平成24年度　修士論文

メディア情報からの
事象抽出とLinked Data化の研究

～ソーシャルメディアとマスメディアの比較事例～

電気通信大学 大学院情報システム学研究科
社会知能情報学専攻

1151017 越川兼地

主任指導教員: 大須賀昭彦 教授
指導教員: 田原康之 准教授
指導教員: 太田敏澄 教授

平成25年2月21日 (木) 提出

概 要

近年様々な情報源 (ソーシャルメディア・マスメディア) に容易にアクセスできるようになった．ソーシャルメディアの普及スピードはとどまることを知らず，多様な意見・考え方にインターネットさえあれば容易に触れることが可能になった．しかし Twitter 上でのデマの拡散や，従来型のマスコミにおいても偏向報道や情報操作などが行われているなどの疑いの声が一般ユーザにまで届いてきている．本来の情報収集のあり方である多角的な観点から情報を入手して，自らの観点で特定の話題について理解することがますます重要になっている．このような背景から，昨今ソーシャルメディアとマスメディアとの対比が今取り沙汰されている．

そこで本研究では，ユーザに代わって両メディアから事象情報を抽出し，見える化するサービスを提案する．具体的には，ニュース記事，ツイートから特定の話題に関する情報 (5W1H 等) を条件付き確率場 (Conditional Random Fields(CRF) を用いて) 抽出し，それらを Linked Data 化することで比較，探索，見える化を容易にする．本稿では，2 つの話題に関する Linked Data ネットワークの比較事例を提示する．1 つ目は輸送機オスプレイに関する話題，2 つ目は及び 2012 年に行われた衆議院議員総選挙・東京都知事選挙に関する話題である．最後に本サービスの有効性について議論する．

目 次

図 目 次

表 目 次

第1章　はじめに

近年様々な情報源 (ソーシャルメディア・マスメディア) に容易にアクセスできるようになった．ソーシャルメディアの普及スピードはとどまることを知らず，多様な意見・考え方にインターネットさえあれば容易に触れることが可能になった．しかし Twitter 上でのデマの拡散や，従来型のマスコミにおいても偏向報道や情報操作などが行われているなどの疑いの声が一般ユーザにまで届いてきている．本来の情報収集のあり方である多角的な観点から情報を入手して，自らの観点で特定の話題について理解することがますます重要になっている．このような背景から，昨今ソーシャルメディアとマスメディアとの対比が今取り沙汰されている．

そこで本研究では，ユーザに代わって両メディアから事象情報を抽出し，見える化するサービスを提案する．具体的には，ニュース記事，ツイートから特定の話題に関する情報 (5W1H 等) を条件付き確率場 (Conditional Random Fields(CRF) を用いて) 抽出し，それらを Linked Data 化 [1] することで比較，探索，見える化を容易にする．

次世代のデータ共有方法である Liked Data という概念に注目が集まっている [2, 3]．Linked Data とは，様々な情報源のデータが RDF で記述され，それらが結びついてつくられるデータの集合である．Linked Data の最大の特徴は，異なるデータサイトのデータが容易につながりあうことができる点である．それを可能にしている仕組みは Web 空間でリソース（資源）を一意に指定することのできる識別子 URI（Uniform Resource Identifier）を RDF 記述内に用いている点が挙げられる．これにより，データセットを超えて相互に参照しあうことを可能にしている．

本稿では，2 つの話題に関する Linked Data ネットワークの比較事例を提示する．1 つ目は輸送機オスプレイに関する話題，2 つ目は及び 2012 年に行われた衆議院議員総選挙・東京都知事選挙に関する話題である．最後に本サービスの有効性について議論する．

本論の構成を以下に示す．まず 2 章で事象を表現するための属性の定義について述べ，

事象情報を意味ネットワークを使った表現方法をを説明し，3 章で自然言語集合から事象情報を構造化された状態で抽出する方法及び Linked Data 化への変換方法を提案する．4 章 5 で提案手法の有効性を示すため評価実験を行う．続く 5 章において本研究での提案手法である事象抽出手法をメディア比較へ応用した事例を示す．6 章で提案手法やその周辺の技術の関連研究に触れる．最後に 7 章で本論文のまとめと今後の検討課題について述べる．

第2章　事象を表現するための属性の定義

本論文の課題は，各メディアに投稿される自然言語の文章から事象属性を構造化された状態で自動的に抽出し，抽出した事象属性を事象毎にまとめ，意味ネットワーク形式 (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として) に変換することで元の自然言語の文意を構造化された情報であるネットワークとして事象情報を表現することである.

Nguyen らの先行研究 [4] では 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) のように定義した．Nguyen らは行動情報を抽出することを目的としていたため動作を含まない文は処理対象としていなかった．本論文で抽出する事象とは世の中で起こっている事柄を指しており，“京王線 新宿～調布間・運転見合わせ中．/ 調布駅混んでる．” といった動作を表す表現を含まない文中にも世の中の事象を表すのに十分な情報が含まれていると考え，Nguyen らの行動属性を拡張し，表 2.1 のような事象属性ラベルを定義した.

この自然言語から事象情報を表現するネットワークに変換する例として，以下の 2 文に対し，事象情報を抽出しネットワーク化したものを図 2.1，図 2.2，図 2.3 に示す.

例 1 昨日太郎は秋葉原で iPhone5 を購入したので，幸せそうだった。

例 2 沖縄○×新聞によると，昨夜鈴木氏は沖縄県庁に出向き，知事らに一連の不祥事について弁明した.

例 3 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。

表 2.1: 事象ラベル 一覧

事象ラベル	意味	先行研究の行動属性との関係
Subject	主題	Who に該当
Activity	動作	Action に該当
Object	動作の目的語	What に該当
Time	事象が起こる時刻及び場面	When に該当
Location	事象が起こる場所	Where に該当
Cause	事象が起こる原因	BecauseOf に該当
Next	ある事象の次の事象	Next に該当
Target	動作の対象主	新定義
Status	主題の状態	新定義
Quoted source	情報の発信元	新定義
Regard	事象を捉える立場/観点	新定義
Modifier	修飾句	新定義
Case	事象が起こる条件	新定義

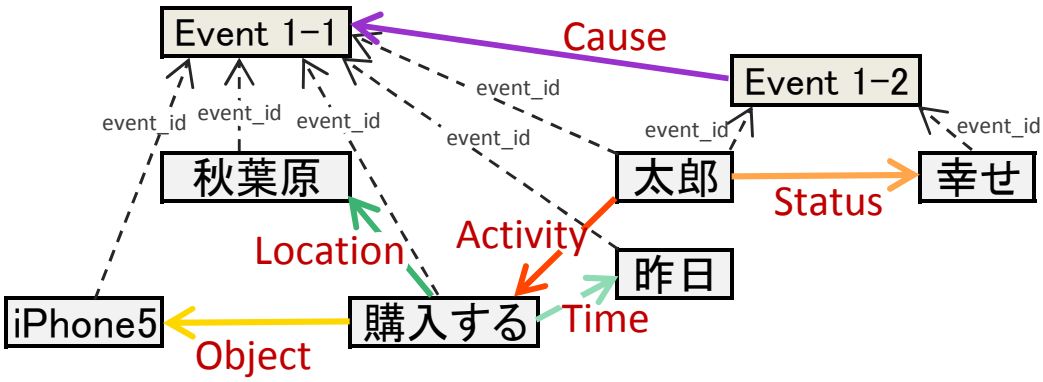

図 2.1: 事象ネットワーク例 1

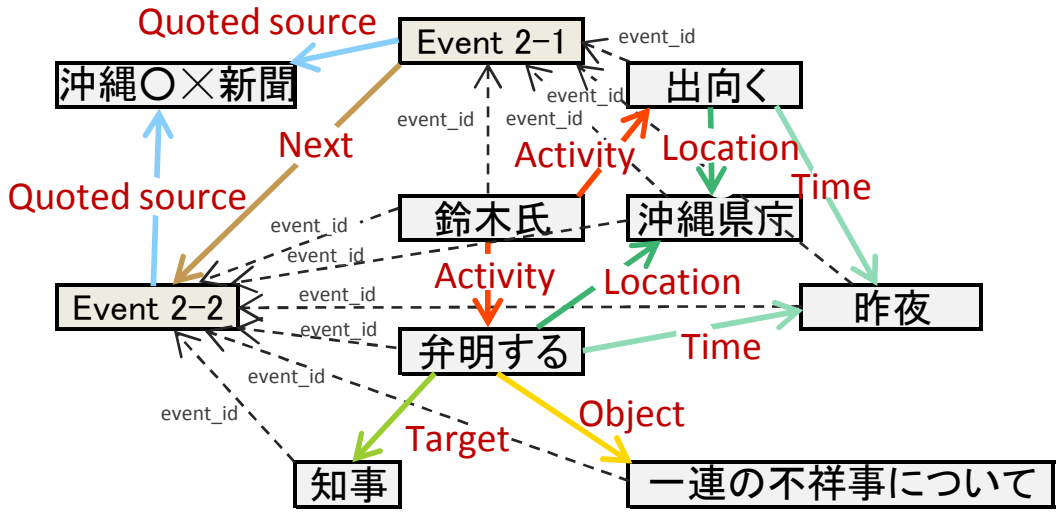

図 2.2: 事象ネットワーク例 2

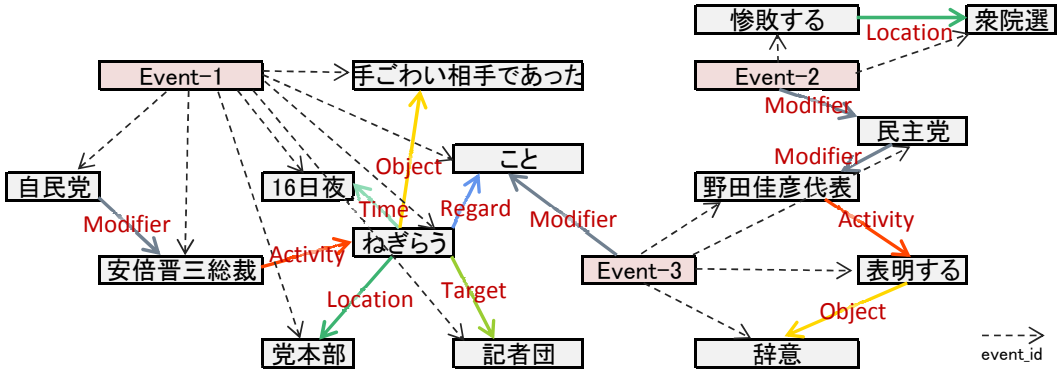

図 2.3: 事象ネットワーク例 3

次章以降では，この事象情報の抽出方法及び，ネットワークへの変換方法について述べていく．

第3章　事象情報抽出手法の提案

自然言語から事象情報を抽出し，Linked Data として出力する本手法のフローチャートを図 3.1 に示す．なお対象言語は日本語を想定している．

3.1　前処理

本節では，事象抽出精度向上のための前処理について説明する．

3.1.1　Twitter 特有表現への対処

Twitter で特定の機能を用いるために投稿するツイート内に含める特有の文字列 (リプライ/ユーザ名/ハッシュタグ/URL/RT/QT) が形態素解析器で解釈されない傾向にあったため，正規表現でそれらの Twitter 特有表現の文字列を検出し，こちらが指定した文字列で置換した後，形態素解析器のユーザ辞書に指定文字列の登録を行い正しく形態素解析処理が行われるように施した．表 3.1 に特有表現への対処をまとめた．

3.1.2　鍵括弧への対処

ニュース記事によく頻繁に見かけられる「」,『』,【】などの鍵括弧は，文構造を複雑にする傾向があり，事象情報をより正確に抽出するために,「鍵括弧内の文字列」と「括弧外の文字列」を分割することで文構造の単純化を測り，事象属性ラベルの推測精度の向上及び事象属性ラベルの付与作業の容易化を狙った．具体的には前に紹介した前処理と同様に正規表現で鍵括弧と該当した部分を指定文字列で置換した後，形態素解析器のユーザ辞書に登録するといった処理を施した．

