日本ソーシャルデータサイエンス論文誌
第 1 巻　第 1 号（2017 年 3 月）

特集：データサイエンスの現状と将来

ビッグデータからの情報抽出とその応用

中原 孝信[†1]　羽室 行信[†2]

概要：インターネットの普及とともに，SNS やセンサー技術など新しいサービスや技術が一般的になり，身近に利用できるようになったことで，これまでに比べて膨大な量のデータが蓄積されている．近年ではそのようなビッグデータをビジネスに活用するための取り組みが注目されており，ビッグデータを分析し得られた情報から，経営戦略の策定，市場の調査・分析、商品・サービスの品質改善，そして，業務の効率化など，さまざまな業務への適用が試みられている．ビッグデータを用いることで，これまでには知ることのできなかった現象の把握や予測精度の向上などが期待されているが，ビッグデータには多くのノイズが含まれており，有用な情報抽出のためにはノイズうまく扱う必要がある．本稿では，出現頻度を利用して，ビッグデータからノイズとなるような分析の信頼性を損なう関係性を除去し，意味のある関係性を抽出する技術であるグラフ研磨を紹介し，ソーシャルデータである Twitter からの意見抽出を目的に 2 部グラフを利用した応用研究を示す．

キーワード：ビッグデータ，情報抽出，グラフ研磨，2 部グラフ，データクリーニング

1. はじめに

情報通信技術は，インターネットの発明以来とどまることのない発展を続けており，革新的な製品とサービスが生み出されている．Google, Facebook などの大手企業は情報通信技術の利点を活かしたビジネスを展開しており，Google は，検索サービス，地図アプリ，メールソフトなどのソフトフェアを無料で提供することで，ユーザを Google プラットフォームに集めている．Facebook はソーシャル・ネットワークによる人と人とのつながりを軸に「いいね」や「タイムライン」などユニークな機能とともに世界中から利用者を獲得している．

これらの企業は無料でサービスを提供し，自ら所有するプラットフォームに集客することで，検索履歴，閲覧履歴，ソーシャル・ネットワークなどの膨大な履歴をビッグデータとして獲得している．そして，ビッグデータからユーザのニーズや興味を把握し，ユーザにクリックされやすい広告を配信するシステムを構築することで，巨万の富を得ている．

インターネットを活用したビジネスモデルでは，製品やサービスだけが収益源ではなく，ビジネスの源泉はデータにある．膨大な数の行動ログや履歴データなどのビッグデータを解析することでユーザのニーズ，好み，行動の特徴を把握し，その情報にもとづいてユーザとコンテンツのマッチングにより収益を挙げている．

今後は、全てのものがインターネットにつながるという概念である IoT への注目から，より多くのセンサーデータが収集され，ビッグデータの利活用は一部の IT 企業だけではなく，家電メーカーや自動車メーカーなどあらゆる分野に広がっていく．更に安倍政権でも「第 4 次産業革命」として IoT，人工知能，ビッグデータなどの革新的技術の活用を成長戦略の 1 つに挙げており，ビッグデータへの期待が高まっている [15].

1.1 ビッグデータの特徴

ビッグデータの特徴は，Volume, Velocity, Variety という 3 つの V で表現されている [1]．Volume は，データ量についての特徴で，Facebook では 1 日のデータ処理量は 600TB に及んでおり，300PB のデータベースにデータが蓄積されている [9]．また Twitter では 1 日に 5 億件以上のツイートが投稿されており，1 秒間の投稿数の最高は 143,199 ツイートで，日本ではおなじみの「天空の城ラピュタ」の放送時である [13]．そして，小売店の最大手であるウォルマートでは，1 時間に 100 万人以上で 2.5PB 以上の取引データが処理されている [14]．このように日々蓄積されているデータ量はますます膨大になっている．

次に Velocity は更新速度を表している．これはデータ量だけではなくその更新頻度がビッグデータを特徴づけており，サーバーのアクセスログやセンサーデータなど時々刻々とリアルタイムに更新されるデータを処理するための必要性が高まっている．

3 つ目の Variety は，データの多様性を表しており，Blog や Email などに記載されるテキスト情報から，SNS に投稿される写真などの画像データ，RFID タグなどのセンサーデータ，スマートフォンから更新される GPS などの位置情報データがあり，これらのデータの大部分は，過去 10 年ほどの間に生まれてきた新しいデータである．

