大規模テキストデータを用いた 社会構造ネットワークモデルの自動抽出

大竹 竜太 [*1]　村田 真樹 [*2]　徳久 雅人 [*2]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*1,*2]{s092012,murata,tokuhisa}@ike.tottori-u.ac.jp

1　はじめに

現在，インターネット上で様々な電子テキストが増加している．これらの電子テキストから有益な情報を取り出すことが望まれている．また地震などの大きな社会的な出来事も多くなり，社会構造を的確に把握する技術が望まれている．そこで本研究では，電子テキストから社会構造の把握に役立つ情報を取得することを試みる．

本研究では, 事物の関係情報をネットワークとしてまとめたものを社会構造モデルと呼ぶ．本研究では，情報抽出の技術を用いて社会構造モデルを自動で構築する．さらに，構築した社会構造モデルにおいて，活性伝搬を行い，モデルにおいてどういう概念が特に重要であるかの分析も行う．本研究では具体的には「地震」に関わる社会構造モデルを扱う．

関連研究として，松尾らは Web 上の情報からの人間関係ネットワークを構築している [1]．この研究は，人間に関するネットワークを構築するという点が本研究と異なる．また，松尾らは，文書において，単語の同一文中での共起頻度を用いた Small World 構造を用いてキーワードの抽出 [2] を行っている．

本研究の特徴は，社会構造モデルを抽出する手法として，実験データの比較，またノードの抽出方法の比較を行った点である．また地震を題材にして作成した社会構造モデルのネットワークにおいて活性伝搬を行い，地震が起きた際に重要となる可能性のある概念を抽出したことである．

2　提案手法

提案手法では，電子テキストから社会構造モデル (事物の関係情報のネットワーク) を構築する．社会構造モデルのネットワークにおいて，活性伝搬を行い，ネットワーク上での重要な概念を考察する．

2.1　社会構造モデルの構築

提案手法では，まず最初に構築したい社会構造モデルの主となる概念を，キーワードとして設定する．そのキーワードに関係した電子テキストを抽出する．そのテキストにおいて，キーワードと関係性の強い単語を抽出する．関係性が強いとされた単語と関係性が強い単語も抽出する．キーワードと抽出した単語をノードとしたネットワークを構築し，そのネットワークを社会構造モデルとする．

キーワードに関係性の強い単語の決定には条件付き確率か TF-IDF を用いる．モデルのエッジには，条件付き確率か TF-IDF のスコアに基づく値を付与する．

より詳細な社会構造モデルの構築方法を以下で説明する．

キーワードとなる単語を単語 a とする．まず単語 a を含んだ記事群を抽出する．抽出された記事群を記事群 A とする．形態素解析を用い記事群 A から名詞のみを抽出する．その際に一文字，ひらがなのみ，数字のみの単語を除外する．記事群 A 内で抽出された単語の出現頻度をそれぞれ求め，抽出した名詞群の上位 100 単語をモデルのノードの候補とする．

ノードの候補の中から，条件付き確率と TF-IDF のどちらかを用いて，実際にノードに用いる単語を選定する．

条件付き確率を用いる方法を説明する．A を単語 a を含んだ記事群，B をノード候補の単語を含んだ記事群とし，$n(A)$ は単語 a を含んだ記事数，$n(A \cap B)$ は単語 a とノード候補の単語が同じ記事内で共起した記事数であるとし条件付き確率を式 1 で表す．

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{n(A \cap B)}{n(A)} \tag{1}$$

この値が大きいノード候補の単語をモデルのノードとして用いる．

TF-IDF を用いる方法を説明する．TF は抽出された対象テキスト内でのノード候補の単語の出現回数，DF は新聞データ内でのノード候補の単語の出現記事数とし，N は新聞データの総記事数とし TF-IDF を式 2 で表す．

$$w = TF * \log \frac{N}{DF} \tag{2}$$

この値が大きいノード候補の単語をモデルのノードとして用いる．

上記方法で取得したモデルのノード n は，単語 a のノードからつながるノードとする．単語 a のノードからノード n への重みは，ノード n を取得する際に得られた条件付き確率か TF-IDF の値に基づく値を利用する．