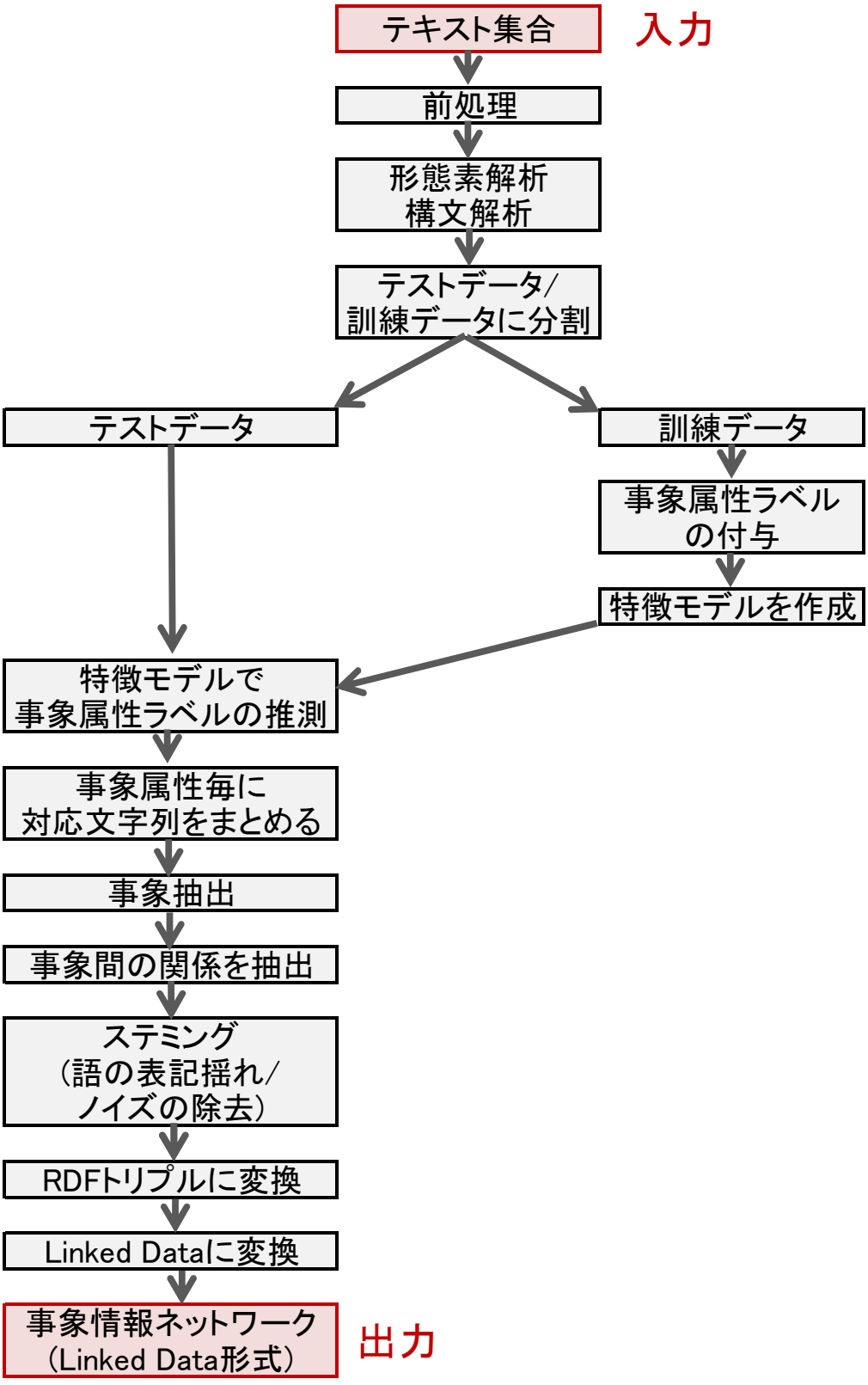

図 3.1: 事象情報抽出手法のフローチャート

表 3.1: Twitter 特有表現への対処

特有表現	対処	特有表現を含む例文	置換後の文字列	正規表現
ユーザ名	文字列置換	昨日 @hogehoge さんに会った。	昨日 [SYSTEM:USER] さんに会った 改行	`@([_0-9a-zA-Z]{3,20})`
リプライ	空白置換	@hogehoge おはよう	おはよう	`^(@([_0-9a-zA-Z]{3,20}))\s(@([_0-9a-zA-Z]{3,20})\s?)*`
RT	改行	RT @hogehoge : おはようございます。	改行 おはようございます 改行	`RT\s@([_0-9a-zA-Z]{3,20})\s?:?`
QT	改行	こんにちは。QT @hogehoge : こんにちは。	こんにちは 改行 改行 こんにちは 改行	`QT\s@([_0-9a-zA-Z]{3,20})\s?:?`
URL	改行	明日は晴れのようです。http://bit.ly/UaQM2	明日は晴れのようです 改行 改行	`(https?|ftp)(:\/\/[-_.!~*'()a-zA-Z0-9;\/?:\@&=+\$,%#]+)`
ハッシュタグ	改行	自民党圧勝 (・∀・) # 政治	自民党圧勝 (・∀・) 改行	`(\s|　)?(#|＃)([_a-zA-Z0-9]+|[一-龠]+|[ぁ-ん一]+|[ァ-ヴー]+)(\s|　|$)|`

3.1.3　形態素解析器用の辞書構築

次工程の CRF を利用した事象属性の推測において形態素解析器および構文解析器を用いた結果を利用する．そこで前処理結果の反映及び解析器自身の解析精度を向上させるべく以下の形態素解析用の辞書を用意した．

前処理用　前述の前処理 (3.1.1 項，3.1.2 項) で置換した文字列

Wikipedia 見出し語　適切な形態素での分割を実現するために，ウェキペディア[1]の 2012 年 12 月時点の見出し語 計 1,342,099 語に対し品詞を名詞としてユーザ辞書を構築した．

なお形態素解析器には MeCab[2]，構文解析器には CaboCha[3]を用いた．

[1] http://ja.wikipedia.org/
[2] https://code.google.com/p/mecab/
[3] http://code.google.com/p/cabocha/

3.2　CRF を用いた事象抽出

本項では条件付き確率場 (Conditional Random Fields)(以下 CRF) を用いた自然言語からの事象抽出手法について説明する．まず CRF について簡単に紹介し，事象情報への適用方法について説明する．具体的にはテストデータと訓練データの形式，特徴モデルを生成する方法を述べる．次に CRF の結果から文節情報や品詞情報を用いた事象抽出のためのヒューリスティックルールについて述べ，最後に Linked Data への変換方法について説明する．

3.2.1　条件付確率場（Conditional Random Fields）とは

条件付確率場（Conditional Random Fields）とは，John D. Lafferty ら [5] が提案した系列ラベリング問題に適用した識別モデルである．

図 3.2: 入力 x と出力 y

図 3.2 入力データを x（例えば，文），出力データを y（例えば，固有名詞）とするとき，特徴モデルによって実現したいのは，x が与えられたときに対応する y が正しく出力されるということである．条件付確率場 (以下 CRF) は一つの指数分布モデルで，この出力系列 $y = y_1, y_2, ..., y_n$ の入力列 $x = x_1, x_2, ..., x_n$ に対する条件付確率 $P(y|x)$ を表す．

$$P(y|x) = \frac{\exp(\alpha, \Phi(x, y))}{\Sigma_{y \in Y} \exp(\Phi(x, y))} \tag{3.1}$$

但し，$\Phi(x, y)$ は系列 $y = y_1, y_2, ..., y_n$ 上のパスの全ての特徴ベクトルを足し合わせたものであり，α はモデルのパラメータである．そして CRF において，新たな x が与えられたときの出力予測 $\hat{y}$ は

$$\hat{y} = argmax_{y \in Y} P(y|x) \tag{3.2}$$

となる．この出力は，Viterbi アルゴリズム [6] を用いることで効率良く解くことができる．

CRF は識別モデル (discriminative model) であり，重複する特徴をモデルに組み込むことができる．通常の識別モデルとの違いは，出力が出力集合の部分集合ではなく，系列となる点である． CRF は，品詞付与 [5]，テキストチャンキング [7]，固有表現抽出 [8]，形態素解析 [9] などといった系列ラベリング問題に適用され，いずれにおいても高い精度を示している．

3.2.2　訓練データ・テストデータの作成

テストデータ形式への変換

前述の前処理を行ったテキストに対し，CRF の入力となるテストデータ形式に変換する必要がある．テストデータに用いる情報は，以下のつの情報を用いた．

1. データ元参照情報 (文章の元記事/元ツイートへの参照)
2. 鍵括弧の対応付け情報
3. 文節情報 (文節 ID / 係り受け先文節 ID)
4. 表層形
5. 品詞 ID

それぞれの情報の取得方法について説明する．“データ元参照情報”については，データベースから処理対象の文章を読み込むときに文章情報と共に取得する．“鍵括弧の対応付け情報”については，前処理 3.1.2 項で述べた鍵括弧置換処理時に取得する．“文節情報”，“表層形”及び “品詞 ID”については，前処理を適用した後のテキストに対して形態素解析・構文解析した結果から得る．

上記の情報を用いて，自然言語から前処理をした後に，表 3.2 のようなテストデータ形式に変換する．このテキストデータは tsv 形式となっており，要素の左から “データ元参照情報”，“鍵括弧の対応付け情報”，“文節 ID”，“係り受け先文節 ID”，“表層形”，“品詞 ID” を表している．この変換したデータを特徴モデル構築時に使う訓練データと残りをテストデータに分割する．

表 3.2: テストデータの例

データ元参照情報	鍵括弧の対応情報	文節ID	係り受け先文節ID	表層形	品詞ID
758-17	0	0	3	一時	67
758-17	0	1	2	帰宅	36
758-17	0	1	2	の	24
758-17	0	2	3	たび	66
758-17	0	2	3	に	13
758-17	0	3	4	朽ち	31
758-17	0	3	4	て	18
758-17	0	3	4	いく	33
758-17	0	4	5	自宅	38
758-17	0	4	5	を	13
758-17	0	5	-1	見	31
758-17	0	5	-1	た	25
758-17	0	5	-1	。	7

訓練データ形式/事象属性ラベルの付与

分割した訓練データ用のテキストデータに対して，人手で 2 で定義した事象属性を付与 (ラベリング) することで，訓練データを作成する．訓練データの例を 表 3.3 に示す．ここで用いている記号『M（Modifier）は修飾句の一部,『B（Begin）』は表現の始まり,『I（Inside）』は表現の途中,『O（Outside）』は表現以外を表す．(表 3.3 は，説明簡易化の理由で “データ元参照情報”，“鍵括弧の対応付け情報”,“文節情報”は省略している．)

表 3.3: 訓練データ例

表層形	品詞ID	事象属性ラベル
一時	67	M
帰宅	36	M
の	24	M
たび	66	B-Time
に	13	O
朽ち	31	B-Activity
て	18	I-Activity
いく	33	I-Activity
自宅	38	B-Object
を	13	O
見	31	B-Activity
た	25	I-Activity
。	7	O

3.2.3　特徴モデルの構築

構築した訓練データに対してどのような素性を与えるかを定義するためのテンプレートファイルを作成した．提案手法が利用する素性は “表層形”，“品詞 ID”，訓練データに付与されている “事象属性ラベル”の 3 つである．テンプレートファイルでこれらの素性を使った特徴を定義する．そして長い文に対応させるために，サイズ 7 のウィンドウ（%x[-3,*]～%x[3,*]）を採用する．図 3.3 にテンプレートファイルの全体を示す．但し，%x[i,j] は現在の位置からの相対位置で i 行目の j 番目の列の要素を指す．また，U**はテンプレートの記号である．このテンプレートファイルと訓練データを用いて，特徴モデルを構築する．

```
# Word column
U00:%x[−3,4]
U01:%x[−2,4]
U02:%x[−1,4]
U03:%x[0,4]
U04:%x[1,4]
U05:%x[2,4]
U06:%x[3,4]
U07:%x[−1,4]/%x[0,4]
U08:%x[0,4]/%x[1,4]

#POS column 's junction
U22:%x[−3,5]/%x[−2,5]/%x[−1,5]
U23:%x[−2,5]/%x[−1,5]/%x[0,5]
U24:%x[−1,5]/%x[0,5]/%x[1,5]
U25:%x[0,5]/%x[1,5]/%x[2,5]
U26:%x[1,5]/%x[2,5]/%x[3,5]

# POS column
U09:%x[−3,5]
U10:%x[−2,5]
U11:%x[−1,5]
U12:%x[0,5]
U13:%x[1,5]
U14:%x[2,5]
U15:%x[3,5]
U16:%x[−3,5]/%x[−2,5]
U17:%x[−2,5]/%x[−1,5]
U18:%x[−1,5]/%x[0,5]
U19:%x[0,5]/%x[1,5]
U20:%x[1,5]/%x[2,5]
U21:%x[2,5]/%x[3,5]
```

図 3.3: CRF で用いた素性を定義するテンプレートファイル

構築した特徴モデルを使って，テストデータの事象属性ラベルの推測を行う．

3.2.4　事象ラベルの推測結果からの事象抽出

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．前節において事象ネットワークの例として挙げた図 2.3 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) CRF 出力からここで説明した事象を抽出する過程は次の 3 ステップから成る．表 3.4 に CRF の出力例を示す．以下にこの出力例について説明する “表層形”，“品詞 ID”，“文節 ID”，“係り受け先文節 ID”，は形態素解析，構文解析の解析結果から得られる情報である．

“データ元参照情報” 原文が保存されているデータベースへの参照情報を示しており，事象ネットワーク構築後に事象と原文との紐付けを行う際に用いる情報である．

“鍵括弧の対応情報” 事象属性と文中の対応文字列を同定する過程において，ニュース記事で頻出する鍵括弧を含んだ文に対して適切に処理するため，鍵括弧の内部の文は別に処理している．事象ネットワーク構築後に鍵括弧が存在した文から得た事象と鍵括弧内部の文から得た事象を紐付けるために用いる情報である．

“事象属性ラベル” 形態素毎に事象属性の対応関係を文字列を事象属性ラベルとしている．事象属性ラベルは文字「B」,「I」と各事象属性の組み合わせと文字「O」,「M」で表現し，それぞれの頭文字は「B（Begin，表現の始まり）」,「I（Inside，表現の途中）」,「O（Outside，それ以外）」,「M（Modifier 修飾句）」を意味している．

表 3.4: CRF 出力例

データ元参照情報	鍵括弧の対応情報	文節ID	係り受け先文節ID	表層形	品詞ID	事象属性ラベル
329-10	0	0	1	半年	67	M
329-10	0	0	1	間	51	M
329-10	0	0	1	の	24	M
329-10	0	1	2	運航	36	B-Object
329-10	0	1	2	休止	36	I-Object
329-10	0	1	2	を	13	O
329-10	0	2	4	余儀なく	10	B-Activity
329-10	0	2	4	さ	31	I-Activity
329-10	0	2	4	れ	32	I-Activity
329-10	0	2	4	た	25	I-Activity
329-10	0	3	4	大きな	68	O
329-10	0	4	5	要因	38	B-Subject
329-10	0	4	5	が	13	O
329-10	0	5	-1	高速料金	38	B-Status
329-10	0	5	-1	だ	25	O
329-10	0	5	-1	。	7	O

CRF の出力形式から事象を抽出するヒューリスティックルールはまず，各事象属性に対応する文字列のまとめ (3.2.4 項)，次に，意味的関連がある事象属性でまとめることで事象を抽出し (3.2.4 項)，最後に，事象間の関係を抽出 (3.2.4 項) することで，CRF の出力形式 (例:表 3.4) から構造化された事象情報 (例: 表 3.9) を得る．

以下にこの 3 つの工程について説明する．

各事象属性に対応する文字列のまとめ

CRF の出力結果から事象を抽出するにあたり，まず CRF の出力形式 (入力例:表 3.4) から各事象属性に対応する文字列にまとめ，修飾句の修飾先の事象属性を同定する必要がある (出力例:表 3.5).

そこで，事象属性ラベルに付与されている先頭の頭文字「B」,「I」,「O」,「M」に着目し，以下のヒューリスティックルールを用いて各事象属性に対応する文字列をまとめる．

1. 抽出する属性，および対応する文字列は BIO タグが B で始まる単独の形態素，もしくは複数の連続した形態素で構成される．

2. 修飾句は単独もしくは連続する形態素で構成される．

3. 修飾句は自身の修飾句の次の形態素から始まる属性を修飾する．

上記のルールをアルゴリズムにしたものをソースコード 3.1 に示す.