ビッグデータを用いることで，これまでに知ることのできなかった現象の把握や予測精度の向上などが期待されて

†1 専修大学（連絡先：nakapara@isc.senshu-u.ac.jp）
†2 関西学院大学

おり，ビッグデータから未来の行動に有用な情報と知識の抽出が求められている [8]. しかしながら，ビッグデータにはノイズが多く含まれている．それは，観測データの欠陥だけではなく，ソーシャルデータの内容に信頼性や信憑性がないものも含まれている．したがって有用な情報抽出のためにはノイズを見つけて除外することが重要である．また，センサーデータなどの生データは，それだけでは役に立たないものが多く，例えば GPS の位置情報は他のランドマークなどのデータと統合することで意味を持つため，オープンデータなど外部データの整備も合わせて必要になる．

ビッグデータから有用な情報を抽出するためには，1) ビッグデータにアクセスし処理できる基盤技術，2) マイニングアルゴリズム，そして 3) ビッグデータから得られる情報をドメイン知識にもとづき意味解釈することが重要である．

本稿では上記 3 つの観点を考慮したビッグデータからの応用として，1) 大規模 CSV データ分析プラットフォームである NYSOL1を利用する．2) 出現頻度にもとづく関係性を考慮して意味のあるグラフ構造を抽出する技術であるグラフ研磨を紹介する．そして応用研究として 3) ソーシャルデータである Twitter からの意見抽出とその解釈を示す．次節以降の応用研究は，2016 年人工知能学会全国大会の報告内容 [6] をもとに加筆・修正を行った．

1.2 育児休業を対象とした Twitter からの意見抽出

1992 年に育児休業法が施行されてから育児休業制度の導入は広がっており，2014 年には従業員数が 30 人以上の事業所では 94.7%で育児休業制度が規定されている．そして，女性の育児休業取得率は 86.6%になっている [11]. しかしながら，第 1 子の出産を機に有職女性の 54.1%が退職しており [10]，出産・育児を経た就業継続はいまだに困難である．

少子高齢化により日本の労働力人口は減少しており，労働力人口を確保するためには，現在働いていない人に働いてもらうか，働いている人の離職率を下げる必要がある．特に第一子出産による女性の離職率は高く，その時期の離職率を下げ就業継続を高めることが労働力人口を維持するためにも重要となる．このような中，安倍政権は待機児童数ゼロの実現を掲げたり，女性活躍推進法を新たに制定したり，成長戦略の中核に女性の活用を据えている．女性にとって働きやすい環境を提供し女性の就業継続率を上昇させることは，重要な課題の 1 つである．

本研究では，育児休業（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方法を紹介する．そして，育休に対する率直な意見や，育児と仕事の両立のために必要な政策などの情報を得ることを試みる．2016 年 2 月 15 日に投稿された匿名ブログでは，保育園の入園選考に落ちたことに対して国に不満をぶつけた内容が，子育てをしている母親らの共感を集め，待機児童問題に関して国の政策を動かす程の大きな反響を得ている [12]. このように SNS やブログでは日々膨大な投稿が行われており，その中に埋もれている重要な意見や，多数の意見を要約して提示することには意義がある．

これまでにも著者らは，安倍首相の育休 3 年の要請という発言 (2013 年 4 月 18 日) によって，ツイートの話題がどのように変化したかを捉える方法を提案した [5].

そこでは，単語間の関係性を表す類似度グラフを構築し，そこから密部分グラフを単語クラスタとして抽出することで，文章要約を実施した．

本研究では，単語間の関係性を一般グラフではなく，格フレームを用いた 2 部グラフで表現し，2 部グラフの研磨手法を適用する．そして，研磨後の 2 部グラフから要素の重複が少ない極大 2 部クリークを抽出し，それらをトピックとして利用する．最終的にそのトピックを含むツイートをクラスタリングすることで文書の要約を行う．

2. 手法

本研究では，図 1 に示す方法でツイートの文章要約を実施する．まず，(1) ツイートを構文解析し，格助詞句と用言句のペアからなる格フレームを抽出する．そして，格フレームを 2 部グラフで表現する．次に (2) 2 部グラフにデータ研磨手法を適用する．データ研磨はグラフのクリーニング方法の 1 つであり，グラフから極大クリークを列挙する際に，同じようなクリークが多数列挙されるという重複問題を解決するためにグラフのクリーニングを実施する．(3) データ研磨後の 2 部グラフから極大 2 部クリークを列挙し，得られた 2 部クリークをトピックとして利用する．そして，(4) そのトピックを含むツイートをクラスタリングすることで要約を行う．最後に比較手法から得られた要約と比較するために，(5) アンケート調査を実施し，提案手法の性能を評価する．

