

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

Experiment of Social-Mass Media Comparison with Linked Data

田代 和浩^{*1} 王 冕^{*1} 越川 兼地^{*1} 西村 悟史^{*2} 森田 武史^{*3} 長野 伸一^{*4}
 Tashiro Kazuhiro Oh Men Koshikawa Kenji Nishimura Satoshi Morita Takeshi Nagano Shinichi
 清 雄一^{*1} 中川 博之^{*1} 田原 康之^{*1} 川村 隆浩^{*1} 大須賀 昭彦^{*1}
 Sei Yuichi Nakagawa Hiroyuki Tahara Yasuyuki Kawamura Takahiro Ohsuga Akihiko

^{*1}電気通信大学情報システム学研究科

Graduate School of Information Systems, University of Electro-Communications

^{*2}大阪大学産業科学研究所

Department of Information and Physical Sciences, Graduate School of Information Science and Technology, Osaka University

^{*3}青山学院大学社会情報学部

Department of Social Informatics, Aoyama Gakuin University

^{*4}株式会社 東芝 研究開発センター

Corporate Research & Development Center, Toshiba Corp.

Recently, the amount of information is growing explosively. Due to the expansion of the WWW, various opinions are uploaded to social media and access to mass media information becomes easy. On the other hand, there are suspicion of biased coverage in mass media and spread of false rumor in social media.

In this paper, we propose a system which extracts and visualizes the information of a certain event from both medias as network. This system extracts the event information (WH-phrases) from tweets and news articles, and converts it to Linked Open Data. Then, it searches the comparable points by SPARQL queries using network structure such as node and edge. Finally, we show efficiency of this system through users' evaluation.

1. はじめに

近年、人が触れる情報量は爆発的に増加している。WWWの普及によって様々な意見がソーシャルメディア上にあげられ、マスメディアの情報にも容易にアクセスできるようになった。特に著しいソーシャルメディアの台頭により、従来のマスメディアにおける偏向報道や情報操作の疑いが強まっているが、ソーシャルメディアにおいてもデマ情報の拡散などの問題がある。本論文では LOD を用いて、ソーシャルメディア・マスメディアの情報を見える化し、両メディアの比較を行うシステムを提案する。両メディアから抽出した事象情報 (5W1H 等) を事象ネットワークとして表現し、その特徴を比較することで、メディア比較を行う。それに加えて、ノード・エッジなどのネットワーク構造情報を利用した SPARQL 検索により、比較ポイントの検出を行う。検出された比較ポイントのユーザー評価を通して、メディア情報を LOD で表現することの有効性を示す。

2. メディア情報の Linked Data 化手法

本論文で扱う“事象”という用語の意味は実世界の出来事を表す情報 (e.g. 「日経平均株価が上昇している。」、「安倍氏と東京駅で握手をした。」) を指す。本研究では、これら事象情報を計算機が扱いやすい整理された情報にするために、各メディアが伝える事象情報 (ニュース記事・ツイートなど) から、セマンティクス情報を考慮した構造化された情報として事象情報を再表現する (Linked Data 化)。Linked Data 化とは、各

メディアに投稿された自然文から事象属性 (2.1 節で説明) を抽出し、それらをまとめて意味ネットワーク (各エッジは抽出された事象属性、各ノードは事象属性に対応するキーワード) を形成することを意味している。

2.1 事象属性の追加

事象情報を構造化された情報として再表現するために、Nguyen らが定義した行動属性をベースに拡張し、事象を表現する上で適した事象属性を定義した。本節では定義した事象属性について説明する。

