジの数になった時点でのグラフをユーザの関心として出力する手法を構築した．単語ペアの優先度の定義とエッジの数の決め方を以下に述べる．

a） 単語ペアの優先度

コーパス間では PMI の大きさに差があり，比較が難しい．そこで，各コーパスごとに，PMI の高い順に単語ペアに順位を付け，どのコーパスによる順位かは関係なく順位の小さい単語ペアほど優先度が高いものと定義する．優先度の高い単語ペアを順にエッジで接続する．

b） エッジの数

エッジの数は本手法のパラメータであり，不適切な値を設定すると被験者の評価が大きく下がってしまうことが考えられる．そこで本実験では，被験者が 2. で述べた実験で作成したグラフのエッジ数を，本手法で設定するエッジ数に設定した．

4.2.2 グループ生成手法

入力された単語集合に対して PMI のスコアに基づく距離関数を用いた階層的クラスタリングを適用し，得られた単語クラスタ群をグループとして出力する手法を構築した．以下では，単語間の距離の定義や，クラスタ間の距離関数の定義，および，閾値の決め方について述べる．

a） 単語間の距離

本手法では，PMI による順位を距離として用いた．これは，グラフのときと同様に，コーパス間で PMI の大きさに差があるためである．複数のコーパスを用いる手法の場合は，一つの単語ペアに対して複数の順位が与えられるため，単純に単語ペア間の距離の大小を比較することができない．そこで本実験では，単語ペアに割り当てられた順位の内，最も上位の順位を距離として定義した．なお，比較した順位同士が等しい場合は，次に高い順位同士を比較することで距離の大小を定義した．

b） クラスタ間の距離関数

クラスタ間の距離関数として最遠隣法を用いた．最遠隣法はクラスタ間の任意の単語ペアのうち，最長の距離をクラスタ間の距離として用いる手法である．2. の実験において被験者がグループで自身の関心を表現した際に，単語数の近いセットを複数作る傾向があったことから，各クラスタの大きさが揃いやすい最遠隣法が適切であると判断した．

c） 閾　　値

入力として与えられた単語集合に対して全ての単語ペアの距離を計算し，その中央値を閾値として定義した．この閾値を用いることで，距離の遠すぎる単語ペアが最遠隣法によって併合されることを防ぐ．

4.2.3 階層グループ生成手法

入力された単語集合に対して，グループと同様に一階層目のクラスタ群を生成した後，二階層目以降は別の距離関数を用いて，指定された階層まで，指定された数のクラスタを各階層に生成する手法を構築した．二階層目以降に用いる距離関数と階層数および各階層のクラスタ数の決め方を以下に述べる．

表 6　グラフの評価結果（関心を表現できているか）

被験者	手法 WTO	手法 W	手法 T	手法 O	手法 WT	手法 WO	手法 TO	2. の実験の評価
B	4	**5**	1	1	3	3	4	8
C	3	2	2	2	3	**4**	3	4
D	1	1	1	1	1	1	1	2
E	4	4	6	**7**	5	5	5	5
F	5	3	6	3	3	3	4	8
G	6	5	7	**8**	**8**	6	5	7
H	4	2	3	**5**	2	3	**5**	4
I	**8**	4	4	4	7	3	**8**	7
J	5	8	3	7	3	**9**	6	6
Avg	4.44	3.78	3.67	4.22	3.89	4.11	**4.56**	5.67

a） 二階層目以降の距離関数

二階層目以降の距離関数には最近隣法を用いた．一階層目で閾値までクラスタの併合を終えた場合，残るクラスタ群の内，任意の二クラスタ間の最長距離は全て閾値となるため，ほとんど関係のない単語ペアによりクラスタが併合されやすい．一方で，最近隣法ではクラスタの併合を最短距離を用いて行うため，PMI の順位が高い単語ペアによりクラスタが併合される．その結果，最遠隣法よりも意味的に近い二クラスタを併合できると考えられる．

b） 階層数とクラスタ数

階層数は，被験者が 2. で述べた実験で作成した階層グループにおける階層数と同数とした．二階層目以降は被験者が作成した階層グループの各階層におけるセット数と等しくなった段階で併合を止め，次の階層のクラスタ生成に移行する．また，二階層目以降は一段下の階層で得られたクラスタのみが併合されるようにし，他階層のクラスタ数が変化しないようにした．

4.3 実 験 手 順

2. の実験の被験者 10 人の内 B〜J の 9 人に再現実験に参加してもらった．各被験者について，以下の手順で実験を行った．

（1） 被験者が，機械的に再現した関心の図を，自身が表現した関心の図と合わせて閲覧した．

（2） 被験者が，自身が表現した関心の図につけた評価と比較しながら，提示された図が自身の関心を「正確かつ詳細に表現できているか」を 10 段階（1:表現できていない〜10:表現できている）で評価した．

（3） 以上の手順を，三種類のデータ構造と七種類の情報源の組合せによって構築された計 21 手法について繰り返した．

5. 再現実験の結果と考察

5.1 再現可能性の考察

5.1.1 グラフの再現可能性

表 6 にグラフの評価結果を示す．表 6 より，被験者ごとに最も高い評価値を達成した手法を確認すると，その評価値は，10 人中 6 人の被験者について 2. の実験の評価値以上となっている．このことから，被験者によっては PMI を算出するためのコーパスを適切に選択できれば，機械的に関心を表現した場合でも，被験者が自身の関心を表現した場合と同等かそれ以上に