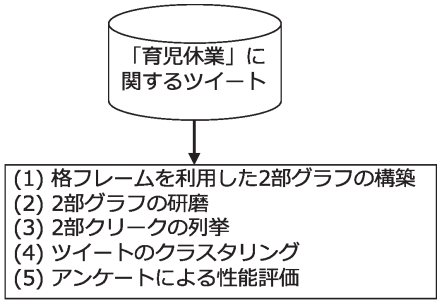

図 1　分析の概略図

1 http://www.nysol.jp

2.1 データクリーニング

SNS などのソーシャルデータには，多くのノイズとなるデータが含まれており，Twitter データでもそれは同様である．まず分析上意味のない用語やツイートを取り除く必要があり，それは分析内容に依存する．

一般的に自然言語処理では，単語（形態素）を対象に処理を行うが，その際にはストップワードを除く必要がある．ストップワードは，あまりにも一般的な語で分析精度の向上のためには除外せざるを得ない語である．日本語では，助詞や助動詞などの「は」「が」「です」や「それ」などの指示代名詞である．また，除外すべきノイズとなるツイートも存在しており，本研究では「育休」「育児休暇」を検索語としてツイートを取得しているが，育休は「体育休み」や「保育休み」などの語にも一致するため「体育休」や「保育休」に一致したツイートは除外する必要がある．

また Twitter には Bot と呼ばれる自動投稿プログラムによる投稿も多く含まれており，それらの投稿には意味がないため Bot からの投稿は除外する必要がある．Bot による投稿は Twitter のスクリーン名に「bot」と含まれている場合が多く，そのようなスクリーン名を持つアカウントは除外する．リツイートは投稿の持つ影響力を評価する上では重要であるが，意見の要約では同じ文章は必要ないためリツイートを利用する必要はない．

2.2 格フレームを用いた 2 部グラフの構築

これまで文章を表現するために最も利用されてきた方法の 1 つは bag-of-words (BOW) であり，単語の出現だけを考慮したベクトルで文章を表現する方法である．BOW による表現は非常にシンプルで，ときに有用な結果をもたらすが，単語の出現順序や文章の構造を無視しているため，文章の意味を表現する場合にはその点が問題となる．一方で，格フレームは，ガ格やヲ格などの格助詞句と，動詞や形容詞などの用言句のペアによる表現で，「育休を，取得する」「保育園が，一杯だ」など，格フレームによって文章の意味を表すことができる，

本研究では，ツイートから格フレームを抽出するために，日本語の自然言語処理ソフトである KNP [4] を利用し，格解析を実施することで格フレームを抽出する．そして，得られた格フレームを 2 部グラフで表現する．2 部グラフとは，グラフ$G = (V \cup U, E)$の任意の頂点集合VとUが枝で接続されたグラフである．抽出した格フレームを構成する格助詞句と用言句をそれぞれV, Uとして頂点を枝で結ぶことで 2 部グラフを生成する．

2.3 2 部グラフの研磨

これまで著者らは一般グラフを対象にしたデータ研磨を提案してきた [3]. データ研磨のアイデアは，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ことで，本質的な構造を失うことなくグラフを明確化するものである．このことにより列挙されるクリーク数を削減する効果が得られる．本研究では，データ研磨を 2 部グラフに対して適用することで，2 部グラフを明確化し，重複の少ない極大 2 部クリークを列挙する．

2 部グラフの頂点集合を$V = \{v_1, v_2, \dots, v_m\}, U = \{u_1, u_2, \dots, u_n\}$とし，$X(v)$を$v$に隣接する$U$の頂点集合とする．また，$X(u)$を$u$に隣接する$V$の頂点集合とする．2 部グラフの研磨アルゴリズムを Algorithm 1 に示す. ここで示すアルゴリズムは，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．

Algorithm 1　2 部グラフ研磨アルゴリズム

```
1: function BIPOLISHING(G = (V ∪ U, E), σ1, σ2)
2:     V, U: 頂点集合，E: 辺集合，σ1, σ2: 類似度下限値
3:     V', U', E' = ∅        ▷頂点集合，辺集合の初期化
4:     for all v ∈ V do
5:         S = ϕ
6:         for all v' ∈ V do
7:             if sim(X(v), X(v')) ≥ σ1 then
                                ▷ 接続関係の類似する頂点を保存
8:                 S = S ∪ {v'}
9:             end if
10:        end for
11:        for all u ∈ U do
12:            if sim(X(u), S) ≥ σ2 then ▷ 接続関係が似て
                いれば枝を張り，似ていなければ張らない
13:                E' = E' ∪ {(u, v)}
14:                V' = V' ∪ {v}
15:                U' = U' ∪ {u}
16:            end if
17:        end for
18:    end for
19:    return G = (V' ∪ U', E')
20: end function
```

