The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

事象情報グラフによる複数文書中の時系列情報可視化

Graph-based Visualization of Timeseries Information in Multiple Texts

河田 裕成[*1] Hironari Kawada
赤石 美奈[*2] Mina Akaishi
細部 博史[*2] Hiroshi Hosobe

[*1] 法政大学大学院情報科学研究科 Graduate School of Computer and Information Sciences, Hosei University
[*2] 法政大学情報科学部 Faculty of Computer and Information Sciences, Hosei University

Newspapers and other media often describe particular topics over some periods by giving a series of texts that introduces new information and corrects previous information. By analyzing such texts, we can understand how a topic changed over time and how the media treated it. However, analyzing such texts manually is time-consuming. To solve this problem, we propose a method for visualizing texts that describe a topic changing over time. For this purpose, we first generate an event information graph from such texts by using an exiting method. Then we transform it into two different graphs. The first graph shows changes in the topic that we obtain by separating the initial graph and comparing the resulting time-dependent graphs. The second graph depicts a summary of the topic that we obtain by extracting high-frequency words from the initial graph. The results of our experiment show that our method can visualize changes and important information described in the texts.

1. はじめに

新聞等のメディアが報道する時事情報は，時間と共に刻々と変化する．新たな情報が次々と報じられる中で，時には過去の情報が間違いだったと判明することもある．こうした時系列的な情報変化に関する文書群を分析することで，時事情報に関する状況の推移と，メディアにおけるその扱いを知ることができる．しかし，長期間に渡って報じられた時事情報について，文書群中の情報変化を人手により分析する作業には手間と時間がかかる．

そこで，本論文では，文書中の情報を表現する有向グラフに対してグラフ操作を行うことによる，情報変化の可視化手法を提案する．具体的には，先行研究の田代らの手法[田代 2013]を用いて文書に含まれる実世界の出来事の情報をグラフ化した後，それらの時系列的な比較と重要部分の抽出を行い，それぞれを別のグラフとして可視化する．これにより，時系列情報に関する情報変化や，時系列全体を通して出現する重要な情報の分析支援を試みる．

2. 関連研究

情報の時系列的変化の可視化手法としては，伊藤らの研究[伊藤 2013]がある．伊藤らは，係り受け解析により動詞句と名詞句のペアをイベントとして抽出した．また，イベント群の時系列的な構成や頻度の変化を 3 次元的に可視化した．伊藤らの手法を用いることで，特定キーワードに関する話題の変化を可視化することができる．しかし，キーワードとその関連単語の間にある関係は明示されず，ユーザの推測に任せている．

田代ら[田代 2013]は，文書から事象を抽出し，単語間の関係をネットワークとして可視化する手法を提案した．ここで，田代らの定義する「事象」とは，実世界の出来事を表す情報のことである．具体的な手法は，まず，係り受け解析とCRFによって文書中における 5W1H 等の事象属性句を推測する．その後，人手で定義したルールによって事象属性句間の関係を推測し，事象情報として LOD により表現する．田代らはこの手法を用いて特定テーマに関する文書群中の事象情報を可視化し，メディアによる情報の差異を分析した．一方で，田代らは情報の時系列的変化は研究の対象としていない．本論文では，田代らの手法を用いて文書中の事象情報を可視化する．加えて，文書の公開時期を考慮することで，文書の題材に関する情報の推移を分析する．

連絡先：河田裕成，e-mail：17t0009@cis.k.hosei.ac.jp

3. 提案手法

本論文では，時系列的な情報変化を扱うために，特定テーマに関する，情報公開日時の明示された文書を分析対象とする．その例としてニュース記事，歴史年表，SNS への投稿文等が挙げられる．図 1 に示すような提案手法の流れを説明する．

まず，2 章で述べた田代らの手法を用いて文書から事象情報を抽出し，文書中の事象情報を有向グラフとして得る．この事象情報グラフを 2 通りのアプローチで操作することで，時系列的な情報変化と文書のテーマに関する重要情報をそれぞれ可視化する．以下，3.1 節において，事象情報グラフの概要を述べる．その後，事象情報グラフを用いた，情報変化の可視化手法を 3.2 節で，重要情報の可視化手法を 3.3 節でそれぞれ説明する．