ソースコード 3.1: 事象属性を束ねる/修飾句の修飾先の属性を特定するアルゴリズム (Python)

```python
def save_attr(attribute_list, attr, attr_str, m_str, before_attr):
  """属性ラベルを追加する"""
  if not attr_str == "" and not before_attr == None:
    if not m_str == "":
      m_attr = "M-" + attr
      attribute_list.append((m_attr, m_str))
    attribute_list.append((attr, attr_str))

def grouping(tokens):
  """属性ラベルを束ねる"""
  def ini_variable():
    """変数を初期化する関数"""
    return "", "", None
  # 変数の用意
  attribute_list = [] # 属性ラベルと文字列の対応関係を格納するための変数
  m_str, attr_str, before_attr = ini_variable()
  for t in tokens: # 形態素を回す
    tag, attribute = t.get_tag() # BIOM tag, 属性ラベルを得る
```

```
    if tag == "M":
      # 修飾句に対しての処理
      attr_str, before_attr  = ("", None)
      m_str += t.surface # + 表層形
    else:
      if tag == "B": # B:Begin
        save_attr(attribute_list, attr, attr_str, m_str, before_attr)
        m_str, attr_str, before_attr = ini_variable()
        attr_str += t.surface # + 表層形
        before_attr = attribute
      elif tag == "I": # I:Inside
        attr_str += t.surface # + 表層形
      elif tag == "O": # O:Other
        save_attr(attribute_list, attr, attr_str, m_str, before_attr)
        m_str, attr_str, before_attr = ini_variable()
  save_attr(attribute_list, attr, attr_str, m_str, before_attr)
  return  attribute_list
```

このルールを例:表 3.4 に適用すると，表 3.5 のように各属性に対応する文字列を得ることができる．表 3.5 の “M-Object”という事象属性は直後の事象属性 Object を修飾する (Modifier) ことを意味している．

表 3.5: 事象属性毎の文字列

事象属性	対応文字列
M−Object	半年間の
Object	運行休止
Activity	余儀なくされた
Subject	要因
Status	高速料金

事象毎の事象属性のまとめ

前工程では各事象属性に対応する文字列をまとめたが，図 2.3 のように 1 文に複数の事象が存在することが多々あるため，安易に前工程で得られた結果を一事象としてみなしてはならない．文意を踏まえた上で事象毎に意味の関連がある事象属性をまとめ，各事象を構成する必要がある．

そこで，文節間の関連情報である係り受け解析の結果と事象属性の Activity(行動)，Status(状態) に着目し，以下のような事象毎に事象属性をまとめるルールを導出した．説明時に用いる例に，本工程の入力として CRF の出力結果 3.4 及び各事象属性情報 3.5 を想定した．

1. 事象属性 Activity/Status が出現する文節を文意の区切りとみなし，事象属性 Activity/Status が出現する文節 ID(*e_id* とする) のリストを作る．尚，末尾の文節 ID がリストに含まれていない場合は加える．以降，本リストを *e_list* とする．
   例: *e_list* ⇒ [2,5] (対応文字列:“余儀なくされた”，“高速利用金だ。”)

2. *List* を文節 ID の昇順に並べ替える．要素の *e_id* に着目し，その *e_id* を係り受けしている文節 ID をまとめる．尚，既に他の *e_id* としてまとめた文節 ID は含めないものとする．
   例: *e_id*: 2 ⇒[0,1,2] (*event*1 とする), *e_id*: 5 ⇒ [3,4,5] (*event*2 とする)

3. *e_id* 毎に束ねられた文節 ID に含まれる事象属性の組み合わせを一事象とみなし，各事象を抽出する．
   例: *event*1 [0,1,2] 中に含まれる事象属性 ⇒ [M-Object:“半年間の”, Object:“運行休止”, Activity:“余儀なくされた”] *event*2 [3,4,5] 中に含まれる事象属性 ⇒ [Subject:“要因”, Status:“高速料金”]

事象間の関係を抽出

前工程では意味的関連のある事象属性をまとめることで，事象を抽出した．この段階で事象情報は構造化されているが，事象情報をよりリッチな情報にすべく事象間の関係抽出を試みる．

一般に実世界に起こる事象は他の大多数の事象と相互作用しあいながら生じるものであることが知られている．Radinsky らは New York Times[4]の 150 年間の過去のニュース記事を対象に事象情報を抽出し，因果関係に着目した予測木を作ることで，ある一定の条件において人間よりも高い精度で未来に起こる事象を予測できることを示した．[10]

Radinsky らの研究で事象間の関係の有効性を示したように，メディア比較・分析のタスクにおいても，事象間の関係情報が重要な情報になると考え，事象間の関係を抽出する．そこで，比較的多く観測できた事象間の関係を表す修飾句，目的語句，要因句，条件句，並列句，主語句を同定するヒューリスティックルールを導出した．

[4]http://www.nytimes.com/

前項で定義した事象属性 Activity/Status が出現する文節 ID である *e_id*，*e_id* の係り受け先文節 ID を *link_id*，*e_id* の次の文節 ID を *next_id* とする (次の文節が存在する場合，*next_id* = *e_id* + 1 が成り立つ).

・修飾句になる条件

link_id = *n_id* であり，かつ，*e_id* の末尾の形態素と対応する事象属性が Activity であり，かつ，*e_id* の品詞が “名詞” または “接頭詞” である場合，*e_id* が属する事象は *next_id* の先頭の語の修飾句である．

例文: 父が購入したバットを試合で使った。

event: (**修飾句 M-Object**:(Subject:父, Activity:購入する), Object:バット, Location:試合, Activity:使う)

・目的語句となる条件

link_id = *n_id* であり，かつ，*e_id* の末尾の形態素と対応する事象属性が Activity であり，かつ，*e_id* の品詞が “動詞” である場合，*e_id* が所属する事象は *next_id* の先頭の動詞の目的語句である．

例文: 天気が回復するかわからない。

event: (**目的語句 Object**:(Object:天気, Activity:回復する), Activity:わからない)

・要因句となる条件

link_id に Activity 属性が存在し，かつ，*e_id* の末尾の文字列が因果関係の要因を表す手がかり表現（後述）に該当する場合，*e_id* が属する事象は *link_id* が属する事象の因果関係句である．

例文: 公務を怠ったとして、減給処分になった。

event_A: (Object:公務, Activity:怠る)

event_B: (Object:減給処分, Activity:なる)a

event_A は *event_B* の要因句である．(*event_B* – *cause* → *event_A*)

・条件句となる条件

link_id に Activity 属性が存在し，かつ，*e_id* の末尾の文字列が条件句を表す手がかり表現（後述）に該当する場合，*e_id* が属する事象は *link_id* が属する事象の条件句で

ある．

例文: 目的税にした社会福祉のためにつかうのであれば国民も納得するだろう。

event_A: (M-Object:(Object:目的税, Activity:する), Object:社会福祉, Activity:つかう)

event_B: (Subject:国民, Activity:納得する)

event_A は *event_B* の条件句である．(*event_B – case* → *event_A*)

・並列句となる条件

link_id に Activity 属性が存在し，かつ，*e_id* が属する事象が因果関係句である条件にも条件句である条件にも該当しない場合，*e_id* が属する事象と *link_id* が属する事象の関係は並列である．

例文: 市民 6 人が 3 日、市長給与の返還などを求め、市に監査請求した。

event_A: (Subject:市民 6 人, Time:3 日, M-Object:市長給与, Object:返還, Activity:求める)

event_B: (Subject:市民 6 人, Time:3 日, Target:市, Activity:監査請求する)

event_A と *event_B* は並列句であり，*event_A* の後に *event_B* が起こる．(*event_A – next* → *event_B*)

・主語句となる条件

link_id に Status 属性が存在し，かつ，*e_id* の末尾の文字列が係り受け先の主語句となる手がかり表現（後述）に該当する場合，*e_id* が属する事象は *link_id* が属する事象の主語句である．

例文: 景気対策をやるのが前提だ。

event: (**主語句 Subject**:(Object:景気対策, Activity:やる), Status:前提)

最後に，事象属性に対応する文字列を主語・目的語に，事象属性を述語に対応させることで Linked Data におけるトリプル｛主語，述語，目的語｝を構成する．トリプルへの変換ルールを表 3.10 に示す．また，上記例に適用し，事象ネットワークを構築した例を図 3.4 に示す．尚，動詞等は原型に戻し，日本語シソーラス辞書 WordNet[5]を用いて可能な限り同義のノードをまとめる（名寄せする）．

[5]http://nlpwww.nict.go.jp/wn-ja/

表 3.6: 因果関係句を表す手がかり表現 17 種

手がかり表現
なり
ても
には
に伴う
おらず
生じて
上で
ため
ために
として
により
によって
が受けて
の効果が
を背景に
の影響も
の影響が

表 3.7: 条件句を表す手がかり表現 4 種

手がかり表現
あれば
場合
すれば
ければ
けれども

表 3.8: 係り受け先の主語句となる手がかり表現 3 種

手がかり表現
のが
のは
のも

表 3.9: 抽出された事象情報

Event ID	事象属性	対応文字列
329-10-1	M-Object	半年間
	Object	運行休止
	Activity	余儀なくされる
329-10-2	M-Subject	*329-10-1*
	Subject	要因
	Status	高速料金

表 3.10: トリプル変換ルール

関係名	属性	対象属性	トリプル例	組合せ数
基本属性（主題）	Subject	Activity, Status	$(Subject_{str}, Activity, Activity_{str})$, $(Subject_{str}, Status, Status_{str})$	2
基本属性（動作）	Activity	Object, Location, Time, Target, Regard	$(Activity_{str}, Object, Object_{str})$, $(Activity_{str}, Time, Time_{str})$,...	5
基本属性（状態）	Status	Location, Time, Regard	$(Status_{str}, Location, Location_{str})$, $(Status_{str}, Time, Time_{str})$,...	3
事象間のつながり(因果関係, 条件, 並列, 引用)	Event_id	Cause, Case, Next, Quoted_souce	$(Event\_id_{str}, Cause, Cause_{str})$, $(Event\_id_{str}, Quoted\_source, Quoted\_source_{str})$,...	4
データ元参照	Event_id	Subject, Time, Status, Object, Location, Time, Target, Regard, (Cause, Case)[6]	$(Activity_{str}, Activity, Event\_id_{str})$, $(Time_{str}, Time, Event\_id_{str})$,...	10

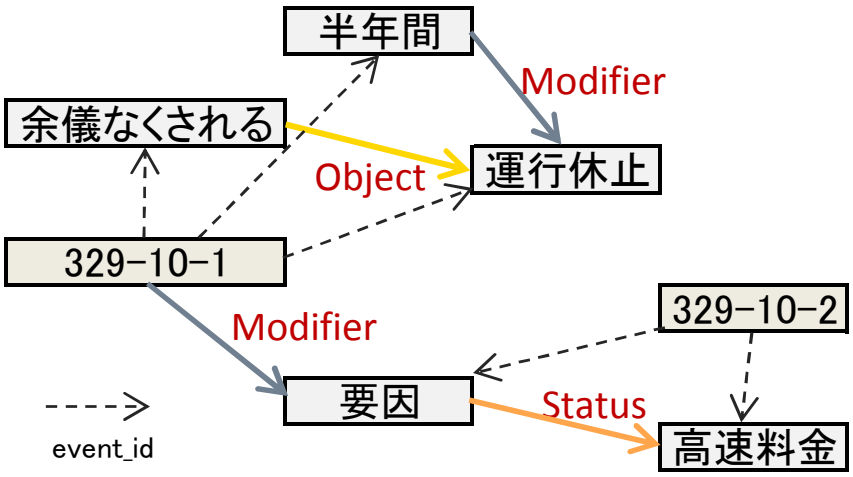

図 3.4: 事象ネットワークの例

第4章　事象抽出精度に関する評価実験

本章では，提案した事象抽出手法の有効性を検証するために評価実験を行う．以下の 2 つの精度を算出する．

1. CRF による事象ラベルの推測精度
2. ヒューリスティックルールによる CRF の結果からの事象抽出精度

4.1　CRF による事象ラベルの推測精度

訓練データとテンプレートファイルから生成された特徴モデルを用いた事象属性ラベルの推測精度を算出をするために，まず各訓練データ (Twitter データ，記事データ，Twitter データ+記事データ) を人手による事象ラベルの付与をすることで用意した．訓練データの概要を表 4.1，表 4.2 に示す．訓練データのうちの 9 割をそのまま訓練データとし，残りの 1 割をテストデータとした 10-交差検定を行い事象属性ラベルの推測精度を算出した．結果を，表 4.3，表 4.4 に示す．なお評価値 (PRECISION，RECALL，F-measure) の値は，10 回の評価値の平均値で算出している．

表 4.1: 用意した正解データの概要 (1)

訓練データ	文の数	ラベル数	modifier	subject	activity	object
ニュース記事	119	2,232	320	162	423	232
Twitter	175	2,529	254	240	489	253
ニュース記事 + Twitter	294	4,761	574	402	912	485

表 4.2: 用意した正解データの概要 (2)

訓練データ	target	status	location	time	regard	cause	quoted source
ニュース記事	12	39	82	80	37	18	14
Twitter	12	159	47	74	34	5	10
ニュース記事 + Twitter	24	198	129	154	71	23	24

表 4.3: CRF による事象属性ラベルの推測精度 (1)

訓練データ	評価値	modifier	subject	activity	object	target
ニュース記事	F値	82.90%	69.05%	93.38%	75.38%	44.44%
	Precision	80.17%	72.31%	90.77%	73.44%	50.00%
	Recall	86.40%	68.49%	96.25%	79.25%	41.67%
Twitter	F値	71.81%	55.06%	86.21%	61.93%	45.00%
	Precision	68.69%	59.22%	84.48%	61.99%	50.00%
	Recall	77.09%	54.72%	88.32%	63.53%	41.67%
ニュース記事	F値	77.41%	61.35%	89.76%	68.29%	50.89%
+ Twitter	Precision	74.91%	61.90%	88.06%	66.30%	62.50%
	Recall	80.53%	63.43%	91.78%	70.86%	46.67%

表 4.4: CRF による事象属性ラベルの推測精度 (2)