2 部グラフの研磨は，部間の接続関係の類似性に着目したグラフ研磨手法である．まず，Uへの接続関係が頂点vと類似した頂点集合S（v自身も含む）を見つけ出す（6〜10 行目）．Vを格助詞句集合，Uを用言句集合とすると，用言句との結びつき（共起関係）が格助詞vと類似した格助詞句部分集合がSとなる．今度は逆に$u \in U$からVへの接続を調べる．uの接続先の頂点集合 $(X(u) \subseteq V)$ と$S \subseteq V$との類似性を判断し（12 行目），類似していれば，節点v, uを接続し，類似していなければ接続しない．このことにより，お互い

の共起関係において類似した格助詞句vと格助詞句uに枝を張り直すことが可能となる．

以上の操作により，オリジナルの 2 部グラフに枝(v,u)がなくても，お互いの共起関係において類似していれば新規に枝が追加されることになり，逆に，枝(v,u)が存在していても，類似していなければその枝は削除されることになる．このようにグラフ研磨は，部間の接続関係（お互いの共起関係）によってオリジナルの 2 部グラフのグラフ構造を変更するため，一方的な共起関係が省かれ，またサンプリング上共起が少なくなっているような関係性を修復するなど，部間の関係性の明確化およびノイズのクリーニングとしての効果が期待できる．

類似度（7, 12 行目）はさまざまな定義を用いることができるが，本計算では jaccard 係数を利用する．2 つの頂点集合YとZの jaccard 係数による類似度は，式 (1) の通り定義される．

$$\mathrm{sim}(X,Y)=\frac{|X\cap Y|}{|X\cup Y|} \tag{1}$$

上記のアルゴリズムを利用し，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．そして，最終的に得られた 2 部グラフが研磨後の 2 部グラフである．

2.4 極大 2 部クリークの列挙とツイートの要約

格フレームを表した 2 部グラフから密な部分グラフを抽出することで，似た意味を表すクラスタが抽出できていると考え，それをツイート内容の要約に利用する．つまり，研磨後の 2 部グラフから極大 2 部クリークを列挙し，それをトピックとしてツイートの要約を行う．2 部グラフの頂点部分集合$K\subseteq V, H\subseteq U$に対して，$K$の任意の頂点と$H$の任意の頂点の間に枝があるとき，$K$と$H$を合わせた頂点集合を 2 部クリークとよぶ．そして，ある 2 部クリークが他の 2 部クリークに含まれないとき，その 2 部クリークを極大 2 部クリークとよぶ．

データ研磨の特徴の 1 つは，グラフに含まれるノイズを除去し，グラフ構造が明確化されることで，列挙されるクリーク数を大幅に削減できることである．本研究で利用したデータに対しても研磨をおこなわなかった場合には，264,733 の極大 2 部クリークが列挙されるが，研磨後の 2 部グラフから列挙される極大 2 部クリーク数は 1,611 で，約 99%の削減ができている．

文章の要約は，ツイートが持つトピック（極大 2 部クリーク）を用いて内容の類似するツイートをクラスタリングして，意味内容が類似したツイートをまとめることで行う．ツイートaに出現するトピック集合をD_a，ツイートbに出現するトピック集合をD_bとすると，2 つのツイートの類似度は以下の式 (2) で計算する。これは式 (1) と同様に jaccard 係数である．

$$\mathrm{jc}(a,b)=\frac{|D_a\cap D_b|}{|D_a\cup D_b|} \tag{2}$$

$\mathrm{jc}(a,b)$がある閾値σ_3以上の場合にツイート間に枝を張り，類似度グラフを生成する．そして，そのグラフに対して Newman クラスタリング [7] を行うことでツイートのクラスタを生成する．

2.5 手法の評価

提案手法では，ツイートをクラスタリングするための素性を 2 部グラフのデータ研磨と極大 2 部クリークによって生成することを示した. 素性とは文章を特徴づける属性で，例えば，単語，文字，概念などを表す．ここでは，提案手法の有効性を評価するために異なる 3 つの方法で素性の生成を行う．1 つ目は 2 部グラフの研磨を行わずに極大 2 部クリークを列挙し，それを素性とする方法である．2 つ目は BOW を素性として利用した方法，3 つ目はトピックモデルである Latent Dirichlet Allocation (LDA)[2] を利用した方法である．

評価の方法は，代表ツイートを一様ランダムに 1 つ選択し，手法毎にそのツイートを含む同一のクラスタから別のツイートをランダムに選択する．そしてアンケート調査を実施し，各手法から選ばれたツイートと代表ツイートを比較してもらい，代表ツイートに最も近いツイートを選んだ手法を 1 位として, 4 位までの順位をつけてもらう. なお，複数の手法で甲乙つけがたい場合は同一順位を与えてもらうことにした．