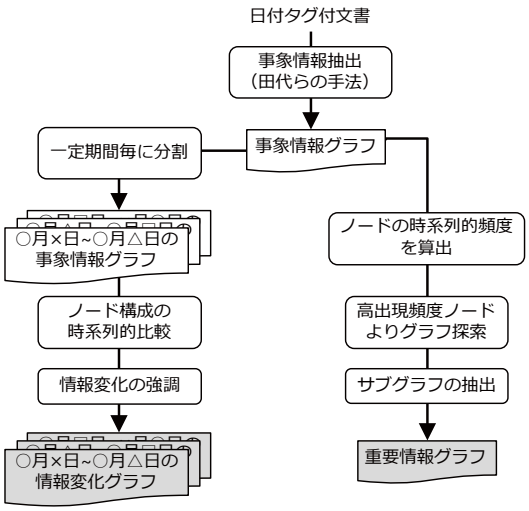

図 1 提案手法の流れ

3.1 事象情報グラフ

田代らの手法による事象情報グラフは，形態素からなる事象属性句をノードとし，「Activity」や「Status」，「Location」など、13種の関係をエッジとしている．これらのエッジは，文書内における事象属性句の係り受け関係や，それを構成する形態素の品詞情報，手がかり語を用いるルールによって設定される．なお，ノードとなる語は形態素解析等により原形化することで，活用による表現揺れを可能な限り吸収している．例えば，「会社員の男が本を買った」という文から生成される有向グラフでは，「会社員」に Modifier 関係で修飾された「男」から，「買う」に動作を表す Activity 関係でエッジが張られる．さらに「買う」は，動作対象を表す Object 関係で「本」に接続される．

3.2 情報変化グラフの生成

時間経過による新たな出来事の発生や，過去の情報に対する補足や訂正がなされることを，時系列的な情報変化とする．これを可視化した情報変化グラフは，時系列中のある期間に出現した事象情報群と，それらが直前の期間からどのように変化したかを表現する．その中で，事象情報は 3.1 節で述べたノードとエッジからなる有向グラフとして表現されるが，各ノードが情報変化の種別を表すように着色されているという点で異なる．ここでの種別とは，直前の時期に出現したノードの①削除，②継続した出現，対象時期における③新規ノードの出現の 3 つである．

ユーザは着色されたノードとそれらの接続関係を見ることで，情報変化を分析する．例えば，②継続出現のノードに対して①削除と③新規出現のノードが状態を表す「Status」関係で接続されている場合，①のノードが表す単語に関して，時間経過による状態の推移が発生したと考えることができる．図 1 左側のフローに示す，情報変化の分析と可視化手法を説明する．

まず，文書の公開日情報を用い，事象情報グラフをユーザの指定する一定期間毎に分割する．各グラフは，対応する期間に出現した事象情報群を表す．次に，グラフ群を時系列順にソートし，直前期間のグラフとのノード構成の差異を計算する．これにより，先に述べた①，②，③が得られる．可視化では，各グラフ中の①のノードを青色，②を黒色，③を赤色に着色する．

3.3 重要情報グラフの抽出

本論文での文書の重要情報とは，文書の扱うテーマを説明するものとする．例えば，自動車事故に関する文書であれば，当事者の名前や事故の起きた状況等が重要情報にあたる．事象情報グラフにおいて，重要情報は文書のテーマに強く関連する単語ノードや，それらと近い距離にあるノードからなるサブグラフとして現れると考えた．そして，文書のテーマに強く関連する単語は，そのテーマを扱う文書群において時系列に関わらず出現しやすいことに注目した．以上より，重要情報グラフは事象情報グラフから抽出したサブグラフに対して，ノードが表す単語の出現時期が判別できるように，ノードの着色処理を行うことによって可視化することとした．

ユーザは着色されたノードによるサブグラフを見ることで，高頻度で出現する単語にどのようなものがあるか知ることができる．また，高頻度語間の関係や，それらを低頻度のノードが表す単語とその出現時期を分析することができる．以下に，図 1 右側のフローで表される，重要情報の可視化手法を説明する．