訓練データ	評価値	status	location	time	regard	cause	quoted source
ニュース記事	F値	47.67%	83.62%	84.17%	28.45%	28.00%	55.56%
	Precision	57.41%	90.36%	98.14%	39.05%	40.00%	66.67%
	Recall	48.32%	78.98%	76.60%	23.61%	25.00%	50.00%
Twitter	F値	29.48%	23.83%	49.24%	21.67%	66.67%	25.00%
	Precision	48.06%	44.44%	59.89%	30.00%	66.67%	25.00%
	Recall	23.81%	21.92%	46.80%	18.33%	66.67%	25.00%
ニュース記事	F値	32.69%	64.30%	70.33%	28.42%	12.50%	16.67%
+ Twitter	Precision	54.92%	79.29%	75.22%	52.50%	12.50%	16.67%
	Recall	24.88%	57.51%	70.15%	24.42%	12.50%	16.67%

4.2　CRF 結果から事象の抽出精度

前節では，CRF の形態素ごとの事象属性ラベルの推測精度を算出したが，本節では 3.2.4 節で説明した次の工程である CRF の結果から事象属性毎に対象文字列を抽出し，事象属性間で関係のある組み合わせで事象属性をまとめ，事象として抽出する工程の精度を算出する．

そこで，事前に用意した文から人手で作成した事象データと表 4.1，表 4.2 で作成した正解データの一部を用いて，ラベルが付与された正解データから事象群の抽出を行い精度を算出した．

表 4.5，表 4.6，表 4.7 にその結果について示す．なお用いた正解データセットは Twitter のデータセットから選定した．個々の事象属性の正解基準は，正解データで用意した文字列と完全一致するかどうかであり，事象の正解は事象属性の組み合わせにおいて，各々の事象属性毎の対応する文字列が正解データと完全一致した場合に限り正解とした．

表 4.5: CRF の出力形式から事象の抽出精度 (1)

	事象	modifier	subject	activity	object	target
抽出すべき数	79	36	32	65	54	4
抽出数	79	30	29	65	55	4
正解数	59	30	28	65	54	4
Precision	74.68%	100.00%	96.55%	100.00%	98.18%	100.00%
Recall	74.68%	83.33%	87.50%	100.00%	100.00%	100.00%
F値	74.68%	90.91%	91.80%	100.00%	99.08%	100.00%

表 4.6: CRF の出力形式から事象の抽出精度 (2)

	status	location	time	regard	cause	quoted source
抽出すべき数	6	18	15	4	8	4
抽出数	6	18	14	4	8	4
正解数	6	18	14	4	8	4
Precision	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Recall	100.00%	100.00%	93.33%	100.00%	100.00%	100.00%
F値	100.00%	100.00%	96.55%	100.00%	100.00%	100.00%

表 4.7: 事象間の関係の抽出精度

	修飾句	Next	Cause	Case	Activity の目的語句	Status の主語句	計
抽出すべき数	24	6	2	2	13	1	48
抽出数	16	6	2	2	12	1	39
正解数	16	5	2	2	12	1	38
Precision	100.00%	83.33%	100.00%	100.00%	100.00%	100.00%	97.44%
Recall	66.67%	83.33%	100.00%	100.00%	92.31%	100.00%	79.17%
F値	80.00%	83.33%	100.00%	100.00%	96.00%	100.00%	87.36%

第5章　メディア比較への適用事例

メディア比較の必要性

Facebook[1]，Google+[2]，mixi[3]などのソーシャル・ネットワーキング・サービスやTwitter[4]などのマイクロブログサービスといったソーシャルメディアの爆発的な普及に伴い，大多数の一般ユーザの声に容易にアクセスできるようになった昨今，テレビ・新聞といった従来型のマスメディアが発信する一部の情報に対して情報操作がおこなわれているのではないかなどの疑問視する声が挙がっている．一方，Twitterのようなソーシャルメディア側でも誤情報が広く伝搬して世間を混乱させてしまったりと，情報過多社会ならではの問題が起こっており，我々一般ユーザは，情報を一方的に信じるのではなくその情報に信憑性があるかどうかを都度判断する必要に迫られている．

我々が提案した手法は事象情報を含む自然言語であればドメインに寄らずに適用でき，単一メディアに限らずに事象情報を構造化された状態で獲得できる．そのため本手法を用いれば，構造化された状態で事象情報を獲得できるため，ある事象についての情報を効率的に収集することができる．複数の情報源から本手法を用いることは多角的な観点から事象を捉えることに等しく，より公正で中立的な情報理解の確立に寄与できると考えている．そこで，本章では提案手法である事象情報抽出手法をメディア比較へ適用することができると考え，いくつかの比較事例を示す．

具体的には，ソーシャルメディア・マスメディアの双方のメディア情報から提案手法である事象情報の抽出手法を適用することで，ソーシャルメディア・マスメディアの双方から得られる事象ネットワークを可視化し，見える化を行うことでメディア比較を実現する．

適当な比較事例の話題として，以下の2点の話題を選んだ．

---

[1] http://facebook.com/
[2] https://plus.google.com/
[3] http://mixi.jp/
[4] http://twitter.com/

- 輸送機オスプレイの導入問題[5] (比較対象期間: 2012/04/01 から 2012/08/16 まで)
- 第 46 回衆議院議員総選挙・2012 年東京都知事選挙 (比較対象期間: 2012/12/04 から 2012/12/19 まで)

5.1　データセットについて

本節では事象ネットワークの情報源となるデータセットについて述べる．ソーシャルメディア・マスメディアの対象メディア及びデータセットの収集方法について説明する．

対象メディア・データセット構築方法

ソーシャルメディアの対象メディアとして，マイクロブログサービス最大手の Twitter [4] を選定し，投稿されるツイートをソーシャルメディアのデータセット対象とし，一定期間 Streaming API[6]にて取得したツイートからユーザリストを作成し，そのユーザリストから無作為に Twitter のユーザ ID を取り出し，順次 REST API[7]にてユーザの過去ツイートを遡れる限り[8]取得することでデータセットを構築した．

一方，マスメディアの対象メディアとして，全国紙と称される 5 紙を運営する新聞社 5 社及び，キー局と称されるテレビ局 5 社に日本放送協会 (NHK) を加えた 6 社が運営するニュースサイトに投稿されたニュース記事を対象にした．5.1 にマスメディアのデータセット対象にしたニュースメディア一覧を示す．各ニュースメディアの最新記事の一覧ページもしくは，サイトの検索機能を用いて記事本文の HTML まで辿り，各サイトごとに作成した記事内容を抽出するスクレイピングプログラムを作成し，記事内容を取得することでマスメディアのデータセットを構築した．

---

[5]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．

[6]GET statuses/sample:
https://dev.twitter.com/docs/streaming-apis

[7]GET statuses/user_timeline:
https://dev.twitter.com/docs/api/1/get/statuses/user_timeline

[8]2013/01/24 時点では，過去 3200 ツイートまで遡って取得可能．

表 5.1: マスメディアのデータセット対象としたニュースメディア一覧

ニュースメディア	運営元	URL
YOMIURI ONLINE	読売新聞	http://www.yomiuri.co.jp
朝日新聞デジタル	朝日新聞	http://www.asahi.com
毎日 JP	毎日新聞	http://mainichi.jp
日本経済新聞 電子版	日本経済新聞	http://www.nikkei.com
MSN 産経ニュース	産経新聞	http://sankei.jp.msn.com
NHK NEWSWEB	NHK	http://www3.nhk.or.jp/news/
日テレ NEWS24	日本テレビ	http://www.news24.jp
テレ朝 news	朝日テレビ	http://www.tv-asahi.co.jp/ann/
TBS News i	TBS	http://news.tbs.co.jp
TXN NEWS	テレビ東京	http://www.tv-tokyo.co.jp/biz/
FNN	フジテレビ	http://www.fnn-news.com

話題:「オスプレイ」に関するデータセット

話題「オスプレイ」に関するソーシャル・マスメディアのデータセット概要を表 5.2，5.3 に示す.

なおマスメディアのデータセットに関しては記事収集時に各ニュースサイトの検索機能を用いて“オスプレイ”というキーワードに該当した記事ページを取得を試みたが，取得した時期の関係で古い記事が削除されていることがあり，一部取得した記事数にばらつきが生じたため，その中でも比較的多くの記事が取得できたニュースサイト 4 件 (MSN 産経ニュース，朝日新聞デジタル，日テレ NEWS24，FNN) を対象マスメディアとした．なお公平な比較をするために，データセットを構成するツイート及び記事は 2012/04/01 から 2012/08/16 までの期間に投稿されたものに限定し，データセットの対象期間を各比較対象のデータセット間で統一した.

表 5.2: 話題:「オスプレイ」に関するデータセット概要 (ソーシャルメディア)

期間	ツイート数	ユーザ数
2012 04/01-08/16	12,097,617	11,880

表 5.3: 話題:「オスプレイ」に関するデータセット概要 (マスメディア)

期間	MSN 産経ニュース	朝日新聞デジタル	日テレ NEWS24	FNN	計
2012 04/01–08/16	231	116	110	78	**535**

話題:「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

話題「第 46 回衆議院議員総選挙・2012 年東京都知事選挙」に関するソーシャル・マスメディアのデータセット概要を表 5.4，5.5 に示す．なおオスプレイのデータセットと同様にデータセットの対象期間は衆院選の開示日の 2012/12/04 から執行日の 2012/12/16 までとデータセット間で同一にした．また一記事あたりの分量が多く，情報量の多い事象情報が得られるという観点からマスメディアの対象メディアは表 5.1 中の新聞社が運営するニュースサイトに絞り，各サイトに用意されている「政治」カテゴリーで上記期間で投稿された記事を選挙に関する話題を扱う記事として取得した．

表 5.4: 話題:「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概要 (ソーシャルメディア)

期間	ツイート数	ユーザ数
2012 12/04–12/09	1,105,894	15,394
2012 12/10–12/16	1,412,024	15,507
2012 12/17–12/19	526,028	13,860
計	3,043,946	16,293

表 5.5: 話題:「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概要 (マスメディア)

期間	日本経済新聞電子版	毎日JP	朝日新聞デジタル	MSN 産経ニュース	YOMIURI ONLINE	計
2012 12/04–12/09	2	78	0	0	0	**80**
2012 12/10–12/16	262	169	120	51	27	**629**
2012 12/17–12/19	178	108	101	77	47	**511**
計	442	355	221	128	74	**1,220**

5.2　特定キーワードによるフィルタリング

Twitter のデータセットはほぼランダムに構築されている．比較する話題に関するツイートに絞ることでネットワーク比較を効率化する．ヒューリスティックなキーワード辞書を

話題毎に作成し，事象ネットワークを構築する前にデータセットを絞込みを行った．各話題毎のデータセットの通過率を表 5.6 に示す.

表 5.6: 各話題に関する絞り込みをするために関するデータセットに対して行った特定キーワードによるフィルタリング結果

話題	期間	フィルターを通過したツイート数	通過率
オスプレイ	2012 04/01-08/16	3,084	0.0255%
衆院選 都知事選	2012 12/04-12/09	30,692	2.78%
	2012 12/10-12/16	80,880	5.73%
	2012 12/17-12/19	27,818	5.29%
	計(2012 12/04-12/09)	139,390	4.58%

なお使用した特定キーワードについては付録の表 A.1，A.2 を参照されたい.

5.3　ネットワークの可視化

5.3.1　可視化ツール

ネットワークを可視化するにあたって，2 種類の可視化ツールを用いた.

rdf-gravity[9]　RDF ファイルを読み込むことができる可視化ツール.

Gephi[10]　ネットワーク描画に特化したオープンソースウェア．RDF ファイルをそのまま読みことはできないが，本研究においては Gexf 形式[11]への変換を行なうことでネットワークを描画した．Gexf 形式にはネットワークのレイアウトからノード・エッジ大きさ・色を指定することが可能であるため，効果的な可視化を実現するためにノード・エッジの色及びノードの大きさ，エッジの太さを指定して描画した.

話題「オスプレイ」の比較においては，両者の可視化ツールを用いてメディア比較を行ったが，プラグインを柔軟に追加できることと可視化性能の優劣から話題「第 46 回衆議院議員総選挙・2012 年東京都知事選挙」の比較時においては Gephi のみを用いた.

[11] http://gexf.net/format/

5.3.2　ノードの大きさ/エッジの太さ

Gephi を利用した可視化時に，ノードの大きさ及びエッジの太さを以下のように定義した．

ノードの大きさ

ノード i が参照しているノード数を $reffer_i$，ノード i を参照しているノード数: $refferd_i$ とすると，ノード i の大きさを $Size\_of\_Node_i$ は以下のように算出している．なお対数の底 a は 1.5 とした．

$$Size\_of\_Node_i = \log_a(reffer_i \times 0.2 + reffered_i \times 0.8)$$

エッジの太さ

ノード i からノード j への関係 r の有効エッジの頻度を $Frequency\_of\_Edge_{ijr}$ とすると，ノード i からノード j への関係 r の有効エッジの太さ $Weight\_of\_Edge_{ijr}$ は以下のように算出している．なお対数の底 a は 1.5 とした．

$$Weight\_of\_Edge_{ijr} = \log_a(Frequency\_of\_Edge_{ijr})$$

ノードの大きさ及びエッジの太さを決める際に対数を用いたのは，“単純に参照されている/この種類にエッジが複数あるという数値”と “大きさ/太さが比例する”場合では，大きく/太くなりすぎてしまうと考えたからである．

話題「オスプレイ」ノードの色およびエッジの色

話題「オスプレイ」に関するネットワーク可視化時に用いたノードの色及びエッジの色を表 5.7，表 5.8 に示す．

ノードの色およびエッジの色

話題「選挙」に関するネットワーク可視化時に用いたノードの色及びエッジの色を表 5.9，表 5.10 に示す．

表 5.7: 話題「オスプレイ」のメディア比較時のノードの色 一覧

ノードのメディア間の構成比率 (マスメディアの構成割合)	色
0%	
37.5 % 未満	
62.5 % 未満	
100 % 未満	
100%	

表 5.8: 話題「オスプレイ」のメディア比較時のエッジの色 一覧

関係	色
Activity	
Object	
Time	
Location	
Cause	
Target	
Status	
Quoted source	
Event_id	

表 5.9: 話題「選挙」のメディア比較時のノードの色 一覧

メディア	色
ソーシャルメディアでのみ出現するノード	
共通ノード	
マスメディアでのみ出現するノード	

表 5.10: 話題「選挙」のメディア比較時のエッジの色 一覧

関係	色
Subject	
Activity	
Object	
Time	
Location	
Cause	
Next	
Target	
Status	
Quoted source	
Regard	
Modifier	
Case	
Event_id	

5.4　ネットワーク考察:「オスプレイ」

ネットワーク概要

可視化を行った様子を図 5.1 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，なお「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 (もしくはされたことが) 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，表 5.7 で記述した通り，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，“琉球新報”ノードが紫色であるように，沖縄の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ノードの色から読み取れる．表 5.11 に抽出した事象数などの情報を示す.