3. 手法の適用

本研究では，2012 年 10 月から 2015 年 1 月 1 日の期間で，「育休」「育児休暇」のどちらかを含むツイートデータ約 28 万件（13 万ユーザー）を利用し，クリーニング後のデータは約 20 万ツイートを対象とした．

3.1 2 部グラフ研磨の結果

最初に格フレームを抽出し，2 ツイート以上に出現する格フレーム 37,003 種類を分析対象として 2 部グラフを生成した．Vは 13,891 種類の格助詞句で，Uは 4,628 種類の用言句であった．この 2 部グラフに対して 2 部グラフの研磨を

2 計算は R の LDA パッケージを利用した．

適用した結果を表 1 に示す．

表 1 はσ_1とσ_2の値をそれぞれ 0.1 ずつ変化させた場合に得られた極大 2 部クリーク数を示している．全体的な傾向は，各閾値を大きくすると得られる極大 2 部クリーク数は少なくなっている．これは，σ_1が大きくなると，用言句への接続関係が強く類似した格助詞句だけが選択されるため，選択される格助詞句が少なくなるためである．そしてσ_2が大きくなると，選択された格助詞句の多くが共通する用言句への関係を持っていなければ枝が削除されるため，疎な 2 部グラフになる．したがって，列挙される極大 2 部クリーク数も少なくなる．

表 1　研磨の閾値と極大 2 部クリーク数の関係

σ_1 \ σ_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	2,139	2,332	2,021	1,755	1,634	1,329	1,130	1,028	945
0.2	2,779	2,501	2,006	1,710	1,578	1,229	1,052	991	938
0.3	3,891	3,027	2,409	1,987	1,870	1,376	1,229	1,193	1,144
0.4	4,261	3,261	2,665	2,215	2,112	1,519	1,405	1,373	1,339
0.5	4,274	3,299	2,737	2,308	2,210	1,604	1,492	1,460	1,426
0.6	4,329	3,409	2,884	2,445	2,374	1,699	1,608	1,590	1,573
0.7	4,326	3,420	2,900	2,465	2,396	1,730	1,642	1,624	1,607
0.8	4,324	3,420	2,902	2,468	2,399	1,734	1,645	1,627	1,610
0.9	4,321	3,421	2,903	2,468	2,399	1,734	1,646	1,628	1,611