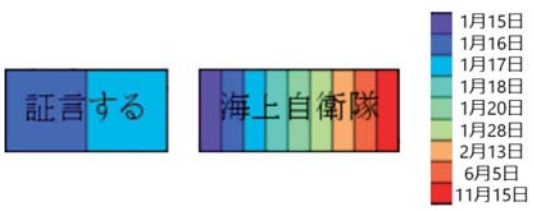

図 2　単語の出現時期を示すノードの着色例

始めに，文書の公開日情報を用いて，単語の時系列的な出現頻度を計算する．これは，時系列の各期間での単語の出現有無を 2 値で求め，それらを合計したものである．次に，出現頻度が閾値以上であるノードを求める．その後，これらのノードを始点とし，そこから指定距離以内で到達できるノードを探索する．最後に，探索によって得られたノードと始点ノードから構成されるサブグラフを抽出する．なお，ノードの出現頻度に関する閾値やグラフ探索の距離は，パラメータとして変更可能である．

可視化では，文書の公開時期ごとに色を割り当て，各ノードを出現時期に対応する色で着色している．図 2 に着色例を示す．ここでは，ノード「証言する」は時系列的に初期の 2 文書のみに出現し，ノード「海上自衛隊」は全ての文書において登場した．よって，「証言する」は青色 2 色で，「海上自衛隊」は青色から緑色，そして赤色までのグラデーションとなるように着色されている．

4. 適用実験

実際の文書群に対して本手法を適用し，情報の変化と重要情報を可視化できているか評価する．本論文では，2014 年 1 月 15 日に発生した海上自衛隊の輸送艦「おおすみ」とプレジャーボート「とびうお」の衝突事故を対象とし，実験を行った．運輸安全委員会の船舶事故調査報告書[*1] によると，この事故では，左舷前方から接近してきた「とびうお」と，それに対し右転して回避しようとした「おおすみ」が衝突したとされる．その結果，転覆した「とびうお」の乗員 2 名が死亡した．この事故に関して，2014 年 1 月から 11 月までに公開された毎日新聞社の新聞記事 9 日分 14 記事を人手で収集し，事象情報を可視化した．図 3 に田代らの手法による事象情報グラフの全体図を示す．このグラフは，431 個のノードと 885 のエッジからなる．提案した 2 つの手法による可視化結果について，以下に述べる．

4.1 情報変化グラフ例

可視化における記事の分割期間は 1 日単位とし，可視化を行った．提案手法では，時系列的に分割した期間毎に事象情報グラフが出力される．本節では，出力例として 2014 年 1 月 16 日に公開された記事に関する情報変化グラフを図 4 に示し，評価を行う．このグラフは，1 月 15 日から 16 日にかけて発生した，記事中の事象情報変化を可視化したものである．

図 4 のグラフには，124 の色付きノードが存在する．その情報変化種別ごとの内訳は，情報の削除が 43，新規出現が 64，継

図 3　田代らの手法で生成された事象情報グラフ

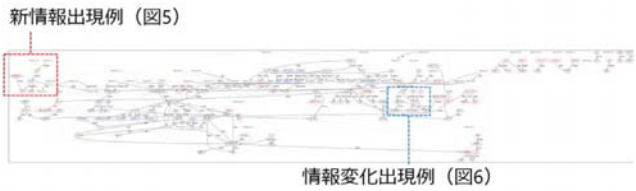

図 4　2014 年 1 月 16 日の記事に関する情報変化グラフ

*1　http://jtsb.mlit.go.jp/jtsb/ship/detail.php?id=6270

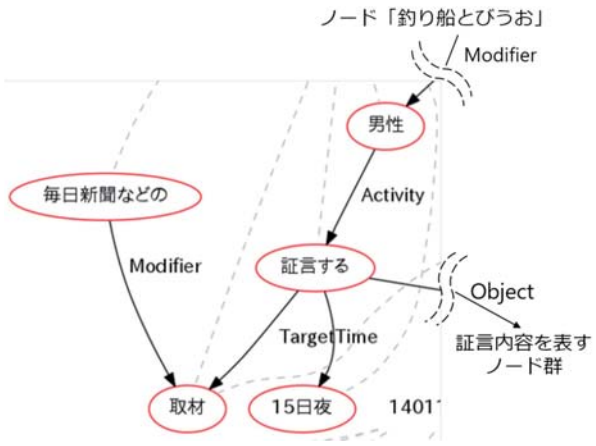

図 5 情報変化グラフ中の新規情報出現例

続出現が 17 ノードである. なお, 継続出現として可視化された 17 ノードのうち 9 つは事故の当事者名や「衝突事故」など, 事故の基本的な情報を表すもので, 図で示した 1 月 15 日から同 16 日までの期間に特有な情報は半数以下であった.