表 5.11: 「オスプレイ」に関する事象ネットワークの概要

メディア	ノード数	得られた事象数	ラベル計	Subject	Activity	Object	Target	Status	Time	Location	Cause	Quoted source
Twitter	4,218	6,846	15,703	2582	6129	4636	16	1130	520	548	13	129
MSN産経ニュース	2,134	3,790	9,669	1258	3733	3277	174	91	554	565	16	1
朝日新聞デジタル	1,339	1,920	5,003	706	1894	1650	102	48	255	332	14	2
日テレ NEWS24	294	409	1,198	145	408	340	12	1	139	149	3	1
FNN	917	1,712	4,148	564	1666	1387	71	10	149	292	9	0

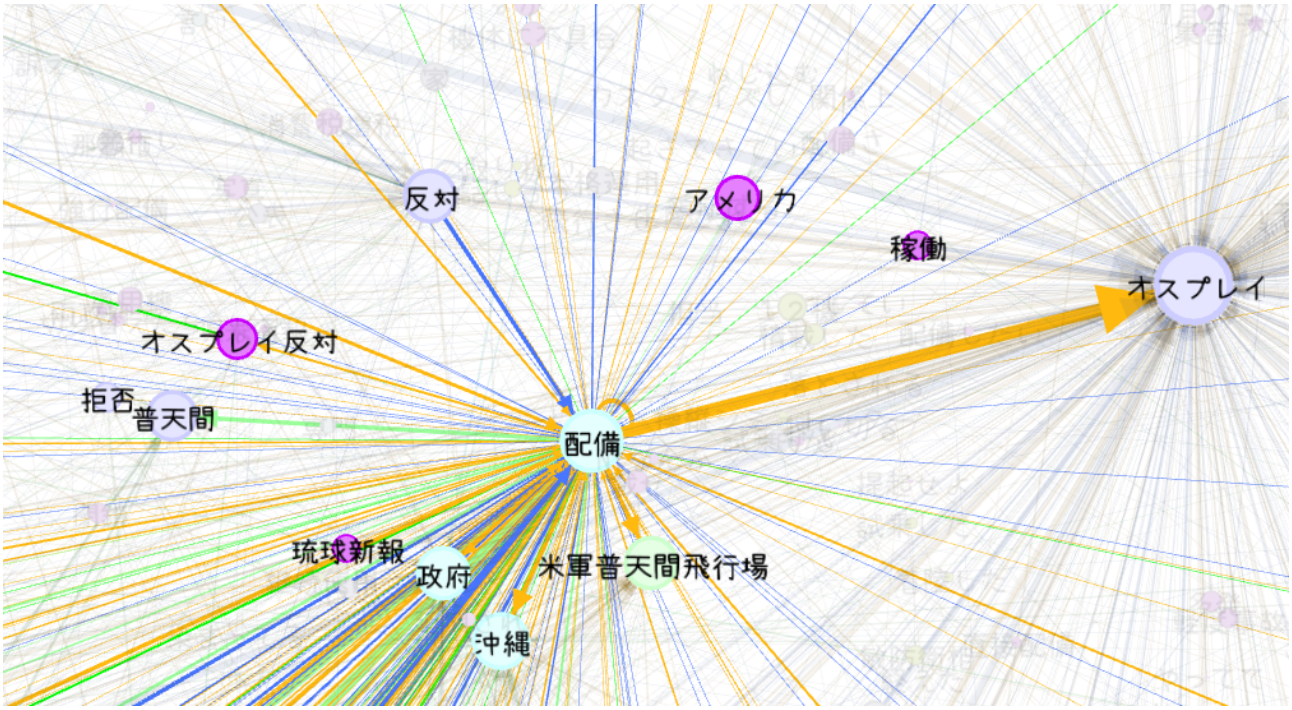

図 5.1: オスプレイに関する事象ネットワーク (「配備」ノードに着目)

① 話題の多様性

表 5.11 の各ネットワークを構成するノード数から及び，各メディアの事象ネットワーク図 5.4(Twitter)，図 5.4(新聞紙)，図 5.4(FNN) から空間に対してのノード数の密度の違いが読み取れる．図 5.5 は Twitter と新聞紙メディアの両者の事象ネットワークを可視化した様子の外観である．ノードの色の定義を示した表 5.7 で記述した通り，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表している．その対極にある黄緑色のノードは新聞紙メディアでのみ出現する語を表している．そのことを踏まえて改めてこのネットワークを見てみると，明らかに紫色のノードが多いことが読み取れる．これらは，マスメディア側では各イベントや事件に対して考え方・意見の統一化をしないといけないのに対し，ソーシャルメディア側では，コンテンツは多くの一般ユーザの多種多様な意見が含まれた投稿から構成されているため，その話題/意見/表現の多様性がノード数の多さに反映されていることがわかる．