Nguyen らは、Web 空間から人間の行動情報を抽出する際に行動属性と行動間の遷移を (行動主: Who, 動作: Action, 対象: What, 場所: Where, 時刻及び場面: When, 行動間の遷移ラベル 次: Next, 後: After, 行動間の因果関係: Because of) として定義した。しかし、行動抽出を目的としていたため動作を含まない文は対象としていなかった。それに対し、本論文における“事象”とは、人間の行動を表さない内容も対象としている (例えば、“京王線 新宿～調布間・運転見合わせ中。 / 調布駅混んでる。”など)。そこで、ニュース記事を分析し事象を表現する上で必要十分となる事象属性を Nguyen らの行動属性を拡張する形で 13 の事象属性を表 1 のように定義した。

2.2 CRF を用いた事象抽出

前節で定義した事象属性を用いた、自然文から事象情報を抽出し、Linked Data として出力する本手法の流れを図 1 に示す。尚、対象言語は日本語である。

ここでは、まず前処理として Twitter 特有の文字列の置換や形態素解析辞書への登録を行った後、形態素解析／構文解析を掛け、その結果を CRF に通して事象属性を推測している。また、事象属性と文中の文字列との対応関係を同定する過程 (CRF におけるテストデータ、訓練データの形式、特徴モデルの生成) については先行研究 [1] を参考にして欲しい。

連絡先: 田代和浩 電気通信大学 大学院情報システム学研究科
 大須賀・田原研究室
 〒182-0021 東京都調布市調布ヶ丘 1-5-1
 E-mail: k-tashiro@ohsuga.is.uec.ac.jp

表 1: 事象属性一覧

事象属性	意味	先行研究との対応
Subject	主題	Who に相当
Activity	動作	Action に相当
Object	動作の目的語	What に相当
Time	事象が起こる時刻	When に相当
Location	事象が起こる場所	Where に相当
Cause	事象が起こる原因	BecauseOf に相当
Next	ある事象の次の事象	Next に相当
Target	動作の対象主	追加
Status	主題の状態	追加
Quoted source	情報の発信元	追加
Regard	事象を捉える立場/観点	追加
Modifier	修飾句	追加
Case	事象が起こる条件	追加

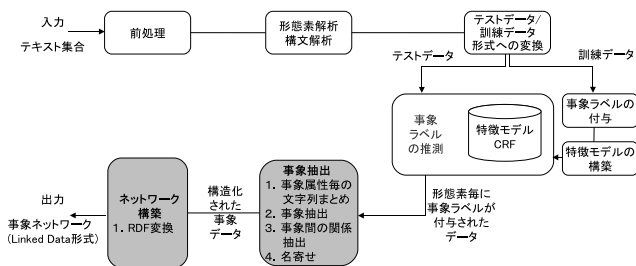


図 1: 事象情報抽出の流れ

次に、CRF の出力（形態素毎に事象属性が付与された情報）から文節情報や品詞情報を手掛かりとしたヒューリスティックルールを用いて“事象”を抽出する。ここで事象とは、1 つ以上の事象属性をエッジとし、事象属性の Activity(行動), Status(状態) を中心とした意味のまとまりを持つ部分グラフである。

事象の例として、以下の例文に含まれる事象情報を前節で定義した新事象属性を加えた事象属性で表現したネットワーク（事象ネットワーク）を図 2 に示す。

例文: 自民党の安倍晋三総裁は 16 日夜、党本部で記者団に対し、衆院選で惨敗した民主党の野田佳彦代表が辞意を表明したことについて「手ごわい相手でもあった」とねぎらった。

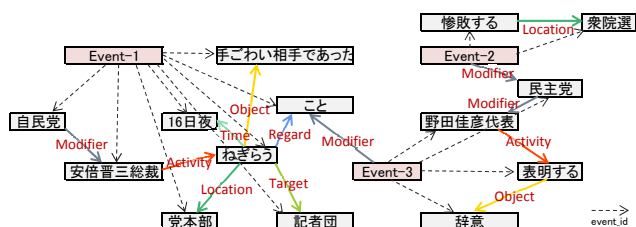


図 2: 事象ネットワーク例

図 2 には、3 つの事象が存在する。（事象 1: 「安倍晋三総裁がねぎらった」こと、事象 2: 「衆院選で惨敗した」こと、事象 3: 「野田佳彦代表が辞意を表明した」こと）

