3M4-2

データ研磨手法を用いたTwitterユーザの関係構造変化の検出

Text Summarization and Change Detection in Word Similarity Structure using Twitter Data

前川浩基 *1 Hiroki MAEGAWA　内田将史 *2 Masashi UCHIDA　大内章子 *2 Akiko OUCHI　宇野毅明 *1 Takeaki UNO　羽室行信 *2 Yukinobu HAMURO

*1国立情報学研究所 情報学プリンシプル研究系
Principles of Informatics Research Division, National Institute of Informatics

*2関西学院大学 経営戦略研究科
Institute of Business and Accounting, Kwansei Gakuin University

In this paper, we propose a new method to summarize the contents of text posted on Twitter about "parental leave", by which we successfully retrieve some interesting topics. The proposed method consist of two major parts. First, we summarize the Twitter text by retrieving some important cluster of words as a maximal clique on a similarity graph of words. We propose a novel approach, "graph polishing" in it. The similarity graph is reconstructed in order to reduce noise and clarify a latent structure on the original graph. Second, we visualize the change of Twitter contents in time series using a Sankey diagram, which is a visualizing method for directed acycric graph (DAG) stressing a flow of stream. In our case, vertex is corresponding to the word cluster, and we draw a stream as a degree of change of the word structure on consecutive days.

1. はじめに

男女雇用機会均等法 (1986 年)，育児休業法制定 (1992 年) など各種法律が整備され，女性の雇用環境は改善してきており，今やほとんどの企業が育児休業制度を導入している．それにもかわらず，第 1 子出産を機に有職女性の 3 分の 2 が退職しており，出産・育児を経た就業継続はいまだに困難である [3]．このような中，安倍政権が成長戦略の中核に女性の活用を据え，育児休業 3 年を企業に要請している．育児休業制度について，女性の就業継続を促進するという研究が多い一方で，企業が制度を取得する可能性のある女性の採用を抑制するといった否定的な研究結果も報告されており [5]，休業期間の長期化については意見が分かれている．

そこで本研究では，育児 (以下，育休) についての Twitter 投稿に注目し，一般の人々の声を要約することを試みる．具体的には，安倍政権の育休 3 年の要請という発言 (2013 年 4 月 18 日) によってユーザの話題がどのように変化したか，そして性別や子供の有無によって内容がどのように異なるかを解析する．Twitter ユーザの偏りから，そこで語られる内容が世論を代表するかどうかは疑わしいが，多様化するメディアの一つとして，育休に関する話題の解析は意義深いと考える．

話題の要約には，トピック抽出と文書要約の技術を用いる．トピック抽出においては，投稿数の急激な増加をバーストとして検知し，バースト時の投稿からトピックを抽出する方法が主流である [6, 2]．本研究でも，話題の変化検出に同様の手法を適用するが，単なる投稿頻度の確率分布の変化をモデル化するのではなく，ユーザが利用する単語の類似構造の時系列変化に注目する．そして構造変化を Sankey ダイアグラムによる視覚化することで話題の変化を検知する．

一方で話題の要約であるが，例えば，Filatova ら [1] は，少数の文書単位 (例えば文) でより多くの概念単位 (例えば単語) を被覆することを目的とした，ナップサック制約付最大被覆問題として定式化した．著者らも同様の方法を Twitter の投稿内容の要約に適用し一定の効果を確認している [4]．ただし，文書単位としては，ツイートから単語間の類似度グラフを構築し，その密部分グラフをクラスタとして複数抽出し，それらを文書単位として用い，単語クラスタを要約として出力している．本稿でも同様の手法を用いている．

連絡先: 羽室行信，関西学院大学経営戦略研究科，兵庫県西宮市上ケ原一番町 1-155，hamuro@kwansei.ac.jp

2. 手法

本節では，育休 3 年関連ツイートの分析で利用する手法について論じる．図 1 に示されるように，大きく 2 つの流れがあり，一つは，ツイートの投稿内容の変化を検出する方法であり，他方が，その変化が生じた前後で投稿された内容の変化を要約する方法である．いずれの手法においても，その中心はツイートに出現する単語の共起情報に基づいた単語の類似構造に着目し，その構造の変化を検知・要約するものである．以下では，そこで利用される中心的手法として，類似度グラフ，グラフ研磨，構造変化の視覚化について示す．