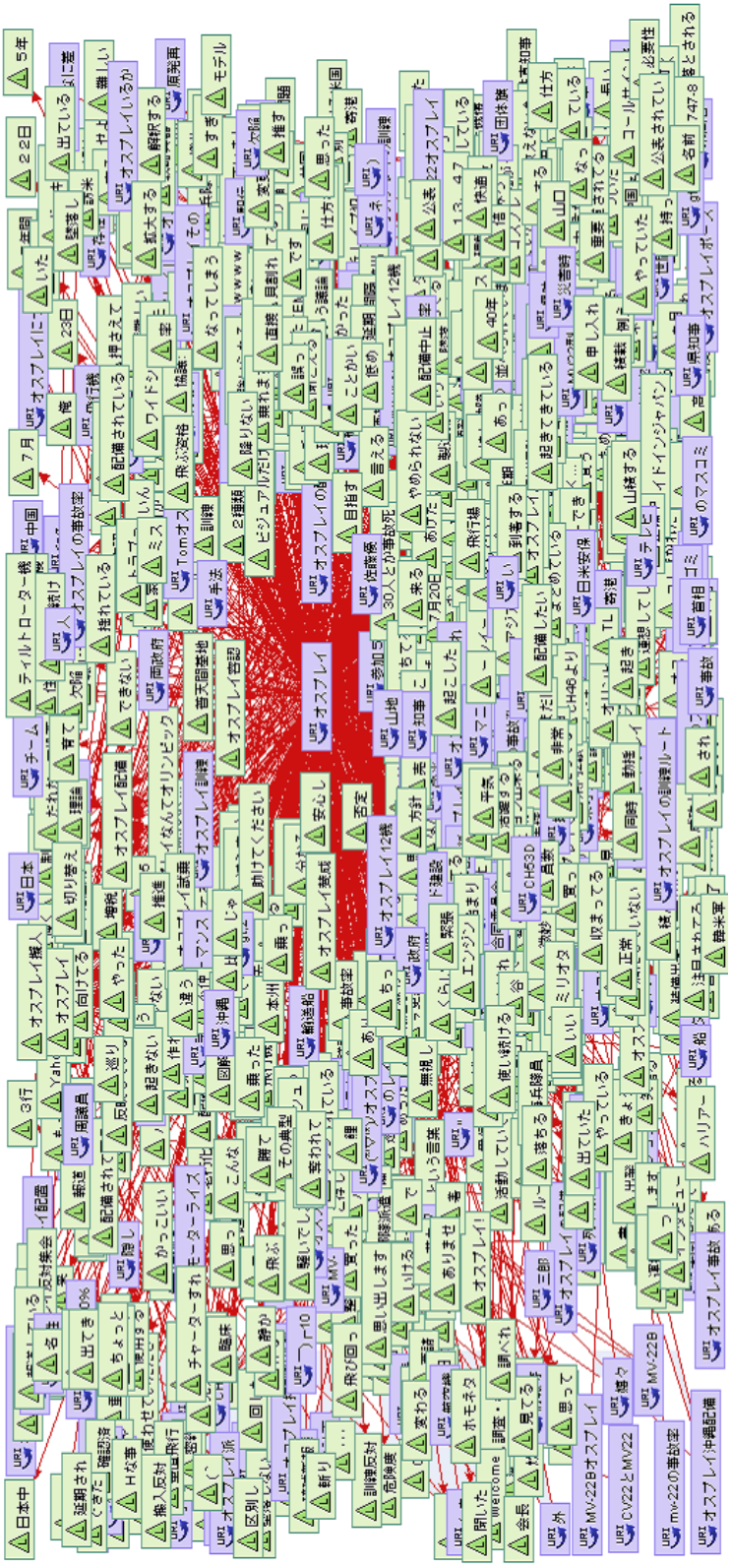

図 5.2: ソーシャルメディア (Twitter)から構築したオスプレイに関する事象ネットワーク

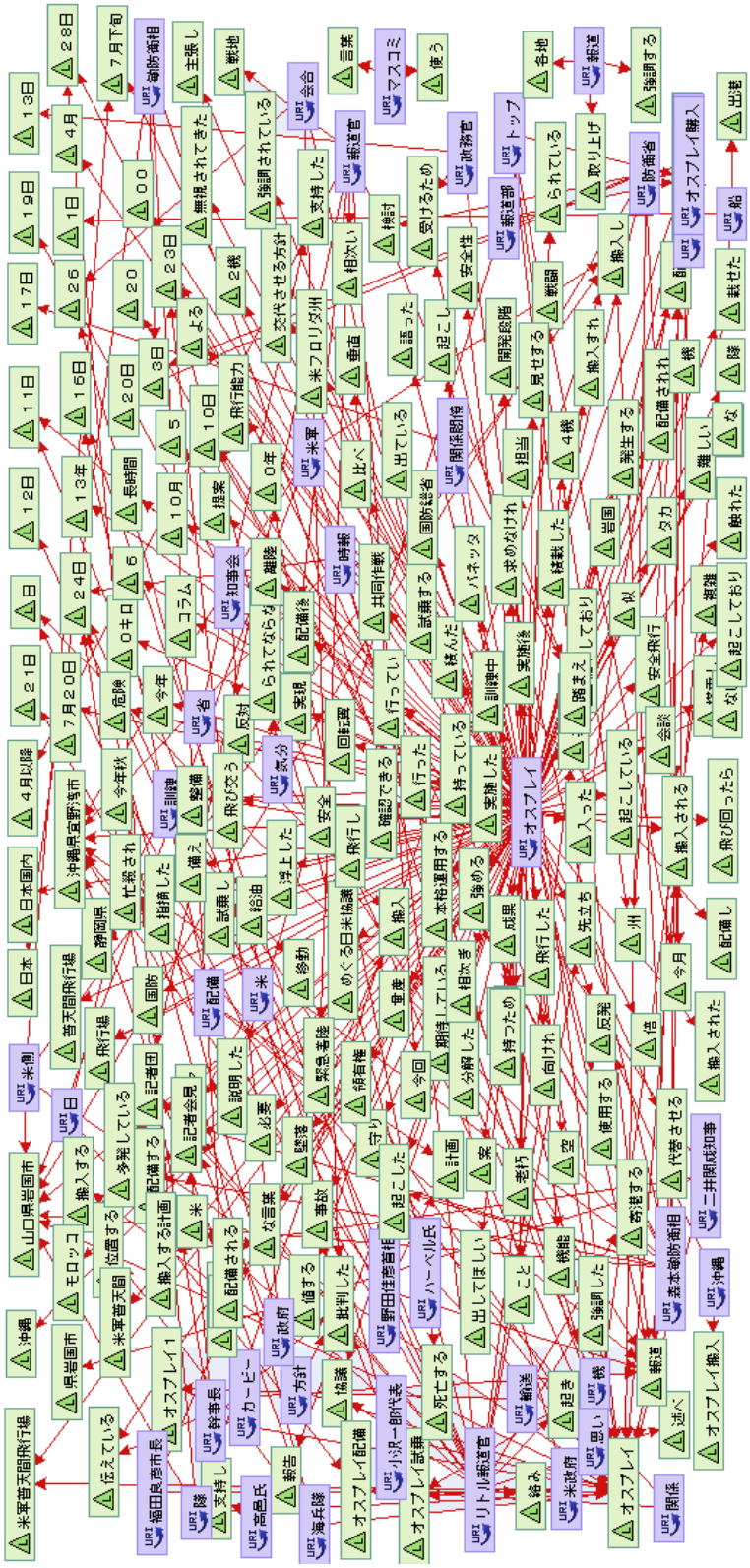

図 5.3: 新聞社のニュースメディア (産経ニュース) から構築したオスプレイに関する事象ネットワーク

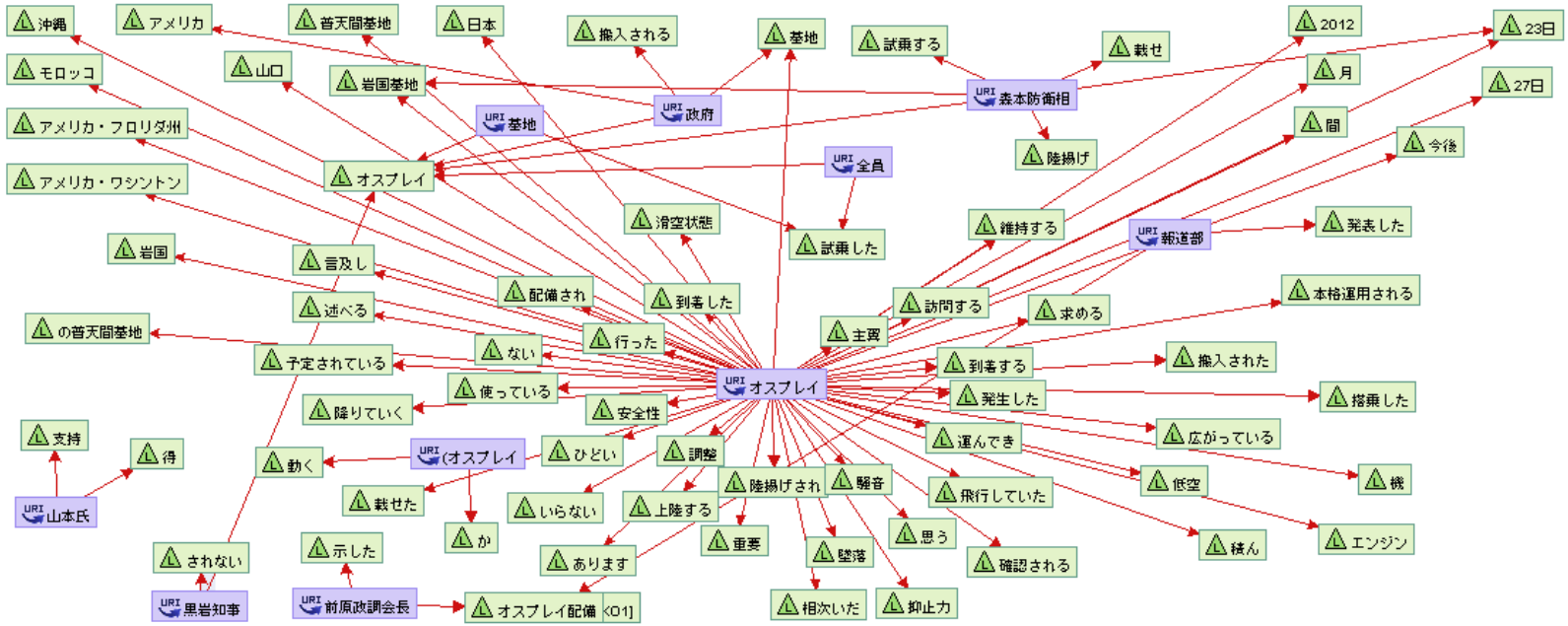

図 5.4: TV 局のニュースメディア (FNN) から構築したオスプレイに関する事象ネットワーク

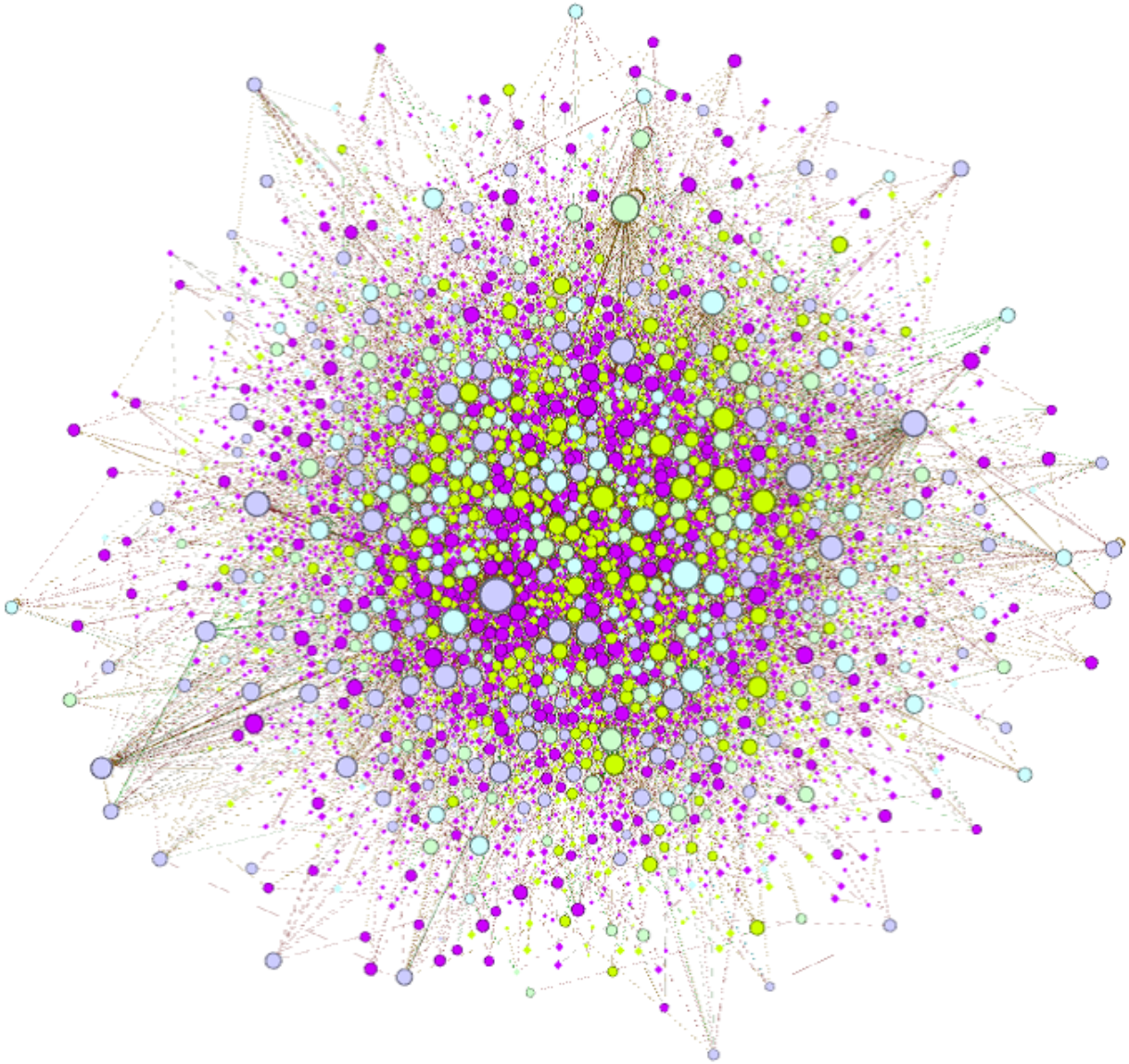

図 5.5: オスプレイに関する事象ネットワークの外観 (紫色のノードがソーシャルメディアでのみで出現する語を表し，黄緑色がマスメディアでのみで出現する語を表す．)

② 少数派意見へのアプローチ - 「賛成意見」

一般的に頻度は意思決定の重要な判断材料になりうるが，時に少数の意見の中に物事の本質が含まれていて，頻度分析の類いでは，これらのマイナーな意見はノイズとして判断し情報を取りこぼしてしまうことがある．しかし，本手法は割合としては少ないものにも

重要な情報が含まれている可能性があると考え，マイナーな意見が最終的なネットワークに現れるように配慮している．

5.6 は Twitter 側のデータセットから構築した事象ネットワークの一部を切り取ったものである．この図は，マスメディア側にはないオスプレイ配置に対するマイナーな意見 (賛成意見) が Twitter 側のネットワークに反映されていることを示している．

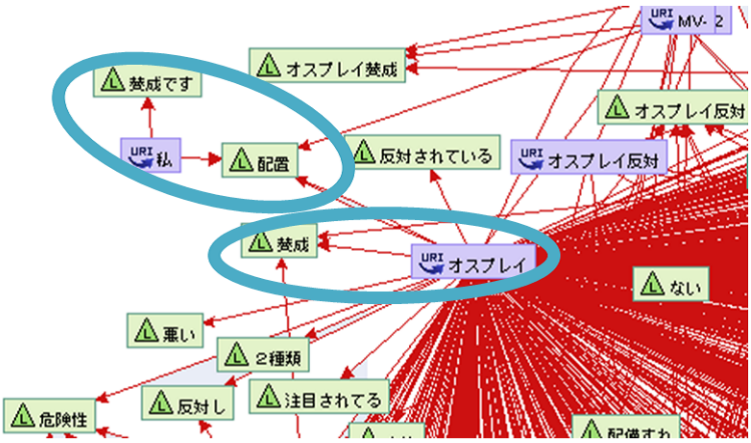

図 5.6: オスプレイに関するサブネットワーク (賛成意見)

③ マスメディアにはない情報 - 「2 種のオスプレイの存在」

図 5.7 は Twitter 側のデータセットから構築した事象ネットワークの一部を切り取ったものである．この図に着目するとマスメディア側のネットワークにはない話題,「MV-22」,「CV-22」の話題がソーシャルメディア側の事象ネットワークに顕著に現れているのが伺える．これらの文字列は，オスプレイの機種を表していて，MV-22(正確には MV-22B) が主に輸送を目的とした機種で事故率は 1.93 であり，海兵隊所属のヘリを含む航空機の平均事故率 2.45 より低い．これに対し，CV-22(正確には CV-22B) は特殊作戦を目的とした機種で事故率は 13.47 と高い．両者オスプレイと言う名称は同じではあるが，用途・事故率の観点からみてまったく違う機体であるのは明らかである．ちなみに日本に来るのは輸送型の MV-22 であり，約 40 年前から使われている現行の CH-46 ヘリコプターの事故率 2.43 より低く，事故率の観点で考えると現状より安全になるのに関わらず，マスコミでの報道 (主に TV 局) では，両機種「MV-22」,「CV-22」を同一に扱い，事故率が高い「CV-22」の事故

率を持ち出し，事故率の観点で頻繁に議論を展開し「オスプレイは危険である」という印象を植え付けているかのように思われる．また，Gephi を用いた可視化したネットワークにも事故率と型番の関係がソーシャルメディアの情報のみに確認できた．図 5.8 では 2 種の型番の違う機種の存在が明らかに読み取れ，図 5.9，図 5.10 においては，オスプレイの事故率が複数存在することを示唆し，日本に配備された MV-22 の事故率が低いことも明示されており,「事故率」ノード付近のネットワークから読み取ることができる．2 種の型番の存在，日本に配備される機種が事故率が低いことの情報もいづれもソーシャルメディアからの情報であった．本研究で構築した事象ネットワークを見ることで，Twitter 側の事象ネットワークからそれぞれの機種「MV-22」,「CV-22」，現行ヘリ「CH-46」を表すノードと事故率を表すノードが Status 属性でリンクされており,「MV-22」,「CV-22」関連のノードの頻出などの理由から，本システムで構築した事象ネットワークがマスコミの印象操作の疑いに気付くまでに達しなくとも，頻出する各々の機種に対して興味を持ち，調べるのに十分な動機を与える結果になっていると考えている．

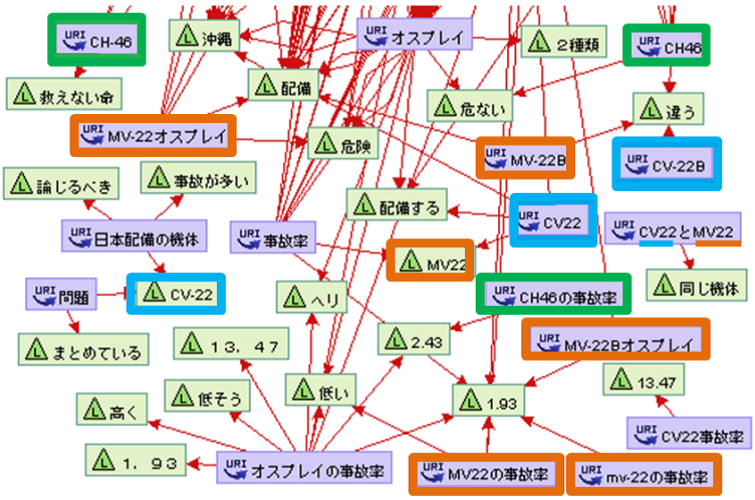

図 5.7: オスプレイに関するサブネットワーク (「MV-22」,「CV-22」)

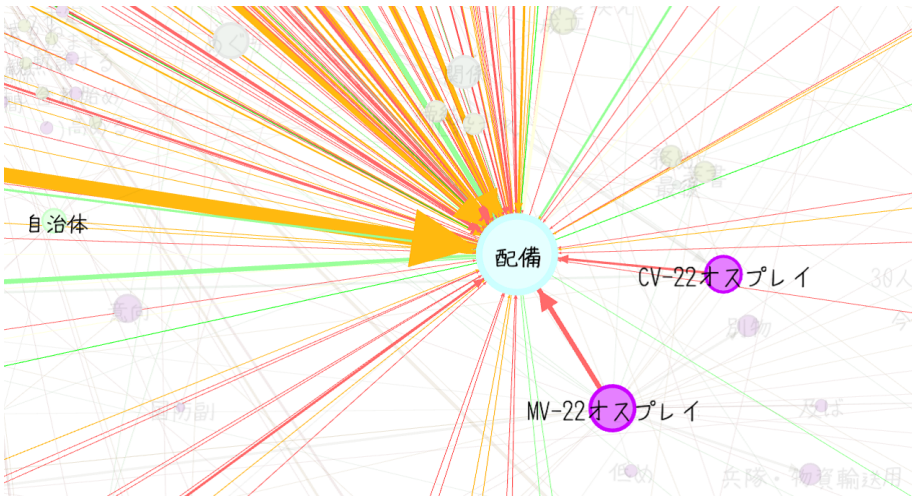

図 5.8: 「配備」ノードを Activity 属性で参照する「MV-22 オスプレイ」,「CV-22 オスプレイ」ノード (型番を示すノードがソーシャルメディアでしか出現しないことを示している．)

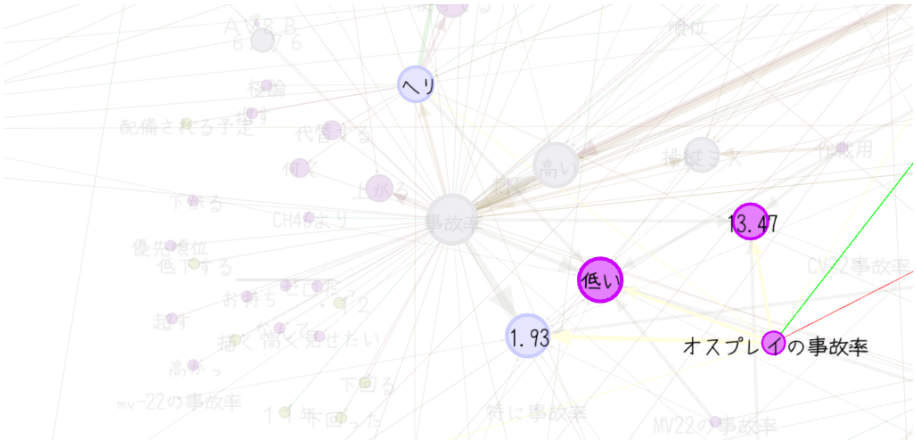

図 5.9: “オスプレイの事故率”が 2 種類 (“1.93”，“13.47”) 存在することを示唆するネットワーク.