上記の方法を繰り返し用いることで，モデルのノードとして選定した単語と関係の深い単語もモデルのノードとする．ノードの候補として選定した単語を上記の単語 a として，上記の手法で新たにモデルのノード候補を作成し，ノード候補中から条件付き確率か TF-IDF を用いてさらにノードとして用いる単語を選定する．

2.2 活性伝搬

活性伝搬では，社会構造モデルの各ノードが活性値を，そのノードに連結している他のノードに伝搬させる [3]．各ノードの活性値の変化によって考察を行う．本研究での活性伝搬は，式 3 により行う．

$$A(t) = C + ((1-\gamma)I + \alpha R(t))A(t-1) \qquad (3)$$

ここで，$A(t)$ は活性回数 t の語の活性値を表すベクトル，C はモデルに注入される活性値を表すベクトル，I は $A(t-1)$ の活性値を $A(t)$ に伝搬させる単位行列，$R(t)$ はネットワークの構造を表す伝搬行列であり，$R(t)$ の i 行 j 列の要素 Rij は単語 Wi と単語 Wj の関連の強さを表す (対角成分は 0)．また，γ は活性値の減衰率を表す減衰パラメータ，α はネットワークが単語の活性値に及ぼす影響力の程度を表す伝搬パラメータである．

3 実験データの選定

本節では事前実験として，どのようなデータが社会構造モデルの構築にふさわしいかを調べる．

実験データには，新聞と Wikipedia を用いる．新聞には，毎日新聞 2011 年の 1 年分の記事, 96,630 記事を用いる．また，Wikipedia には 1,602,208 記事が含まれる．

新聞と Wikipedia の比較のためにキーワードを含む記事を抽出し，抽出された記事群内の名詞の出現頻度を利用して単語抽出を行い，比較する．本研究では，キーワードは「地震」と「経済」とした．「地震」と「経済」の両方の単語が同時に出現した記事をキーワードに関連する記事群として抽出する．

抽出された記事群は，新聞データからは 514 記事であり，Wikipedia からは 2818 記事であった．抽出された記事群に出現する名詞を出現頻度順に整理し比較する．結果を表 1, 表 2 に示す．

Wikipedia では，頻度の高い単語であっても，地震，経済に直接関連しない単語が多く得られた．一方新聞データでは，地震や経済と関連の高い「原発」「事故」「安全」などの単語が抽出された．これの理由としては，以下が考えられる．Wikipedia では事柄の説明を単に記載しているだけであり，多くの事柄の説明をしている文章での頻度では関連の高い単語を抽出できなかったと思われる．一方新聞データでは，社会的に大きな事柄については高頻度に記述されるため，頻度により今回扱った地震，経済に関連の高い単語を抽出できたと思われる．

表 1: 新聞データ

単語	出現回数
原発	3604
事故	1594
安全	1570
福島	1477
地震	1371
原子力	1190
日本	1132
号機	1028
経済	970
東電	852
津波	849
大震災	832
政府	778
被災	759
対策	723
首相	686
保安	668
東日本	664
原子	643
評価	589

表 2: Wikipedia

単語	出現回数
放送	48947
日本	47033
番組	25279
東京	21992
テレビ	19350
地震	16774
平成	16533
利用	15941
昭和	15016
都市	14640
現在	14498
選手	14100
世界	13942
開始	13699
学校	13524
地域	13479
研究	13044
時代	12197
野球	11580
情報	11550

以上の結果より，Wikipedia よりも新聞データの方がキーワードに近い単語の取り出しに役立つことがわかった．このため，本研究での以降の実験では，新聞データを利用することにする．

Wikipedia には記事数が多く，抽出する記事群を減らし計算コストを削減するために「地震」「経済」をキーワードとしていた．しかし，新聞データではそこまで記事数を減らして計算コストを削減する必要はないため，以降の実験では，「地震」「経済」でなく，「地震」のみをキーワードとして用いることとする．