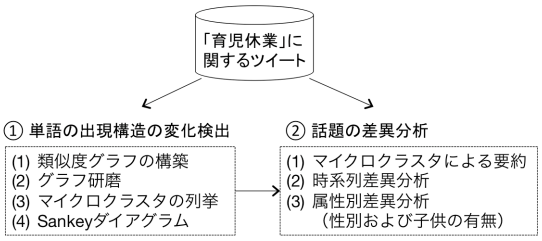

図 1: 分析の概略図

2.1 単語の類似度グラフ

まずツイートは，一週間を単位として一日ごとにずらした移動窓を設定し，それぞれの単位で単語の類似度グラフを構築する．単語の類似度グラフとは，単語を節点で表し，類似した単語間に枝を張った無向グラフである．ここで類似度は，単語の共

起情報に基づいて定義される基準化 PMI(NPMI:Normalized Pointwise Mutual Information) を用いる．2 つの単語 u, v について，各単語の出現確率を $\Pr(u), \Pr(v)$，2 つの単語の共起確率を $\Pr(u, v)$ で表すとすると，NPMI は式 1 の通り定義される．

$$sim(u, v) = \log \frac{\Pr(u, v)}{\Pr(u)P(v)} / (-\log \Pr(u, v)) \qquad (1)$$

これは，単語の出現を独立と考えた場合の共起確率に対する実際の共起確率の比を -1.0 から 1.0 に基準化したものである．この値が 0 より大きければ，2 つの単語は類似していると考える．

2.2 グラフ研磨

単語の類似度グラフは，互いに類似した単語群には密に枝が張られ，逆に類似度の低い単語群には枝は張られにくくなり疎な構造となる．そこで，類似度グラフの密な部分グラフを意味ある単位 (クラスタ) として抽出することで，ツイート内容の要約として利用することが可能であると考える．一般グラフのクラスタリングについては，ニューマンクラスタリング，グラフ分割，極大クリーク列挙など，これまでも様々な手法が提案されてきたが，どの手法も問題点を抱えており，決定打になっていないというのが現状である．

例えば極大クリーク列挙では，現実データにおいては多くの場合，非常に多数の類似した極大クリークが列挙されてしまうという問題がある．列挙された極大クリークの類似関係を用いて，極大クリークを更にクラスタリングするという方法も提案されているが，列挙される極大クリークの数によっては計算量が問題となる．このような問題の多くは，そもそも対象とするグラフにノイズが含まれるために起こる問題とも考えられる．

そこで，最近著者らは，対象とするグラフをクリーニングする「グラフ研磨」手法を提案している．これは，グラフをクラスタリングする前に，枝を張り直すことでグラフを再構成し，できる限り構造を明確化しておこうというものである．直感的には，図 2,3 に例示されるように，枝密度の濃い部分グラフはより濃く，薄い部分グラフはより薄くするというものである．そして，グラフ研磨により，列挙されるクリークの数が劇的に少なくなることがわかっている．

研磨のアルゴリズムを Algorithm 1 に示す．ここに示されたアルゴリズムは，非常に効率の悪い方法ではあるが，理解のし易さを優先させている．効率的なアルゴリズムについては文献 [8] に詳しい．研磨の方法は至ってシンプルで，全ての頂点ペアについて，その類似度がユーザの指定した閾値以上であれば接続し，そうでなければ接続しないというルールに従って，新たなグラフを再構成する．

類似度としては様々な定義を用いることができるが，今回は類似度グラフの構築で用いた NPMI とした．グラフ上での 2 つの節点 u, v の類似度 NPMI は，式 1 の定義において $\Pr(u) = |N(u)|/|V|$，$\Pr(u, v) = |N(u) \cap N(v)|/|V|$ として出現確率を定義したものに相当する．ここで $N(u)$ は節点 u に直接接続のある節点集合を，V は研磨対象のグラフを構成する節点集合を表している．すなわち大雑把に言えば，共通節点の多い節点間に枝が張られ，少ない接点間の枝は切断される．これは，SNS における友達紹介のアルゴリズム (すなわち共通友達の多い友達は友達である可能性が高い) と同様なもので，グラフ構造のプリミティブな変化予測 (リンク予測) を行っているとも解釈できる．