2.3 事象抽出精度の評価

前節で提案したヒューリスティックルールの有効性を検証するため、事象抽出の精度を評価した。尚、ベースとなる CRF

による事象属性の推測精度に関しては、Nguyen らの先行研究 [1] とほぼ同様であり、ニュース等では F 値約 83%, ツイートでは約 72% という結果が得られている。但し、精度は属性によって異なる。ここでは、事前に事象属性が正しく付与されたデータから事象群の抽出を行い、人手で作成した正解と比較することで、事象単位および事象間の関係の抽出精度を算出した。個々の事象の正解基準は、正解データとして用意した事象属性（エッジ）とキーワード（ノード）の組みが全て一致するかどうかである。表 2 に結果を示す。この結果から、先行研究における品詞情報のみで事象情報をまとめる手法に比べ、より高い精度で抽出できることを確認した。

表 2: CRF 出力からの事象・事象間関係の抽出精度

	事象抽出		事象間関係抽出
	提案手法	先行研究	
適合率 (%)	74.7	53.9	97.4
再現率 (%)	74.7	44.3	79.2
F 値	74.7	48.6	87.4

3. メディア情報の比較事例

3.1 オスプレイに関する報道比較

本章では輸送機オスプレイの導入問題に関して、*1 ソーシャル、マス両メディアから集めた情報を対象に、Linked Data 化を行い、両者の比較を行う [2]。

ソーシャルメディアとして Twitter を選定し、事前に作成したオスプレイに関するキーワードリストを用いて関連ツイートを取得した。マスメディアとして主要新聞社 5 社、および主要テレビ局 6 社のニュースサイトの記事を取得した。データの収集期間はいずれも 2012/04/01~2012/08/16 とし、約 3000 ツイート、535 記事を収集した。その上で、rdf-gravity*2 や Gephi*3 を用いて、ネットワークを可視化した。

ノードの色、およびエッジの色を表 3 に示す。また、ネットワークの一部を図 3 に示す。この図から「配備」ノードから「オスプレイ」ノードへ Object 属性で繋がっており、また「配備」ノードから「米軍普天間飛行場」「沖縄」ノードにも Object 有向エッジが伸びていることから、オスプレイが沖縄もしくは普天間飛行場に配備されることがネットワークから読み取れる。また、「配備」に「反対」していること声があることや、紫色のノードはソーシャルメディアでのみ出現する語を意味しているため、沖縄ローカル紙である琉球新報からの情報はソーシャルメディアでしか得られなかったことが読み取れる。以下では、ソーシャル、マス両メディアによるネットワークを人手により比較し、注目すべき観点の洗い出しを行った。

3.1.1 希少事象へのアプローチ

一般的に頻度は意思決定の重要な判断材料になり得るが、場合によっては少数の意見の中に重要な情報が含まれており、頻度分析ではこれらマイナーな意見はノイズとして判断され、取りこぼされてしまうことがある。したがって、本手法ではマイナーな意見も最終的なネットワークに現れるように配慮している。図 4 は Twitter から構築したネットワークの一部である。

*1 存日米軍基地に現行の旧型輸送機に代わり新型輸送機オスプレイを搬入しようとする/した動きに対して、オスプレイの安全性などの観点で導入することに対して賛否がわかれている問題を指す。

*2 <http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html>

*3 <http://oss.infoscience.co.jp/gephi/gephi.org/index.html>

表 3: ノード, エッジの色

ノードにおける マスメディア の構成比率		エッジの種類	色
色	0%	Activity	赤
	37.5 % 未満	Object	黄
	62.5 % 未満	Time	緑
	100 % 未満	Location	青
	100%	Cause	紫
		Target	赤
		Status	青
		Quoted source	黄
		Event.Id	白