ここで, ノードの色に注目することで, より詳細な情報変化の分析を行う. 図 4 のグラフ中における情報の新規出現例を図 5 に, 特定単語に関する情報変化の例を図 6 に示す. ただし, 図 6 中のノードにおける人物名は黒塗りとして伏せてある.

図 5 では,「新聞社などの取材に対し, 男性が 15 日夜に証言した」という情報の新規出現が可視化されている. また, ノード「男性」に向かう Modifier 関係のエッジは「釣り船とびうお」から伸びており,「男性」が事故で転覆した「釣り船とびうお」の関係者であると把握できる. 同様に,「証言する」からの Object 関係のエッジを辿ることで, 証言の内容を知ることができる.

図 6 では, ノード「2 人」に関して「15 日に病院へ搬送された」から「16 日に死亡した」という時間経過による情報の推移が可視化されている. この「2 人」とは, とびうおの乗員のことを指している. しかし, グラフでは「2 人」と「釣り船とびうお」の間にエッジは存在しないため, グラフのみからそうした情報を得ることはできない. 実際の文書において, 図 6 に示した 2 つの情報はそれぞれ, 1 つの文として出現する. しかしながら,「釣り船とびうお」という単語はこれらの文中に存在しなかった. そのため, 事象情報グラフ構築時にノード間の関係設定が期待通りに行われず, 情報の不足が発生したと考えられる.

以上より, 提案手法は, 時間経過による文書中の単語の出現や削除に注目した情報変化を可視化できる. グラフにおいて, 単語を表すノードは事象内の関係を基に接続される. ユーザはエッジに付与された関係ラベルやノードの色が示す情報変化種別の組み合わせを見ることで, 情報がどのように変化したか知ることができる. 一方で, 提案手法は文中で明示的に書かれていない関係を可視化することはできない. また, 情報変化の構造や, グラフ中におけるその範囲はユーザが判断する必要がある.

4.2 重要情報グラフ例

時系列全体の文書からなるグラフに対し, 全 9 日中 8 日以上の頻度で出現した単語ノードから, 距離 2 で到達できるノードからなる重要情報グラフを抽出した. その結果を図 7 に示す. なお, 時系列的な出現頻度によるノードの着色は, 図 2 の右側に示すように, 全 9 日のそれぞれに色を割り当てることとした.

図 7 において, 閾値の 8 回を超える頻度で出現したノードは「海上自衛隊」「輸送艦おおすみ」「釣り船とびうお」「衝突事故」「瀬戸内海」の 5 つである. これらの高出現頻度ノードの描画された部分を図 8 に示す. 高頻度のノードの表す単語や, それら

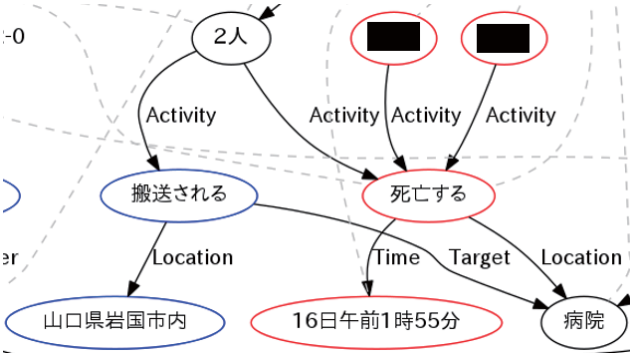

図 6 情報変化グラフ中の情報推移出現例

がどのように接続されているかを見ることで, 文書の対象とするテーマの基本的な情報を知ることができる.

まず, 高頻度のノード間の関係について分析する. 図 8 に点線四角で強調したように,「海上自衛隊」に修飾される「輸送艦おおすみ」と「釣り船とびうお」が, Activity 関係で「衝突事故」に接続されている. 加えて,「衝突事故」は「瀬戸内海」と Location 関係をもっている. 以上のことから, 分析対象の文書が扱うテーマは, 瀬戸内海で発生した海上自衛隊の輸送艦と釣り船の衝突事故であることが分かる.