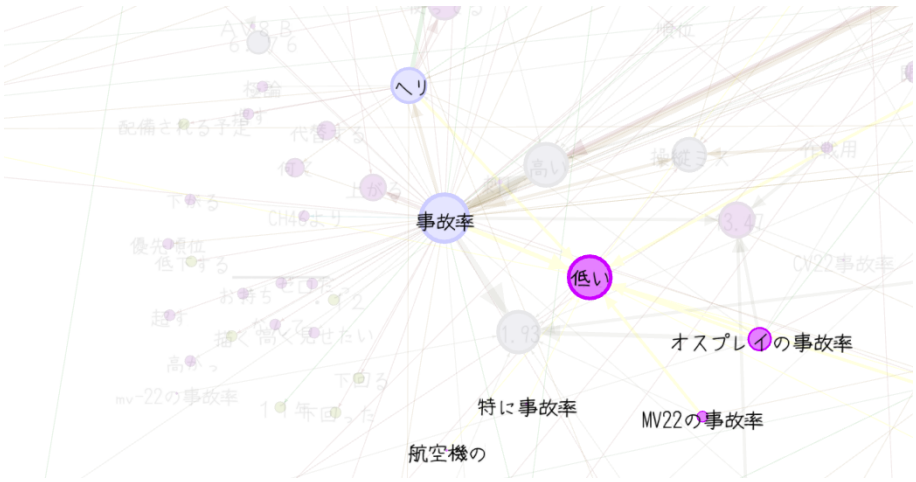

図 5.10: “オスプレイの事故率 (MV22)”がソーシャルメディアで低いと言及されていることを示唆するネットワーク

④ 偏在性に関して (今後の課題)

マスメディアは名の通り大衆に向けた粒度の大きい情報を扱うのに対し，ソーシャルメディアではユーザ毎に情報を発信する目的・対象が多様であることから粒度の小さい情報が得られやすいという特徴がある．その典型的なものとして，地域毎の情報が挙げられる．オスプレイの事例では，配備後の影響度合いの観点で人々の問題に対する関心度が人々の住まいと配備予定地までの距離と強い相関があると考えられる．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 (いわば偏在性) に対しての理解を容易にすることにつながると考えられる．しかし，今回フィルタリングを通過した 3,084 件のツイートの中に位置情報を含むツイートは 5 件と少なく，今回のネットワーク比較によって偏在性のある情報を確認することはできなかった．今後偏在性のある情報をネットワークに反映させるために，データセットの規模を大きくした上での再実験を検討している．例えば，関東地域に住む人はオスプレイの導入問題に関して無関心の人が多い傾向がある．しかし当事者である沖縄地域に深く関心を頂いていると想定できる．他の話題・出来事についても，地理別に物事の影響範囲があると考えられユーザの位置情報を考慮に入れた，事象情報ネットワークを作ることができれば，より深い分析が可能になり価値のある情報を得ることができるようになると考えている．偏在性を考慮できた事象ネッ

トワークをマーケティング分野への適用をした場合，より効果的なマーケティングをおこなうための本ネットワークが重要な材料となることは疑いの余地のないことだろう．

⑤ 因果関係 (Cause) に関して (今後の課題)

話題「オスプレイ」に関する事象ネットワーク構築時には，3.2.4 節で説明した事象間の関係抽出アルゴリズムが未完だったためを適用していない．そのため話題「オスプレイ」に関する事象ネットワークには Cause 関係数が一桁台もしくは，10 数個どまりであった．もし事象間の因果関係を多く抽出することができれば，ある現象が起こる真の根本原因を Cause 関係を辿ることで見つけることができる．つまり，構築された事象ネットワークから現象に対しての原因が究明することができるので，本事象ネットワークを用いれば，実世界の問題に対して適切な解放 (有用な知見) を得ることができると考えている．

5.5　ネットワーク考察:「第 46 回衆議院議員総選挙・2012 年東京都知事選挙」

比較期間を以下の 3 つに分割した．選挙戦前半 (公示から週末まで 12/04-12/09)，選挙戦後半 (2 週目から選挙終了日まで 12/10-12/16)，選挙終了後の 3 日間 (12/17-19) それぞれ期間 A，B，C と呼ぶ．また，期間毎に事象ネットワークを作るのだが，ソーシャルメディアとマスメディアのデータ量にかなり差がある (表 5.12 参照)，公正な比較をするために，同じくらいの規模にする必要があると考え，ソーシャルメディアのデータ量を調整した上で比較時に用いる事象ネットワークを構築した．期間ごとのネットワークの概要が表 5.14 である．3 種類のネットワークをネットワーク A，ネットワーク B，ネットワーク C と呼ぶ．

表 5.12: 期間毎の「選挙」に関するデータセットのトリプル数

期間	トリプル数	
	ソーシャル	マス
A: 12/04−12/09	463,228	9,843
B: 12/10−12/16	1,078,772	49,109
C: 12/17−12/19	392,658	46,106

なお表 5.14 中の “構成比”とは，ノードは 3 種類に分類できソーシャルのみに出現する語を表すノード or マスメディアにのみ出現する語を表すノード or 両メディアで出現する語

表 5.13: 期間毎の「選挙」に関するネットワークの概要

期間	トリプル数		ノード数						
	ソーシャル	マス	ソーシャルのみ	構成比	マスのみ	構成比	共通ノード数	構成比	ノード数計
ネットワーク A (12/04–12/9)	9,860	9,843	3,591	53.8%	2,522	37.8%	567	8.5%	6,680
ネットワーク B (12/10–12/16)	51,373	49,109	18,522	55.8%	11,502	34.7%	3,148	9.5%	33,172
ネットワーク C (12/17–12/19)	43,630	46,106	48,903	66.1%	18,544	25.1%	6,557	8.9%	74,004

という 3 種類．ネットワークを構成する全ノードのうちソーシャルのみに出現するノードは何%の割合で存在するかを表すものである．

①話題の多様性

「オスプレイ」のネットワークと同様に「選挙」関するネットワークでもソーシャルネットワークならではの話題・表現の多様性が確認できた．

- 前述の表 5.14 の構成比に着目すると，ソーシャルメディアでのみ出現する語が 3 つのネットワーク全てにおいてマスメディアより多くかった．その傾向色濃く見られたのは選挙終了後の期間で，ソーシャルのみに出現する語が 66.1%となっていた．．これは衆院選という大きなトレンドが終了し，選挙期間中ユーザの興味が衆院選に集中していたものが，選挙終了という合図で元の興味範囲に戻っていき，選挙期間中に通常時より絞られていた論点が拡散したことで，このような急激にノード構成比が上昇したのではないかと考えている．

- 図 5.11 はネットワーク A の外観を表している．ソーシャルメディアだけで出現する語の青いノードが外側に散在していることがわかり，この図からも多様性の観点が読み取れる．

- 図 5.11 のノードの大きさに着目すると，赤いマスメディアのノードの方が青いノードよりも大きいノードが多いことが読み取れる．これはツイッターに投稿される文章が短く文法の正しさが約束されていないという特徴から，マスメディアの正しく比較的長い文章であることに比べてまとまった事象が得られにくいために，各ノードが参照もしくは参照されることが絶対数的に少なくなっていることもノードの大きさから読み取れる．その半面，マスメディアから得られる事象情報は図 5.12 のようなまとまった情報を得ることができる．ークの外観だけでなくネットワーク内部に

目を向けると，図 5.13 のように，新安倍政権が発足した後に，“拉致被害者の再調査に応じたこと”，“様々なことに言及していること”などの一般事象についての情報を得ることができる．一方で同事象ネットワーク内の別の場所では，図 5.14 のような“２ちゃんねる”[12]がサービスがダウンしている状況にあることがこのネットワークから読み取れるだろう．この情報はソーシャルネットワークから得たものである．複数の情報源からネットワークを構築しているために得られる情報のバラエティが豊かになっている．

5.6　比較観点に基づく注目ポイントの抽出

本節では，前節の予備実験（および，これまでネットワーク構築にあたった経験）から，前節で挙げた多様性，希少性，偏在性，因果関係の 4 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 4 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．次章でも述べるように，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 5.14 にネットワークのノード構成を示す．