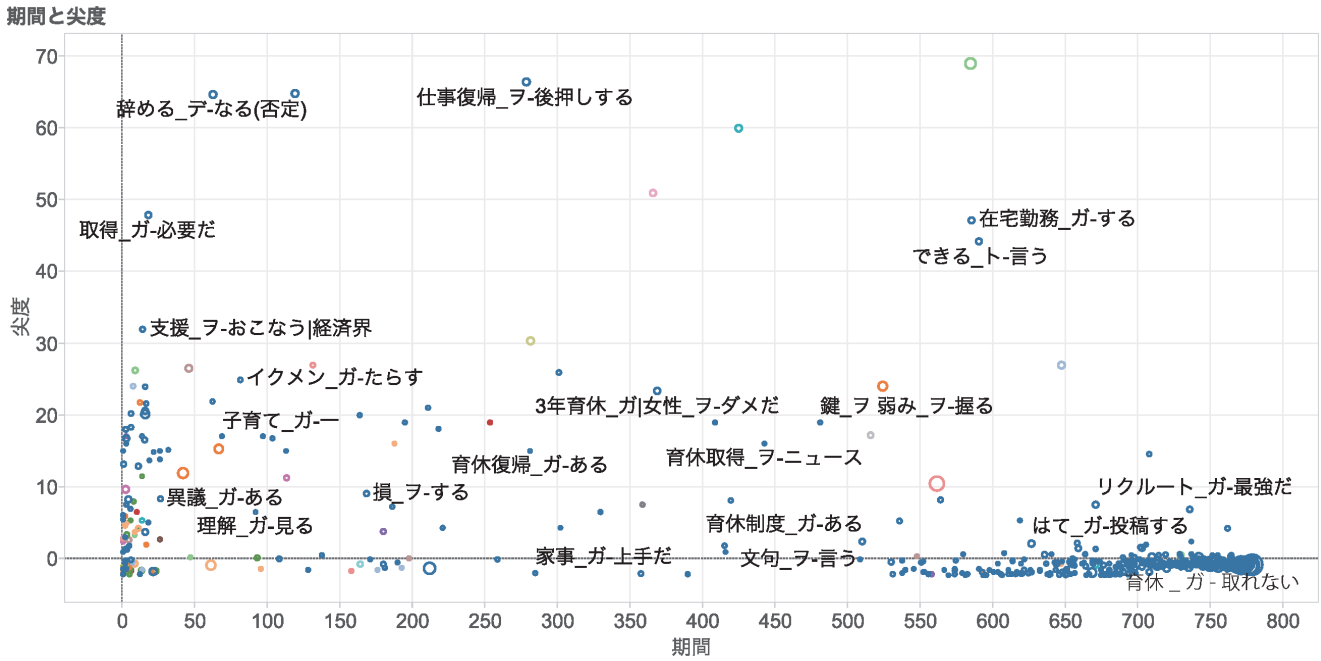

図 2　ツイート要約の結果

3.2　ツイートの要約

本研究では研磨のパラメータとして，最も高い閾値である 0.9 を設定し，得られた 1,611 の極大 2 部クリークを利用してツイートの要約を行った．得られた極大 2 部クリークは，{育休 3 年ガ-る，育休 3 年ガ-批判噴出，育休 3 年ガ-活気，育休 3 年ガ-育児子育て支援} や {夫ガ-育児しない，夫ガ-育児休業取得する，夫ガ-薬局長} のように比較的意味の取りやすいものが多かった．

これらの極大 2 部クリークを持つツイートをクラスタリングした結果が図 2 である．ツイートの類似度グラフを作成する際に利用した閾値σ_3は 0.6 とした．また Newman クラスタリングを利用することでクラスタ数を指定する必要はなく，Modularity Q が最小になるようにクラスタ数が決定される．結果として約 20 万のツイートから約 12,000 のクラスタが構成された．

図の点は 1 つのクラスタを示しており，点の大きさは，クラスタに含まれるツイート数に対応している．図には 15 ツイート以上を含むクラスタのみを示している．また，図の軸は各クラスタに含まれるツイートの投稿日から計算した尖度と期間を表しており，縦軸の尖度が高いと短期間で投稿数が多くなっていることを示している．また横軸の期間は，各クラスタで最初に投稿された日から最後に投稿された日までの期間を表しており，同一の話題がどの程度の期間で展開されたかを示している．

たとえば，尖度が 66，期間が 280 の「仕事復帰 ヲ-後押しする」というクラスタは全体で 67 件のツイートからな

るクラスタである．尖度は他のクラスタに比べて大きい値であり，ある程度まとまった期間に投稿されていることを表している．実際には 2013 年 5 月 21 日，22 日に 60 件のツイートが投稿された．また期間の 280 日は，同じ話題が 280 日に渡って投稿されているが，尖度が高いため多くのツイートは 5 月 21 日，22 日の 2 日間に集中しており，それ以外の期間は少ない投稿が分散していることを表している．

このクラスタは「育休 3 年が仕事復帰を後押しするか？」という内容の記事に対しての意見が投稿されたもので,「問題はこの制度の対象にならない非正規雇用が多い点だ」「中小企業が 99.7%の日本では、まず 3 年も待てない」「浦島太郎になりそう」「二人目できたら 6 年休むのか？」「女性にプラスではなく，子供に何がプラスかをまずは考えるべきだ」など否定的な意見が圧倒的に多く,「私は年齢的にぜひ賛成！」という少数の賛成意見もあった．

また，その横の「辞める_デ-なる（否定）」は,「辞めなくなっただけで」という文節に対応したクラスタで,「女性社員，辞めなくなっただけで戦力になっていない．育休，時短の増加で企業疲弊」という記事に対してのツイートが投稿されており，65 件のツイートからなるクラスタで 2013 年 3 月 14 日，15 日に 57 件の投稿が行われていた．「男性社員にも戦力にならないやつ大量にいるのに」「男女関係なしに，稼げば雇う，稼がないなら解雇でよくないか？」「うちや周りの女性社員は優秀だが」「総合職キャリア組のための施策を腰掛け OL ばかりが使っているから」等の様々な意見が投稿されている．

尖度が低く，期間の長いクラスタとしては,「育休 ガ-取れない」（期間 736，尖度 0）のクラスタで，育休が取れないことに対しての様々な意見を投稿しており,「補填される額では全然足りないから，簡単に育休って取れないんだよね」「出産するからっていって育休取れない会社なら，いずれにせよ出産の前に辞めるという決断を多くの人がするのでは？」「育休は正直怖くて取れない」「養子だと育休が取れないとか，そんな慣習がこの国に在ったとは」「こういう職場にいるから普段は感じないけど、仕事で不利益を被る女性は多いんだろうなぁ 気軽に育休取れないとか考えられない」「中小企業や自営業だと育休は取れないよなー」など多岐にわたる意見が投稿されている．