次に, 高頻度ノードに接続される低頻度ノードにどのようなものがあるか分析する. 例えば,「衝突事故」は「横切る形」で修飾され, Time 関係で「1 月」や「15 日」と接続されている. また,「輸送艦おおすみ」は「右旋回している」を経由して「釣り船とびうお」や「衝突事故」と接続されている. こうした関係から, 事故が発生したのは 1 月 15 日であることや, 輸送艦は右旋回中に釣り船と衝突したという情報が出現したと分かる.

ここで, ノードの色によって表現される単語の出現時期に注目する. 図 9 では,「釣り船とびうお」に修飾された「2 人」は,「搬送される」と「死亡する」に Activity 関係のエッジをもつ. これらのノードの色によれば,「搬送される」は時系列の最初期のみに出現する. そして,「死亡する」は「搬送される」の出現直後と時系列の末期周辺に出現している. これは, 図 6 で示したような情報変化が, 図 9 での点線矢印のようなノード色の変化として現れていると解釈できる. また,「とびうおの乗員が死亡した」という情報は, 始めに出現した後はしばらく文書中で取り上げられなかったことが分かる. 実際, 分析対象の記事では, この情報が出現しなかった 1 月 17 日から 28 日までの記事では, 1 記事を除いて調査で判明した事実や関係者の証言のみを述べていた. 一方で, 例外である 1 記事には, それらに加えて「死亡した乗員 2 人のうち 1 人の通夜が営まれた」という情報が含まれていた. この記事では「2 人」という単語を用いずに以上の情報を述べていたため, ノードの出現頻度計算に誤りが生じたと考えられる.

このように, 提案手法による時系列的な出現頻度を用いたサブグラフでは, 高頻度で出現する情報と, それに付随する低頻度な情報を可視化することができる. 特に, 高頻度ノードが示す単語や, それらの関係は, 3.3 節で定義した重要情報を表すものであった. また, 低頻度ノードは, 重要情報に関連する情報が, 時系列の中でどのように出現したかを示した. ただし, 低頻度の情報にはある一時期のみ注目されたものや後に誤りだと判明したものなど, 様々な種類があることが推測される. 従って, ユーザは低頻度の情報が事実を正しく表しているとは限らないことに注意する必要がある.

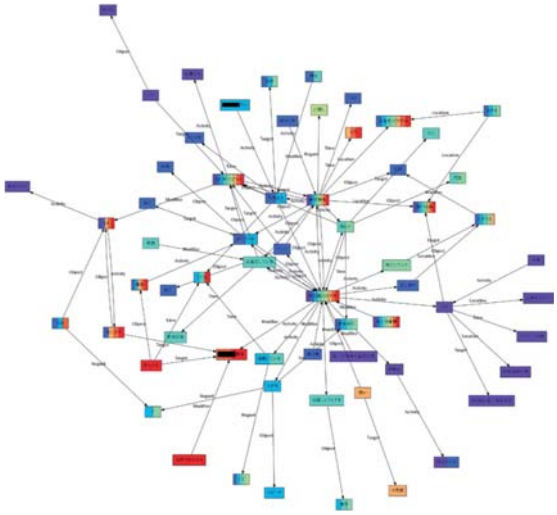

図 7 グラフ探索によって得られた重要情報グラフ

5. 課題

4 章での可視化結果より得られた，提案手法における課題について述べる．提案システムで利用する事象抽出手法に関する問題の 1 つとして，グラフ中におけるノードの多さが挙げられる．その原因としては，「こと」や「する」等，それ自体では多くの情報を持たないノードの存在がある．こうしたノードは他のノードと接続されやすく，グラフ操作による削除が難しい一方，表示が煩雑になる．もう 1 つの問題として，文中における単語の省略に起因する，エッジの設定ミスがある．先行研究と異なり，本論文ではグラフの分割を行うために 1 グラフ当たりの文書数が少なくなりやすい．そのため，グラフ中の情報欠損が比較的発生しやすく，分析の支障となる可能性がある．

これらへの対策としては，田代らの事象抽出手法を改良することによる，ノードとなる単語のフィルタリングや，照応解析による情報の補完などが考えられる．

(1) 情報変化グラフにおける課題

情報変化の可視化では，連続する 2 期間における短期的な変化を計算している．そのため，初期に錯綜していた情報が次第に訂正され，統一されていくといった，長期的視点での情報変化を可視化することができない．また，情報変化を検知する以上の処理を行わないため，情報の矛盾や人物状態の推移等，情報変化の特徴を判別することができない．