表 7　グループの評価結果（関心を表現できているか）

被験者	手法 WTO	手法 W	手法 T	手法 O	手法 WT	手法 WO	手法 TO	2. の実験の評価
B	5	**6**	2	1	3	5	5	8
C	2	2	2	**6**	4	**6**	3	6
D	2	3	2	2	**5**	1	2	7
E	**8**	6	5	6	6	**8**	7	7
F	**7**	5	6	2	1	**7**	1	6
G	4	3	2	4	3	**7**	**7**	5
H	4	3	3	**6**	4	5	4	7
I	**6**	4	4	5	3	**6**	4	6
J	5	6	3	**8**	7	7	6	7
Avg	4.78	4.22	3.22	4.44	4.00	**5.78**	4.33	6.56

表 8　階層グループの評価結果（関心を表現できているか）

被験者	手法 WTO	手法 W	手法 T	手法 O	手法 WT	手法 WO	手法 TO	2. の実験の評価
B	**4**	**4**	3	3	3	**4**	3	8
C	3	4	**6**	**6**	**6**	5	3	8
D	**7**	2	1	4	6	5	4	9
E	**9**	6	5	7	7	**9**	8	9
F	4	5	5	4	**7**	5	2	8
G	**6**	3	3	5	3	5	3	7
H	5	6	6	**7**	5	6	6	8
I	4	**5**	3	4	3	4	3	7
J	5	5	3	5	5	**7**	5	8
Avg	5.22	4.44	3.89	5.00	5.00	**5.56**	4.11	8.00

正確かつ詳細に被験者の関心をグラフで表現できる可能性が高いことが分かる．

5.1.2　グループの再現可能性

表 7 にグループの評価結果を示す．表 7 より，被験者ごとに最も高い評価値を達成した手法を確認すると，その評価値は，10 人中 6 人の被験者について 2. の実験の評価値以上となっている．グラフの場合と同様に，PMI を算出するためのコーパスを適切に選択できれば，機械的にユーザの関心を再現できるといえる．

5.1.3　階層グループの再現可能性

表 8 に階層グループの評価結果を示す．表 8 より，被験者ごとに最も高い評価値を達成した手法を確認すると，その評価値が 2. の実験の評価値以上となった被験者は 10 人中 1 人であった．また被験者ごとに，階層グループで最も高い評価値をグループで最も高い評価値と比較した場合に，階層グループの方が高かった被験者は D,E,H，変わらなかった被験者は C,F，低かった被験者は B,G,I,J であった．2. の実験ではほとんどの被験者がグループよりも階層グループを高く評価していたが，再現実験においては，グループと階層グループの評価値に大きな差はないといえる．以上のことから，今回構築した機械的手法では，被験者の関心を階層グループとして再現することは難しく，階層機能を十分に活用できていないと考えられる．

6. 結　　論

本研究では，Twitter ユーザの関心を適切に表現するデータ構造の検証を目的とした実験を行った．実験では，関心抽出に関する既存研究を参考に，単語集合，グラフ，グループ，階層グループの計四種類のデータ構造を調査した．Twitter ユーザ 10 人を被験者とし，各データ構造を用いて自身の関心を表現した後，それぞれのデータ構造が自身の関心をどの程度正確かつ詳細に表現できているかを評価した．実験結果から，各データ構造の機能が表現可能な関心の性質についての考察を得た．また，認知心理学に基づき定義した関心の二性質の観点から実験結果を再考し，関心の局所性および関心の遷移性を表現することが，関心を正確かつ詳細に表現する上で重要であることが分かった．さらに，機械的に関心を抽出し表現した場合に，各データ構造がどの程度ユーザの関心を正確かつ詳細に表現できるのかを検証することを目的とし，グラフ，グループ，階層グループを対象とする実験を行った．実験の結果から，グラフとグループにおいては，手法をうまく設計できれば，機械的に関心を表現した場合でも，被験者が自身の関心を表現した場合と同等かそれ以上に正確かつ詳細に被験者の関心を表現できる可能性が高いことが分かった．

今後の課題として，今回行った実験の被験者数の拡大が挙げられる．また，本研究の成果を参考に抽出した関心を表現することで，情報推薦などのアプリケーションの性能を改善可能かどうかを検証することが挙げられる．

謝　　辞

本研究の一部は，文部科学省科学研究費補助金・基盤研究 (A)(26240013)，JST 国際科学技術共同研究推進事業（戦略的国際共同研究プログラム），および，文部科学省国家課題対応型研究開発推進事業－次世代 IT 基盤構築のための研究開発－「社会システム・サービスの最適化のための IT 統合システムの構築」の研究助成によるものである．ここに記して謝意を表す．

文　　献

[1] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. Smola, “Scalable distributed inference of dynamic user interests for behavioral targeting,” Proc. ACM SIGKDD Conf., pp.114–122, 2011.
[2] D.M. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum, “Hierarchical topic models and the nested chinese restaurant process,” Advances in Neural Information Processing Systems, vol.16, p.17, 2004.
[3] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” The Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.
[4] A.M. Collins, and E.F. Loftus, “A spreading-activation theory of semantic processing.,” Psychological Review, vol.82, no.6, p.407, 1975.
[5] 箱田裕司, 都築誉史, 川畑秀明, 萩原 滋, 認知心理学 (New Liberal Arts Selection), 有斐閣, 2010.
[6] D. Ramage, S.T. Dumais, and D.J. Liebling, “Characterizing microblogs with topic models,” Proc. ICWSM, 2010.
[7] G. Salton, and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing & Management, vol.24, no.5, pp.513–523, 1988.
[8] T. Vu, and V. Perez, “Interest mining from user tweets,” Proc. ACM CIKM, pp.1869–1872, 2013.
[9] W. Wu, B. Zhang, and M. Ostendorf, “Automatic generation of personalized annotation tags for twitter users,” Proc. Conf. of NAACL HLT, pp.689–692, 2010.
[10] W.X. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, and X. Li, “Comparing twitter and traditional media using topic models,” Proc. ECIR, pp.338–349, 2011.