そして新たに構成されたグラフを入力として同様の研磨手法を繰り返し適用し，グラフの構成に変化がなくなるか，もしくはユーザの指定した最大繰り返し回数に達すれば終了する．最終的に得られたグラフが研磨グラフである．この研磨グラフから列挙された極大クリークを我々はマイクロクラスタと呼ぶ．

Algorithm 1 グラフ研磨アルゴリズム

```
function POLISHING(G = (V, E), σ)
    V : 頂点集合, E : 辺集合, σ : 類似度下限値
    E' = φ; V' = φ          ▷ 研磨後の辺集合と頂点集合の初期化
    for all u ∈ V do
        for all v ∈ V do    ▷ # 全頂点ペア u, v について調べる
            if sim(u, v) ≥ σ then    ▷ 頂点ペア u, v が似ていれば新たに辺として加え，似てなければ加えない
                E' = E' ∪ (u, v)
                V' = V' ∪ u
                V' = V' ∪ v
            end if
        end for
    end for
    return(V', E')
end function
```

グラフを研磨することの利点の一つは，グラフ構造が明確化されるために，グラフに含まれる極大クリークの数が大幅に少なくなることである．本研究で利用したデータにおいても，研磨前のグラフに比べて平均約 89.5%の削減効果が確認されている．そこで研磨グラフから列挙される少数の極大クリークをツイート内容の要約として考えることにする．各極大クリークは，ユーザが投稿するツイートに頻出する単語の共起関係に基いた類似性によって定義され，またグラフ研磨によって，本来ならば表面化しなかったような隠れた関係性を持った単語群を含んだクラスタとなっていると期待される．

2.3 構造変化の視覚化

ツイート内容の変化検出では，前節で導入した単語クラスタ (研磨された類似度グラフの極大クリーク) を構成する単語がどのように時系列で変化したかを考える．グラフ構造の差異を表す指標としては，グラフの編集距離など多くの方法があるが，本研究では図 4 に示されるような Sankey ダイアグラムによって視覚化することで変化を主観的に捉えることにする．Sankey ダイアグラムとは，閉路のない有向グラフ (DAG) を視覚化する手法の一つで，枝の重みとして定義される流量が接点間でどのような割合で流れていくかを直感的に理解することができ，送電ネットワークの視覚化などに利用される．

本研究では，ある期における研磨グラフの各クラスタを節点と考え，次の期の各クラスタと共通する単語数を流量として Sankey ダイアグラムを描画する．図 4 では，節点 (クラスタ) は棒で示されているが，その高さはクラスタに含まれる単語数に対応する．そして同じ期のクラスタは全て同じ水平位置に描画されている．このチャートから，ツイート内容について以下の 3 つの性質を読み取ることができる．

- 内容変化 (枝の錯綜): 枝の分岐が多い場合，単語の結びつきに変化が生じたということを意味し，全体としての投稿内容に何らかの変化が生じたと考えられる．
- 多様性 (節点の高さ): ある期における全ての棒の合計が相対的に長くなるということは，それだけ多様な単語が利用さていることを意味し，意見に多様性が出てきたと考えられる．
- 独立性 (節点の多さ): ある期における接点数が多い場合，単語の結びつきが細分化されたことを意味し，ユーザによって投稿される内容が分化してきたことが伺える．

3. 実験

3.1 実験データ

本研究では,「育休」「育児休暇」のいずれかを含む 2013 年 4 月 10 日から 10 日間につぶやかれた約 26,000 ツイート（約 6,400 ユーザ）を用いた.

3.2 属性推定

制度に対する意見が性別，あるいは子供の有無によって異なるかどうかを知るため，Naïve Bayes による属性推定を行った．まずプロフィール文から「サラリーマン」「ママです」等の表現を含むユーザを選択し,「男性」「女性」および「子供あり」「子供なし」のラベルを付与した．次にツイート本文に含まれる単語を用いて各属性の特徴語を取り出し，属性未知のユーザに対する属性推定を行った．交叉検証の分類表を表 1 および表 2 に示す．性別で 83.7%，子供有無で 78.2%の正答率を得た.