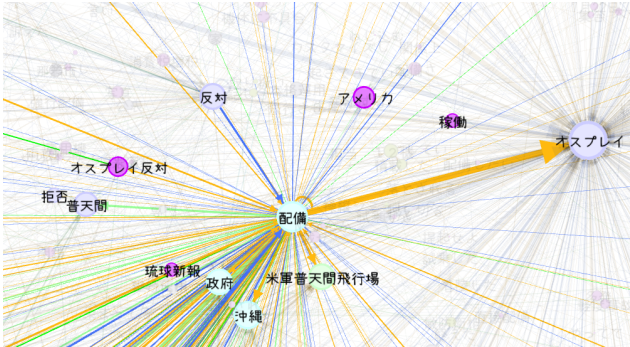


図 3: オスプレイに関する事象ネットワーク (配備ノードに着目)

この図には、マスメディアにはないオスプレイ配置に対する賛成意見が現れている。

更に、Twitter から構築した事象ネットワークの一部（図 5, 6）に注目すると、マスメディアにはないノード「MV-22」、「CV-22」がソーシャルメディアに顕著に現れていることが分かる。これらはオスプレイの型番であり、MV-22（正確には MV-22B）は主に輸送を目的とした機種（事故率 1.93）、CV-22（正確には CV-22B）は特殊作戦を目的とした機種（事故率 13.47）を表している。じつは、日本に配備されるのは輸送型 MV-22 であり、約 40 年前から配備されている現行 CH-46 ヘリコプターの事故率 2.43 より低い。したがって、事故率だけ見れば現状より安全になるのにも関わらず、マスメディアでは両機種を同一に扱い、「CV-22」の事故率を持ち出して「オスプレイは危険である」という主張を繰り返しているようにも伺える。

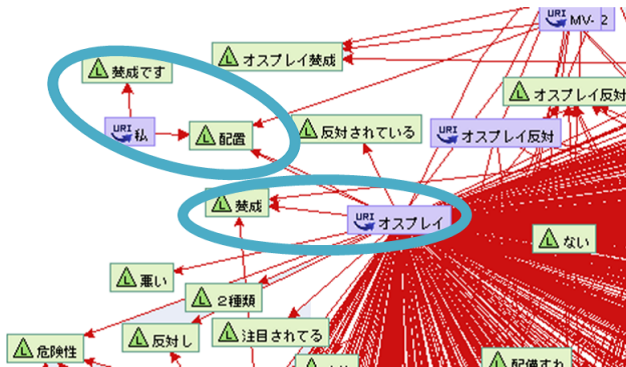


図 4: オスプレイに関するサブネットワーク (賛成意見)

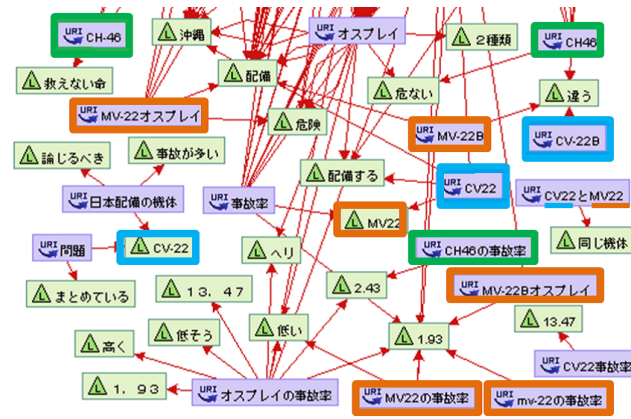


図 5: オスプレイに関するサブネットワーク (MV-22, CV-22)