各クラスタはある程度共通のトピックを持ったツイートでまとめられており，トピックがインデックスの役割をすることで,興味のある話題を選択することが可能である．

そして，詳細な内容はそのクラスタの各ツイートを確認することで，有益な情報が得られる。上述のクラスタの内容からも，制度はあっても実際には育休を取得することのリスクを恐れていることが確認できる．国としては育休制度を浸透させるだけではなく，育休を取得したことによるキャリア形成への影響なども考慮した制度の整備が必要になってくる．

3.3 手法の比較

決定木，SVM，回帰モデルなどの教師あり学習では，データから正しく分類が行われたかを確認することはできるが，クラスタリングなどの教師なし学習では，性能を評価することは難しい．クラスタリングの性能を評価するためには，前節で示したようにクラスタの中身を確認し意味解釈が可能かどうかで判断することはできるが，客観的な評価を与えることは困難である．

そこで本研究では，複数の方法で得られたクラスタを評価するために，アンケート調査を実施した．アンケート調査以外の方法としては，近年クラウドソーシングが身近に利用できるようになってきており，不特定多数の評価者を低コストで集めることが可能になってきている．クラウドソーシングを利用し，人間と計算機による協調から問題解決に繋げる研究も行われ始めている [2].

アンケートによる評価は,2.5 節に示す方法で実施し比較を行った．一人の被験者には，10 ツイートを代表ツイートとして選択し，合計で 10 人の被験者にアンケートを実施した. 提案手法, 研磨なしの極大 2 部クリーク, BOW, LDA の 4 種類の方法から選ばれたツイートと代表ツイートを比較して，内容の近い順に 1 位から 4 位までの順位をつけてもらった．LDA のパラメータは，$\alpha = 0.1$, $\beta = 0.1$でパラメータの更新は Collapsed Gibbs sampling を利用した．

表 2 は，全アンケートの中から 1 つの代表ツイートだけを抜き出したものである．比較 1 は提案手法，2 は BOW，3 は研磨なしの極大 2 部クリーク，4 は LDA である．実際のアンケートでは提示順序はランダムにし，手法も特定できないようにした．順位を見ると比較 1 と 3 は同じ内容のツイートで代表ツイートに最も近いと判断されているため 2 つに 1 位が与えられている．

表 2　アンケートの例

方法	ツイート	順位
代表ツイート	そういや、育休だった方が戻ってくるな。2年ぶりかな？	
比較1（提案手法）	目安ついたのか～！育休の方が戻ってくるのかな？	1位
比較2（BOW）	最近ハマってるチョコの名前を検索したらブログ発見＊ この製品計画に携わっていた女性社員の方は1年半の育 休の後またこの製品の担当に戻ってきたみたい！	3位
比較3（極大2部クリーク）	目安ついたのか～！育休の方が戻ってくるのかな？	1位
比較4（LDA）	『「パタニティ（父性）・ハラスメント」』／「育休を取得 したくてもできなかった」45.5%男性の育児参加を阻む 「パタハラ」と上司の無理解｜ザ・世論～日本人の気持ち ～｜ダイヤモンド・オンライン	4位

表 3 は全代表ツイートに関する結果をまとめたものである．各手法は合計で 100 回評価されており，スコアは，各順位とその頻度の合計を 100 で割った値で，全て 1 位の場合には 1 になる．最も 1 に近い値は提案手法で，続いて研

磨なしの極大 2 部クリーク，BOW，LDA の順であった．この結果から提案手法は他の手法よりも意味の類似したツイートでクラスタが構成されていることを示している．提案手法と研磨なしの方法はどちらもスコアが 1 に近く僅差であり，
研磨なしも比較的意味の類似したツイートでクラスタが構成されている．ただし既に述べたように，素性の数は大きく異なっており，提案手法は 1,611 種類の極大 2 部クリークで，研磨なしは 264,733 種類である．類似したクリークをクラスタリングするためには，素性が多ければ良いわけではなく，2 部グラフの研磨によって，ノイズが除去され重要な格フレームが浮き上がったことによって，有用な結果が得られたと考えられる．

一方で，BOW は格フレームではなく単語（形態素）の出現のみを扱っており，3,676 種類の形態素を利用している．BOW の場合には形態素のある 1 語が共通することによって，クラスタを構成する場合もあり，素性が細かすぎることで文章の類似性が格フレームに比べて劣っている．LDA は，トピック数を多くしすぎるとトピックの解釈が困難なため，合計で 500 のトピックを生成するようにパラメータを調整したが，素性としては粒度が大きく，意味の類似していないツイートがクラスタリングされてしまった．