表 1: 分類表 (性別)

		実際のクラス 男性	女性
予測クラス	男性	706	201
	女性	332	2044

表 2: 分類表 (子供有無)

		実際のクラス 子有	子無
予測クラス	子有	2278	462
	子無	465	1051

3.3 単語の出現構造の変化検出

2013 年 4 月 18 日の安倍首相「育休 3 年」発言を機に，Twitter 上の話題はどのように変化したか，またどのような話題が展開されたかを検出する．本研究が提案する手法と比較するために，まずは単語の出現頻度による評価，次に研磨しないグラフによるクラスタリングを行い，最後に本研究が提案する手法，すなわち研磨したグラフからのクラスタリングを行う.

なお,「育休」「育児休暇」という語を含むツイートは 4 月 18 日以降 150〜450 件/日に急増するが，4 月 17 日以前は 20〜30 件/日と少ない．そこで以下の話題抽出は，当日を含む過去 7 日間のツイートを用いて行う．たとえば 4 月 18 日の話題とは，4 月 12 日〜18 日のツイートから抽出されたものである.

3.3.1 単語の出現頻度の増減

単純な方法として，期間ごとに単語の出現頻度（全ユーザに対する，その単語をツイートしたユーザの比率）を求め，その増減で判断する方法が考えられる．4 月 17 日と翌 18 日を比較したときの，増加率・減少率が高い語の一部を表 3 に示す．どのような単語が多く出現するようになったかはわかるが，そこから特定の話題を見いだすことは難しい.

表 3: 出現頻度が増加/減少した単語（一部）

増加	少子化, 安倍晋三, 待機児童, お金, 話題, 親, 長い, 場合, 3, 半年, 幼稚園, 採用, 知れる, 神話, 疎だ, 離れる, 役所
減少	終日, 週間, 不安だ, 貴重だ, 力, 呟く, 彼女, 悩む, 我が家

3.3.2 研磨しないグラフからのクラスタリング

単語間の類似度グラフを図 2，グラフから抽出したクラスタ (極大クリーク) の一部を表 4 に示す.

クラスタは 77 個抽出され，1 つのクラスタに含まれる単語数の平均は 2.67 である．{ 保育園, 入れる, 子供 } などその話題が理解できそうなクラスタもあるが，{ 会社, 内 }，{ 年, 月 } など 2 語からなるクラスタが 44 個を占めるなど，全体を話題として理解することは難しい．また語数が多いクラスタを見ても，{ ない, 復帰, 有る, 私, 良い }，{ ない, 休む, 保育園, 復帰 }，{ ない, 保育園, 復帰, 有る, 良い } など，重複部分の多いクラスタが多数現れるという問題も散見される.

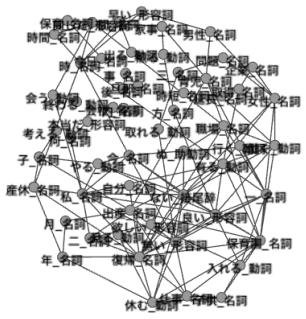

図 2: 単語の類似度グラフ

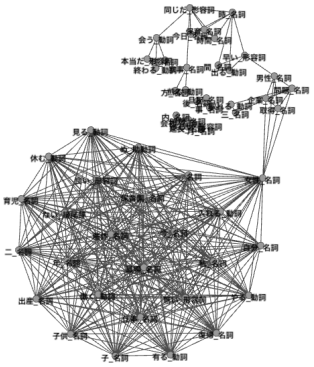

図 3: 単語の類似度グラフを研磨したグラフ

表 4: 類似度グラフから抽出したクラスタ（一部）

{ 会社, 内 }	{ 年, 月 }
{ 保育園, 入れる, 子供 }	{ 出産, 月 }
{ 仕事, 休む, 保育園, 子供 }	{ 二, 欲しい }
{ 事, 後 }	{ 方, いる }
{ 事, 旦那 }	{ 家事, いる }
{ 延長, 時短 }	{ 保育園, 働く, 職場 }
{ 働く, 延長 }	{ ぬ, 働く, 職場 }
{ 同じだ, 時, 時間 }	{ ない, ぬ, 働く }
{ 保育, 同じだ, 時間 }	{ 女性, 男性 }
{ 今日, 保育 }	:

3.3.3 研磨グラフからのクラスタリング

本研究の提案する手法を用い，研磨した類似度グラフを図 3 に，研磨グラフから抽出したマイクロクラスタを表 5 に示す.