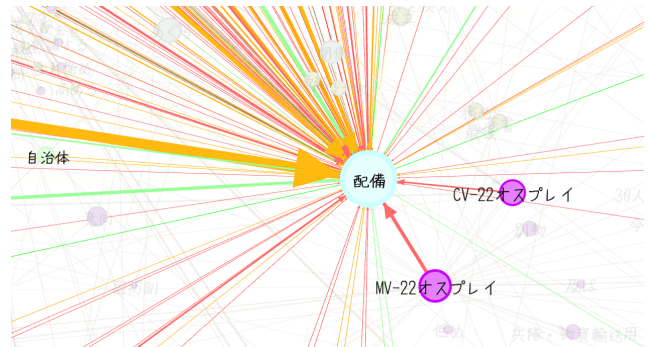


図 6: オスプレイに関するサブネットワーク (配備ノードを Activity 属性で参照している様子。ノードの色がソーシャルメディアでしか出現しないことを示す)

3.1.2 事象の偏在性

例として地域毎の情報の偏在性が挙げられる。オスプレイの事例では、配備後の影響度合いからか、人々の関心は居住地域と強い相関があった。言うまでもなく、当事者である沖縄地域が最も高い関心を示し、関東以北に住む人ほど関心の低い人が多い傾向にあった。そこで、地域や距離を考慮したネットワークを構築することで、その地域特有の問題や現象に対する理解を容易にすることができるだろう。例えば、LOD の活用事例でよく見られるように地図情報と連携させることで、特定の事象への賛成反対意見などの影響範囲を可視化でき、マスメディアと対比して注目すべき情報を見つけることができるだろう。

3.1.3 事象間の因果関係

更に、ある事象が起こる原因を事象間の関係を辿ることで見つけることができる。オスプレイの事例では、先に述べたようにオスプレイ配備に賛成している意見も見られるが、その理由としては米軍が撤退することになれば、中国の脅威が増すとの考えが見られる。更に、中国脅威論の背景には、昨今の尖閣諸島問題に加えて、中国資本が国内の不動産購入を急速に進めていることなどが伺える。このようにネットワークから特定の事象に対する原因（や更にその原因）の可能性を探ることができれば、マスメディアと対比して各主張の裏にある情報の発見を可能とすることを今後検討したい。

3.2 比較観点に基づく注目ポイントの抽出

本節では、前節のオスプレイに関する事例（および、これまでネットワーク構築にあたった経験）から、前節で挙げた希少

性, 偏在性, 因果関係の3点をメディア比較における観点として設定した。そして, 前節とは異なるデータセットから事象ネットワークを構築し, これら3点に基いて注目すべきポイント(ノードとエッジの組み合わせ)を自動的に抽出, それらをユーザに提示することで本研究の目的であるメディア比較支援の有効性を評価する。本研究の特徴はTwitterなどの自然文からキーワード(インスタンス)をキーワード間の関係を表す事象属性(プロパティ)で繋いだグラフ構造(Linked Data)に変換している点であり, 特に本論文では“事象”という単位(1つ以上のトリプルから成る部分グラフ)で一定程度の抽出精度を実現している。そこで, これら事象をSPARQLクエリを用いてグラフ検索することで, 上記の観点に沿ったトリプルの組み(事象)を抽出する。ここでは「第46回衆議院議員総選挙・2012年東京都知事選挙」に関するデータセット(収集機関2012/12/04~2012/12/19, 約14万ツイート, 1220記事)を対象とした。表4にネットワークのノード構成を示す。

表4: 総選挙に関する事象ネットワークに関するデータ
トリプル数

	ソーシャル	マス	合計	
	104,863	105,058	209,921	
ノード数				
	ソーシャル	マス	共通	合計
	71,016 (62.4%)	32,568 (28.6%)	10,272 (9.0%)	113,856

3.2.1 SPARQL クエリによる注目ポイントの検索

クエリーコードは紙面の都合から割愛するが, 各観点について以下のようにエッジを辿ることで注目すべきポイントを検索する。

希少性

ソーシャル, マス両データセットを対象に, event id を手掛かりに各事象を表す部分グラフを抽出し, グラフの形(ノード, エッジの組み)が他と最も類似していない事象を抽出する。但し, ノードの内, 1つは上記, 共通行動/状態/意見を表すトリプルのいずれかに繋がっているものとする(ノイズ除去のため)。