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．

[12] http://www.2ch.net/

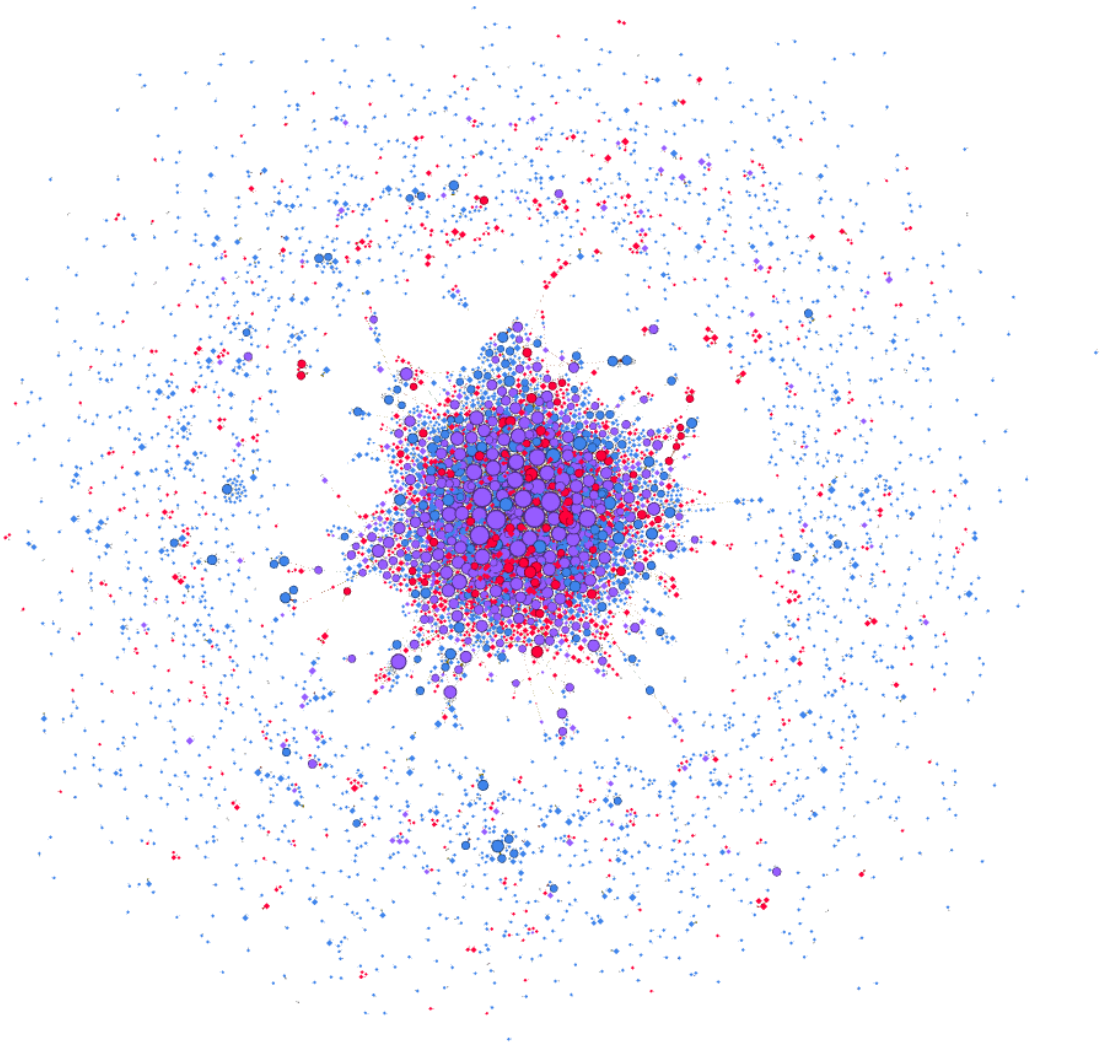

図 5.11: 「選挙」に関する事象ネットワーク A の外観

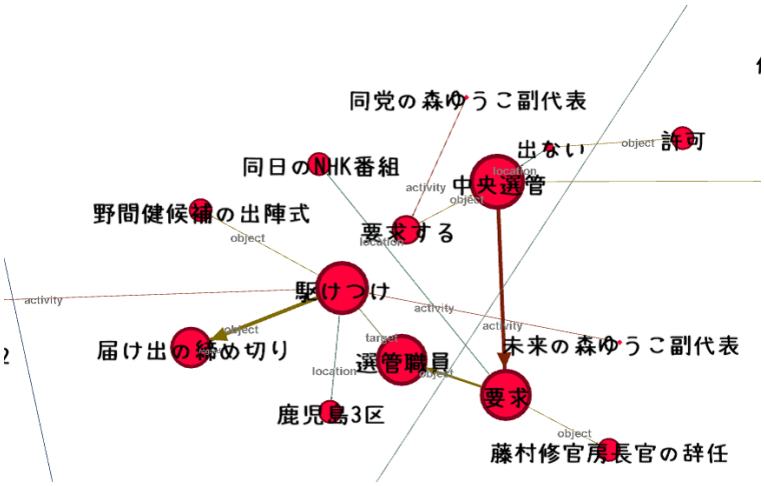

図 5.12: 「選挙」に関する事象ネットワーク A のマスメディアデータから得たまとまった事象情報を表すネットワーク

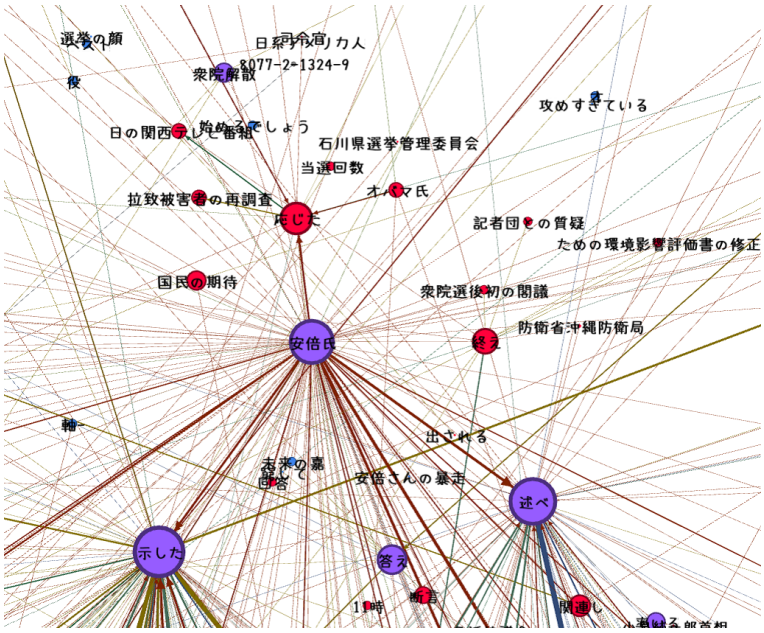

図 5.13: 「選挙」に関する事象ネットワーク B の一般事象へのアクセス

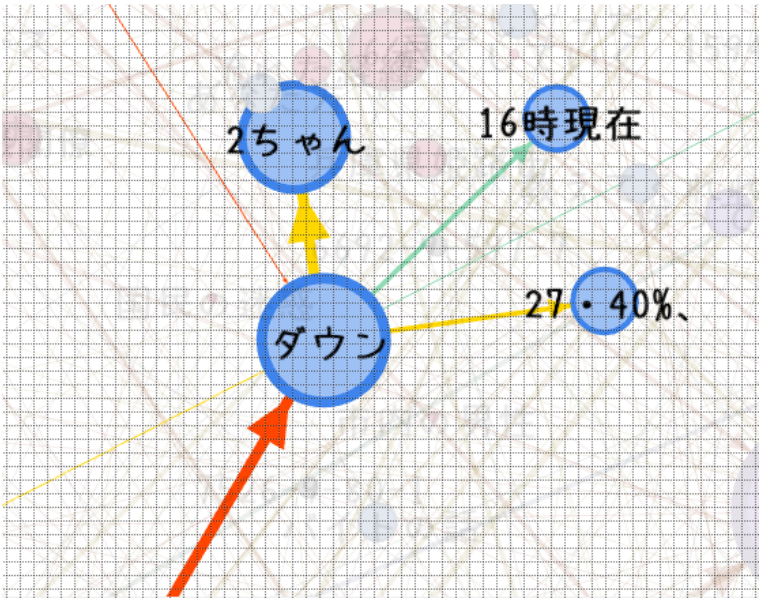

図 5.14: 「選挙」に関する事象ネットワーク B のマイナー情報へのアクセス

表 5.14: 総選挙に関する事象ネットワークに関するデータ

Network information for Election	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

多様性

ソーシャル，マス各データセットを対象に，両データセット共通ノードの中からエッジの参照元，参照先という頻度の観点で頻出するノードを絞り込み，そのノードに代表的なエッジとして Activity/Status 属性で繋がる非共通ノードを検索する．検索された非共通ノード数が多いノードであるほど，多様な行動・状態・意見が確認できる注目すべきノードとみなせる．各注目ノードの周辺トリプルを共通行動/状態/意見，非共通行動/状態/意見（ソーシャルのみに出現，マスのみに出現）の組みとして抽出する．注目ノードと共にこれらの情報を提示することで整理された情報をユーザに提示できると考えている．

例：注目ノード:「自民党」，マスメディア: 行動 155 種 e.g. 大勝する (頻度:3)，離党する (頻度:1)，... / 状態 10 種 / 意見 13 種，ソーシャルメディア: 行動 2556 種 e.g. 期待されてる (頻度:84)，投票する (頻度:60)，... / 状態 463 種 / 意見 379 種 e.g. ダメ (頻度:13)，古い (頻度:2)

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）.

例：(Subject:(Subject:公明党，Activity:全勝，Location:小選挙区)，Status:凄い)–*Cause* →(M-Subject:公明党，Subject:死に票，Status:0%)

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.

事象例：(Location:福島，M-Object:全投票所，Object:投票時間，Activity:繰り上げる)

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.

事象例: (M-Subject:自民党， Subject:安倍総裁， Activity:遊説)← *Cause*– (Subject:安倍氏，Location: 静岡県内，Activity: 移動)–*Next* → (M-Subject:初老， Subject:男性， Target:安倍氏， Activity:注意)–*Cause* → (M-Subject:(Time:後から， Activity:乗る)， Subject:安倍氏， M-Object:(Subject:JR 職員， Activity:抑える)， Activity:座る， Object:席)

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．多様性と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 5.15 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 5.15: 注目ポイントの評価
Evaluation of comparable points

	多様性	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 (%)	65.33%	58.33%	55.08%	51.16%

注目ポイントの評価結果から 4 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 4 つの比較観点が一定の有効性のある手が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が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．

第6章　関連研究

本研究と類似した目的から，Web 空間もしくは Twitter から単語間の関連語，話題抽出を行った研究 [11, 12, 13] がある．これらの研究は，時系列変化における単語の頻度情報，単語間の共起頻度などに着目することで，関連語及び話題抽出が行えることを示した，これらの研究と本研究との相違点としては，上記研究で行われていた単語と関連語という表面的な関係を抽出するのとは違い，本研究では複数ノード (キーワード) と複数のエッジから事象関係をネットワークで表現している．これにより，点 (単語) 同士を関連度でつなげたネットワークよりも，関係間の意味を表現できるようになり (いわば面で関係を表せるようになり) より深い分析や事象のノードをたどることで事象が起きている背景・本質に迫ることができると考えている．

単語・関連語ではなく語句レベルで語句間の関係を抽出し，効果的な可視化を行うシステムを構築した研究として，伊藤らの研究 [14] がある．伊藤らはブログから個人の行動・興味とそれらの対象を係り受け解析を用いて事象として抽出し，入力キーワードに対して関連する事象群の月ごとの変化及び各事象の頻度変化を探索可能にする 3 次元可視化システムを構築した．本研究と比較すると，伊藤らは事象群と文章構造をキーワードを中心としたツリー表現を用いて可視化を行なっているが，事象を構成する各ノード (キーワード) 同士をつなぐエッジの種類までは定義しておらず，各ノード間の関係を判断するのに人が目視で可視化された情報を確認し推測する必要がある．また特定事象に予め固定して詳細に分析することには長けているが，複数の事象を同時に比較するためには複数のキーワードを予め用意する必要があるため，分析者の想定を超えた未知な事象間の関係を発見することは難しいと考えている．それに対して我々の研究では，先に述べたように，各ノード (キーワード) 同士をつなぐエッジに 9 種類の関係 (Subject/Action/What など) を明確にした上で，複数ノードをつないで一つの事象を表現している．これにより，エッジの意味の明確化だけではなく，各事象を表すネットワークが密に繋がることができるような構造になっ

ており，従来より深い知見及び未知性にあふれる知見の獲得が可能になると考えている．

関連の研究の最後に，マイクロブログに投稿される情報をソーシャルセンサーとみなし応用した研究に触れる．代表的な研究として，榊らによる地震検知への応用 [15]，荒牧らによるインフルエンザの流行度合いの見積りへの応用 [16, 17]，などが挙げられる．榊らは Twitter のツイートから地震関連の語が含まれているツイートに対して，いくつかの特徴を定義し SVM を用いてツイートがターゲットイベントであるかを判別し，実際にターゲットイベントが検出および発生場所の特定を行う手法を提案した．そして実際に高精度で地震を検知し，ユーザに地震が到達する前にメールで通知するシステムを開発した [15]．荒牧らはインフルエンザに関連の単語が含まれているツイートに対して，ツイート投稿者がインフルエンザ患者なのかどうかを SVM を用いてインフルエンザ関連の語の前後を特徴語とみなし現実的な精度で判別を行えることを示し [17]，そのモデルを用いてインフルエンザ数を見積り，実際のインフルエンザ患者数と高い正の相関があることを示し，Twitter からインフルエンザの流行を十分に見積もれることを示した [16]．我々の研究においても，実際に Twitter から事象群の抽出を行なっているため，Twitter をソーシャルセンサーとしてみなした研究と考えられる．上記の研究では分類器に学習させる際の特徴にツイートの単語数，キーワードが何番目の単語か，ツイート中に出現する各単語，ツイートは各指定キーワードの前後の単語を特徴にするなど，単語間の共起頻度に着目した特徴を与えており，ツイートを解析する際に表面的な特徴を利用している一方，我々の研究では，前述の通り事象を各ノード (キーワード) 間の関係の意味を与えたネットワークで表現するなどセマンティックな解釈目指している点が異なる．

第7章　おわりに

本稿では事象を表すための事象属性を定義し，自然言語から事象情報を抽出し，Linked Data として構造化した状態に変換する手法を提案した．5 章において，事象情報ネットワークとして可視化することで，メディア比較のドメインへ適用することを示した．いくつかの観点から比較ポイントを提示するエージェントサービスを提案した．また，実験を通じて一定の有効性を確認した．

本手法はメディア比較だけにとどまらず，株価予測を行う際に事象症状を特徴として与える時に用いるなど，用途は無限にある．今後の課題として，その他のシステムの改善点としては，他事例についてのメディア比較，多種な話題に対応する訓練データの汎用化，事象抽出精度の改善，より整理された情報がネットワークで表現できるよう RDF 構造の見直し，比較作業の容易化のための可視化ツールの開発及びネットワークの注目ポイントのの自動検出アルゴリズムの改良などが挙げられる．

参考文献

[1] Tim Berners-Lee. Design issues: Linked data. Jul 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[2] Chris Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked data on the web (ldow2008). In *in WWW 2008*, pp. 1265–1266, Apr 2008.

[3] 武田英明. Linked data の動向. カレントアウェアネス, No. 308, pp. 8–11, 2011.

[4] The-Minh Nguyen, Takahiro Kawamura, Yasuyuki Tahara, and Akihiko Ohsuga. Self-supervised capturing of users' activities from weblogs. *IJIIDS*, Vol. 6, No. 1, pp. 61–76, 2012.

[5] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp. 282–289, 2001.

[6] G.D. Forney. The viterbi algorithm. In *Proc. IEEE*, 第 61 巻, 1973.

[7] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *HLT-NAACL*, pp. 213–220, 2003.

[8] Andrew McCallum and Wei Li.

[9] Taku Kudo, Kaoru, Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphologiaical analysis. In *EMNLP*, pp. 230–237, 2004.

[10] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *WWW*, pp. 909–918, 2012.

[11] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *CHI*, pp. 227–236, 2011.

[12] 藤川智英, 鍛治伸裕, 吉永直樹, 喜連川優. マイクロブログ上の話題抽出とユーザの態度の分類に基づく流言検出支援システム. 第 4 回データ工学と情報マネジメントに関するフォーラム（DEIM2012）, 3 2012.

[13] 風間一洋, 鳥海不二夫, 榊剛史, 篠田孝祐, 栗原聡, 野田五十樹. 東日本大震災時の twitter データを用いた単語間の関係の時系列変化の分析. 第 26 回人工知能学会全国大会, 6 2012.

[14] 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. 電子情報通信学会論文誌 D, Vol. J95-D, No. 7, pp. 1454–1466, 2012.

[15] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, Feb 2012.

[16] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Influenza patients are invisible in the web: Traditional model still improves the state of the art web based influenza surveillance. 2012.

[17] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *EMNLP*, pp. 1568–1576, 2011.

謝辞

本研究を行なうにあたり，ご多忙の中，終始適切かつ丁寧なご指導を下さった大須賀昭彦教授，田原康之准教授，太田敏澄准教授に深く感謝いたします．川村隆浩客員准教授，中川博之助教，清雄一助教にはご多忙の中，週 1 回のゼミを初めとして熱心な研究指導を賜り，研究のみならず様々な面でサポートしていただきました．厚く御礼申し上げます．大須賀・田原研究室の皆様，国立情報学研究所・東京大学の本位田研究室の皆様，早稲田大学の深澤研究室の皆様に感謝の意を表します．最後に，いつも温かく見守ってくれていた家族に心より感謝申し上げます．

研究業績

論文誌

1. **越川 兼地**，川村 隆浩，中川 博之，田原 康之，大須賀 昭彦：メディア情報からの事象抽出と LinkedData 化 - ソーシャルメディアとマスメディアの比較事例 - ，電子情報通信学会論文誌 ソフトウェアエージェントとその応用 特集号 (投稿中)

国際会議

1. T. M. Nguyen, **K. Koshikawa**, T. Kawamura, Y. Tahara, A. Ohsuga: Building Earthquake Semantic Network by Mining Human Activity from Twitter, Proceedings of 2011 IEEE International Conference on Granular Computing (GrC 2011), 2011.

査読付き国内シンポジウム・ワークショップ

1. **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. (ロング発表論文) 学生優秀論文賞

2. グェン ミンティ，**越川 兼地**，川村 隆浩，田原 康之，大須賀 昭彦：震災時の効率的な避難のための行動推薦エージェント Ready for 87%の提案 (2) - 時系列避難行動ネットワークの構築 -，合同エージェントワークショップ＆シンポジウム (JAWS 2011), 2011. (ロング発表論文)

付 録A　キーワードフィルタリングで使用したキーワード

A.1　オスプレイ導入問題

表 A.1: 話題:「オスプレイ」に関する絞り込みをするために使用したキーワードリスト

キーワードリスト (1 語)
オスプレイ

A.2　選挙 (2012 年東京都知事選挙・第 46 回衆議院議員総選挙)

表 A.2: 話題:「第 46 回衆議院議員総選挙・2012 年東京都知事選挙」に関する絞り込みをするために使用したキーワードリスト

キーワードリスト (110 語)					
みんなの党	トクマ	マック赤坂	不信任	与党	世襲
中松義郎	五十嵐政一	信任	候補	党首	入閣
公明	公示	公約	公選法	共産党	出口調査
出馬	制度	協定	参政	参議院	吉田重信
国会	国民	国民新党	外相	大臣	太田和美
宇都宮健児	安倍	官邸	審査	小沢一郎	小泉
山本太郎	市民革命	市議会	当確	当選	投票
捏造	支持	改憲	改革	政党	政府
政権	政治	政策	日本共産党	最高裁	有権者
未来の党	松沢成文	核兵器	椙杜徳馬	橋下さん	橋下徹
次期市長	死票	死に票	生き票	比例	民主
民意	民権	求心力	池上彰	法案	演説
無所属	猪瀬直樹	獲得	由紀夫	町村	発足
白票	真紀子	知事	石原慎太郎	石破	社会党
笹川堯	維新	総理	総裁	社民党	立候補
自公	自民	菅直人	落選	衆議院	衆院
街頭演説	裁判官	議員	議席	辞任	進次郎
過半数	選挙	都知事	野党	野田	開票
首相	鳩山				