表 5: 研磨したグラフから抽出したマイクロクラスタ

{ 会社, 内 }	{ 今日, 保育, 同じだ, 時, 時間 }
{ 事, 後, 旦那 }	{ 企業, 取得, 問題, 女性, 男性 }
{ 延長, 時短 }	{ 三, 取れる, 問題 }
{ 早い, 男性 }	{ 会う, 何, 本当だ, 終わる }
{ 会う, 同じだ }	{ 出る, 早い, 時, 間 }
{ 月, 欲しい }	{ 家事, 方, いる }
{ 同じだ, 家事 }	:

研磨しない類似度グラフからクラスタを抽出した場合に比べ，クラスタ数は 14 に減り，1 クラスタに含まれる単語数は増えている（平均 4.64 語．26 語からなるクラスタを除くと平均 3.0 語）．また似たようなクラスタが複数列挙されるという問題も回避されている.

4 月 10 日から 20 日の Sankey ダイアグラムを図 4 に示す．10 日から 11 日にかけてもクラスタ構造の変化が見て取れるが，クラスタの多様性，独立性の観点から見て 4 月 18 日が突出しているのがわかる．安倍首相の発言を受けてツイート数自体が増加したこともあるが，Twitter で展開される話題に大きな変化があったことが読み取れる.

3.4 話題の差異分析

本節では，マイクロクラスタに含まれる単語の内容から，話題の時系列差異および属性別差異を分析する.

3.4.1 安倍首相発言前後での差異

4 月 17 日には { 出産, 為, 取れる, ある, 本当だ }，{ ママ, 頑張る, 出す, . . . } など，育休を取得している，取得しようと

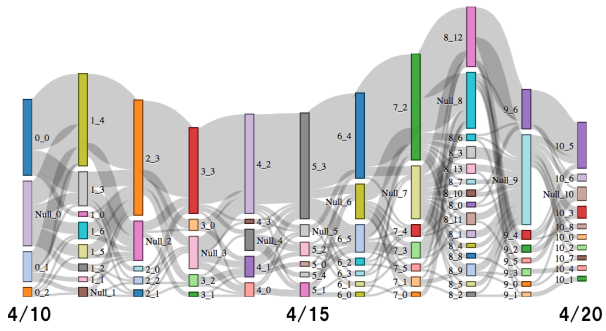

図 4: Sankey ダイアグラム．図中、“Null” へ流れる単語はどのクラスタからも消えた単語で，また “Null” から流れる単語は新たにクラスタに現れた単語を意味する．

しているユーザによると思われる話題が抽出されていた．ところが 4 月 18 日には { 企業, 取得, 問題, 男性, 女性 } や { 3, 取れる, 問題 } など，安倍首相の発言を受けたと思われる話題が現れ，4 月 19 日になると { 男性, 育児 }，{ 自分, 子, 期間, 考える } など，ツイートの増加にあわせて意見の表明や議論が進んでいることを思わせる話題が抽出されるようになった．

3.4.2 男女での差異

4 月 18 日の話題を男女で比較したところ，女性では 8 つの話題が得られ，その中には { 企業, 問題 }，{ 主婦, 大変だ, 感じる, 方, 旦那, 見る } のように意味を解釈しうるものもあった一方，男性では 50 語からなる話題が 1 つ得られたのみで，特定の話題を捉えることはできなかった．安倍首相の発言当日は，男性よりも女性の方が大きく反応したと考えられる．

4 月 19 日になると男性でも話題が得られ，男女それぞれに特徴が見て取れる．たとえば男性には { 一, 労働, 勤務, 安倍晋三, 実現, 待機児童, 支援, 日, 時間, 期間, 経済, 要請 }，女性には { 子, 復職, 期間, 産む, 考える, 自分, 長い } といった話題が現れた．これは，男性はこの育休の延長問題を政治的な話題と捉え，女性は自身もしくはその周辺の問題として捉えているといえる．以下に，男女それぞれのツイート例を示す．