偏在性

ソーシャル, マス各データセットを対象に, Location 属性を含むトリプルを検索し, event id を手掛かりに同一事象内の Subject(に対応する文字列), Activity(に対応する文字列), Object(に対応する文字列)を抽出, データセット毎に Location(に対応する文字列)でソートして出力する。

因果関係

ソーシャル, マス両データセットを対象に, Cause/Case/Next 属性を含むトリプルを検索し, Cause 関係にある両事象内の Subject(に対応する文字列), Activity(に対応する文字列), Object(に対応する文字列)を抽出, 2対1組みとして出力する。

3.2.2 ユーザ評価による有効性の確認

SPARQLを用いた重要ポイントの検索手法の有効性を確認するために前節で紹介したクエリによって返却されたトリプルまたは事象をユーザに提示し, 未知性(「知らなかった」)・嗜好性(「興味深い」, 「面白い」)の2つの評価軸で評価しても

らった。希少性の観点においてクエリで得られた注目ノードと周辺トリプルの情報をユーザに提示し, 注目ノード毎に提示した情報がユーザの嗜好性に一致した場合を正解とした。偏在性と因果関係の観点においてはクエリで得られた事象をユーザに提示し, 嗜好性及び未知性の各評価軸毎に評価結果を得た。大学院生5名による各比較観点毎の適合度の平均値を表5に示す。尚, 総選挙の話題は多岐に及び, 再現率を算出するための十分な正解データの作成が困難であったため, 今回は適合率に留めている。

表5: 注目ポイントの評価

	希少性	偏在性	因果関係
未知性に関する適合率(%)	-	51.74%	42.82%
嗜好性に関する適合率(%)	58.33%	55.08%	51.16%

注目ポイントの評価結果から3つの比較観点の嗜好性に関する適合率が全て50%を超える値を示し, 注目ポイントを推定する上で用いた3つの比較観点が一定の有効性のある手掛かりになることが確認できた。同時に, 用いたクエリはコンテキスト情報を用いずにノード・エッジなどのネットワーク構造情報のみを手掛かりにしているため, ドメインによらずに適用できる汎用性があることも確認することができた。

またクエリから返却された事象には, 事象の抽出結果が正しくないことが多々あり, 注目ポイントの推定結果の精度を下げる要因の一つになっていた。更なる精度向上のためには, 事象抽出精度の改善, クエリの改善, 新たな比較観点の発見, 各観点を元にした機械学習手法の適用などが必要だと考えている。

4. まとめと今後の課題

本論文では, ユーザがメディア情報を多角的な観点から比較することを支援するため, ユーザに代わってソーシャル, マス両メディアから情報を抽出, 見える化し, いくつかの観点から比較ポイントを提示するシステムを提案した。また, 実験を通じて一定の有効性を確認した。

今後の課題としては, 事象抽出精度, 比較観点の抽出精度の改善に加えて, 可視化ツールと比較ポイント提示機能との連携が挙げられる。ネットワークを可視化させた状態で, 注目すべき点をポップアップ, ハイライトなどで注意喚起できれば利便性を高められるものと考えている。また, 本手法はメディア比較にとどまらず, マーケティングや株価予測など事象分析が有効な用途に広く使える技術であると考えており, 今後, 他ドメインへの適用も検討して行きたい。

参考文献

- [1] T. M. Nguyen, T. Kawamura, Y. Tahara, A. Ohsuga: “Self-supervised capturing of users’ activities from weblogs”, Int. J. of Intelligent Information and Database Systems (IJIIDS), Vol. 6, No. 1, pp. 61-76, 2012.
- [2] 越川 兼地, 川村 隆浩, 中川 博之, 田原 康之, 大須賀 昭彦: “CRFを用いたメディア情報の抽出と LinkedData 化 - ソーシャルメディアとマスメディアの比較事例 -”, 合同エージェントワークショップ&シンポジウム (JAWS 2012), 2012.