これらの改善策として，情報変化の対象や変化パターン等，長期的変化の可視化に有用な情報の利用を検討中である．

(2) 重要情報グラフにおける課題

提案した重要情報可視化手法は，ノードがどのような単語を表すかを考慮せず，出現頻度とノード間の距離によってサブグラフを抽出する．そのため，特にサブグラフの末端ノードにおいて，それがどのような事象に出現したのか解釈するのが難しい．例えば，図 7 のグラフでは「衝突事故」が「に関し」と接続されているが，「に関し」の対象となるノードは抽出できていない．

この問題の解決策として，ノード間の関係やエッジの向きを考慮することで，ノードの削除や追加取得を行うことが考えられる．

6. おわりに

本論文では，時系列的に複数作製される文書群における情報の変化と，文書の対象とするテーマに関連する重要情報の

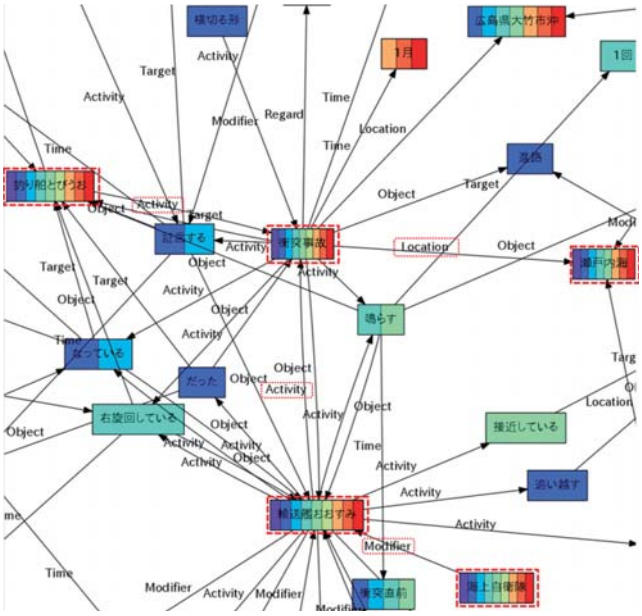

図 8 高出現頻度ノードに関する部分

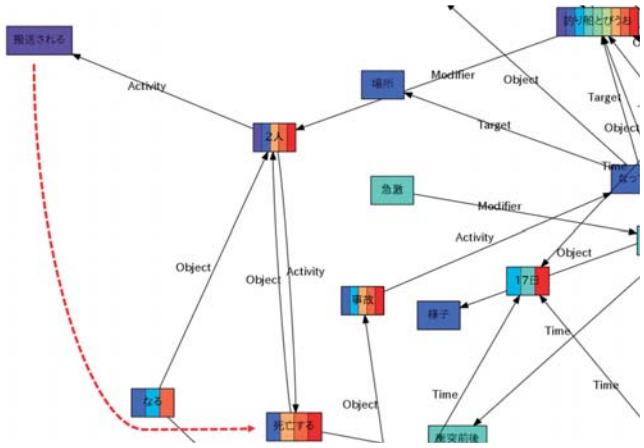

図 9 時間経過に関する重要情報例

可視化手法を提案した．その後，実際の新聞記事に対して提案手法を適用し，可視化結果を評価した．

実際の文書群に適用した結果，提案手法は短期間での情報の新規出現や時間経過による変化を可視化できることを確認した．また，重要情報として文書のテーマを表すような情報や，それに関する様々な情報を可視化できることを示した．

今後の予定としては，人物等に注目することによる長期的な情報変化の可視化や，ノード間の関係などを考慮することによる重要情報のより高精度な抽出などを考えている．

参考文献

[田代 2013] 田沼和浩，王冕，越川兼地，森田武史，長野伸一，清雄一，中川博之，田原康之，川村隆浩，大須賀昭彦: “Linked Data を用いたソーシャルメディア×マスメディアの比較実験”，人工知能学会全国大会論文集，2013.

[伊藤 2012] 伊藤正彦，吉永直樹，豊田正史，喜連川優: “係り受け解析を用いたブログユーザの行動・興味に関する時系列推移 3 次元可視化システム”，電気情報通信学会論文誌 D，Vol. J95-D，No.7，pp.1454-1466，2012.