- 安倍首相は経済団体に、ベアに加えて、育休期間拡大、上場企業に最低 1 名の女性の役員登用を要請するらしい。こんなものを政府に要請させる企業経営者と労組幹部は恥ずかしくないのか。（男性）
- 育休延長もいいと思うけど、仕事から遠ざかっている期間が長くなればなるほど、延長だけでなく、復職しやすい制度の方が需要あると思う。保育園はもちろん、例えば時短の延長とか、病後児保育や学童とか。（女性）

3.4.3 子供有無での差異

子供の有無によっても話題は異なってくる．たとえば 4 月 18 日では，子供のいるユーザからは { 会社, 保険, 児, 内, 出す, 制度, 取れる, いる, 延長, 後, 頑張る }，子供のいないユーザからは { ない, 三, 休暇, 先, 日本, 給与, 育てる, 育児, 違う, 雇用, 難しい } といった話題が抽出された．

子供のいるユーザには「保険」という単語が含まれているが，これは育休期間が延長されたとしても，その間の社会保険料の負担を心配する声である．すでに子を持ち，育休を取得した経験のあるユーザからの声と考えられる．

一方子供のいないユーザからは,「雇用」「難しい」という単語が含まる話題が抽出された．育休が事実上正規雇用者しか利用できない制度であることに疑問を感じている，もしくは子を持ちたいが自身の置かれている労働環境ではそれが難しいという声であろう．

以下に，子供あり，子供なしユーザそれぞれのツイート例を示す．

- 育休 3 年を制定して、幼保一体に入園するようになったら、保育園はいらなくなる？育休手当ては出ないけど、保険や年金は支払わなきゃいけない会社の場合、3 年間払い続けるのしんどくない？（子供あり）
- 三歳まで育休をって…まずどんな職場でも雇用形態を問わず育休をとれるように、が先じゃ？というか、日本全国の企業のせめて半分ぐらいは、労基法完全施行するところからか？（子供なし）

4. おわりに

本研究では，Twitter に投稿されたツイートの話題変化検出，および話題要約の手法として，研磨した類似度グラフからのクラスタリング手法（極大クリーク列挙）を提案した．育休に関するツイートに適用することで，いくつかの興味深いを知見を得ることができ，本手法の有効性を示すことができた．

ただし，本研究では手法の有効性について厳密な定量的検証を行っていない．これはクラスタリングという教師なし学習が本質的に持つ困難性に由来していると考える．今後，要約の有効性についての人間の目による比較実験を積み重ねるなど，手法の実務的かつ意味的有効性を検証していきたい．

5. 謝辞

本研究の一部は，科学技術振興機構 CREST，および ERATO 湊離散構造処理系プロジェクトの補助を受けている．

参考文献

[1] Filatova, E., V. Hatzivassiloglou, “A formal model for information selection in multi-sentence text extraction”, *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp.397–403, 2004.

[2] Fung, G., J. Yu, P. Yu and H. Lu, “Parameter free bursty events detection in text streams”, *Proceedings of the 31st international conference on Very large data bases*, No.12, pp.181-192, 2005.

[3] 今田幸子「女性の就業継続の現状と課題」,『ビジネス・レーバー・トレンド』pp.2-4, 2009.

[4] 中原孝信, 前川浩基, 羽室行信「テレビ番組視聴時における Twitter 投稿からのトピック検知」, オペレーションズ・リサーチ, 58(8), pp. 442-448, 2013.

[5] 乙部由子, 乙部ひさよ「育児休業を取得する女性総合職増加に伴う課題：女性総合職の新卒採用抑制と活躍の場の減少」,『家計経済研究』(99), pp.74-81, 2013.

[6] 高橋佑介, 横本大輔, 宇津呂武仁, 吉岡真治 「ニュースにおけるトピックのバースト特性の分析」, 情報処理学会研究報告. 自然言語処理研究会報告, 一般社団法人情報処理学会, No.6, pp.1–6, 2011.

[7] 高村大也, 奥村学「最大被覆問題とその変種による文書要約モデル」人工知能学会誌 23(6), 505-513, 2008.

[8] 宇野毅明, 中原孝信, 前川浩基, 羽室行信「データ研磨によるクリーク列挙クラスタリング」情報処理学会アルゴリズム研究会報告書, 2014-AL-146(2), pp. 1-8, 2014.