これらの結果から，素性の粒度と質が類似するツイートをクラスタリングするためには重要であり，2 部グラフの研磨によって適切な粒度でかつ重要な格フレームが抽出できていると考えられる．

表 3　手法による結果比較

手法	1位	2位	3位	4位	スコア
提案手法	86	7	5	2	1.23
研磨なし	79	10	7	4	1.36
BOW	77	10	8	5	1.41
LDA	52	4	7	37	2.29

4. おわりに

本稿では，ビッグデータからの情報抽出技術として 2 部グラフを対象としたデータ研磨手法を示した．応用研究では，格フレームを利用した 2 部グラフに，データ研磨を適用することで構造を明確化し，極大 2 部クリークの列挙数が大幅に減少することを示した．また，極大 2 部クリークをトピックとして利用することで，類似するツイートのクラスタリングと有用な情報を抽出できることを示した．育休に関するツイートの要約からは，育休 3 年と仕事復帰に関しては，否定的な意見が多く，育休 3 年という政策は論点がずれていることなど，国民の率直な意見を捉えることができた．

応用研究で示した分析内容は，大規模 CSV データの分析プラットフォームである NYSOL を利用している．NYSOL はデータのハンドリングに優れており，1 億件以上のデータが PC で処理可能である．またデータマイニングに必要なコマンド群も充実しており，京都大学の黒橋研究室で開発された自然言語処理で利用される形態素解析プログラム (JUMAN) や格解析プログラム (KNP) などが NYSOL プラットフォームで利用できる。また宇野 CREST プロジェクト3で開発された 2 部グラフのグラフ研磨アルゴリズムも利用可能である．

今後はビッグデータへの解析ニーズは更に高まることが考えられるが，データを分析することは手段であり，分析から得られた結果をどのように意思決定に役立てるかという点が最も重要である．意思決定に役立つ情報を抽出するためには，試行錯誤をしながらデータ分析を繰り返し行うことが必要不可欠であり，直感的に柔軟な方法で大規模なデータを扱える方法が求められている．

謝辞　本研究の一部は，これまで取り組んできた宇野 CREST プロジェクト，湊 ERATO プロジェクトの研究成果であり，また JSPS 科研費 JP15K17146 の助成を受けたものです．

参考文献

[1] Andrew, M. and Erik, B., “Big Data: The Management Revolution,” *Harvard Business Review* Vol. 90, No. 10 pp.60-68 (2012). (bigData)
[2] 鹿島久嗣，小山聡，馬場雪乃,「ヒューマンコンピュテーションとクラウドソーシング」，講談社 (2016). (Crowdsourcing)
[3] 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). (Uno2014)
[4] 黒橋禎夫，河原大輔，http://nlp.ist.i.kyoto-u.ac.jp/?KNP (KU00)
[5] 前川浩基，内田将史，大内章子，宇野毅明，羽室行信，“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”，人工知能学会全国大会論文集，Vol. 28, 3M-42 (2014). (maegawa2014)
[6] 中原孝信，大内章子，宇野毅明，羽室行信，“データ研磨の 2 部グラフへの適用と Twitter からの意見抽出”，2016 年度人工知能学会（第 30 回），4I1-3 (2016). (ai2016)
[7] Newman, M.E.J., “Fast Algorithm for Detecting Community Structure in Networks,” *Physical Review E*, Vol. 69, 066133 (2004). (Newman)
[8] Wu, X., Zhu, X., Wu, G-Q., and Ding, W., “Data Mining with Big Data,” *IEEE Trans. on Knowl. and Data Eng.* Vol. 26, No. 1, pp. 97-107 (2014).
[9] Facebook, https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/)
[10]厚生労働省,「第 1 回 21 世紀出生児縦断調査（平成 22 年出生児）の結果」,(2011) http://www.mhlw.go.jp/toukei/saikin/hw/shusshoujib/01/
[11]厚生労働省，「平成 26 年度雇用均等基本調査」，(2015) .
[12]日本経済新聞,

3 https://www.jst.go.jp/kisoken/crest/project/45/14531617.html

http://www.nikkei.com/article/DGXZZO76056900T20C14A800000
94/
[13] Twitter,
https://blog.twitter.com/2013/new-tweets-per-second-record-and-
how
[14] Walmart, http://www.economist.com/node/15557443
[15] http://www.kantei.go.jp/jp/headline/seicho_senryaku2013.html