4 社会構造モデルの構築における条件付き確率と TF-IDF の比較

社会構造モデルの構築では，ネットワークのノードに用いる単語の決定のために，条件付き確率や TF-IDF を用いる．本節では，条件付き確率と TF-IDF のうちどちらを利用した方が，より良い社会構造モデルを構築できるかを調べる．

キーワードとして「地震」を用いる．「地震」を単語 a として提案手法を用いて，地震につながるノードに利用する単語を取得する．

提案手法の条件付き確率を用いる方法でノードに利用する単語を取得した結果を表 3 に示す．また TF-IDF を用いる方法で取得した結果を表 4 に示す．それぞれ条件付き確率と TF-IDF の値の上位のものを示している．

TF-IDF を用いた場合には，「津波」「原発」「避難」な

表 3: 条件付き確率

単語	条件付き確率
地震	1.000
日本	0.786
震災	0.707
大震災	0.663
東日本	0.618
被災	0.461
津波	0.448
東京	0.392
福島	0.377
発生	0.358
避難	0.346
被害	0.343
原発	0.323
事故	0.274
宮城	0.254
災害	0.243
岩手	0.220
対策	0.220
キロ	0.211
安全	0.210

表 4: TF-IDF

単語	TF-IDF
地震	15047
津波	8318
原発	7394
避難	6584
被災	5522
福島	4903
電話	4723
大震災	3796
発生	3693
事故	3575
宮城	3550
災害	3517
安全	3295
被害	3237
岩手	3229
東日本	3157
防災	3053
対策	2749
支援	2671
原子力	2623

どの地震が起きた際に特に関連が高いと思われる語が上位に集中した．さらに「電話」という地震が起きた際に注意すべき語も上位に現れた．

一方，条件付き確率を用いた場合は，「日本」「震災」「大震災」など地震には確かに関連があるが TF-IDF を用いた場合ほど関連のないものが上位にきた．この結果より，ノードの抽出には TF-IDF を利用した方がよいことがわかった．

以上の結果より，社会構造モデルのノードの抽出には TF-IDF を利用し，エッジに付与する重みにも TF-IDF のスコアを利用することにする．

条件付き確率を用いる手法が良くない結果となった理由は以下と思われる．もともと高頻度に出現する単語は地震と共起しやすく条件付き確率が高くなる．このため，高頻度で出現するが関連性はそれほど高くない単語が上位に現れたと思われる．

松尾らの人間関係ネットワークの抽出 [1] の際には，ノード間の関連性の取得に閾値つきの Simpson 法を利用するのが良いとされていた．この方法やそれに類似する方法も本研究で試したが条件付き確率と同様の結果となった．

5 TF-IDF を用いた社会構造モデルの構築

キーワードを「地震」として，TF-IDF を用いる提案手法により，社会構造モデルを構築する．キーワード「地震」から得られた単語を単語 a として同様の手順を用いて単語 a と関連性の高い単語を抽出する．これらの手順を複数繰り返し「地震」と直接つながらない単語をノードに持つモデルを構成する．単語 a に対してモデルのノードとして抽出する単語は，TF-IDF のスコア上位 5 単語とする．

単語 a から 5 つの単語へのエッジのスコアは，その 5 つの単語の TF-IDF のスコアから計算される確率で求める．5 つの単語のうちの一つである単語 n へのエッジのスコアは式 4 で表される．

$$score = \frac{\text{単語 } n \text{ の } TF-IDF}{\text{5 単語の } TF-IDF \text{ の和}} \tag{4}$$

この手法により社会構造モデルを自動構築した．「地震」を第一単語群，「地震」から抽出された単語を第二単語群，第二単語群から抽出された新しい単語を第三単語群，同様に第四単語群とする．その抽出結果を表 5，表 6, 表 7 に示す．

表 5: 第二単語群

第二単語群	津波, 原発, 避難, 被災, 福島

表 6: 第三単語群

第三単語群	宮城，事故，原子力，東電，電話， 大震災, 復興，東日本

表 7: 第四単語群

第四単語群	岩手，安全，号機，東京電力，相談， 携帯，ボランティア，東京，首相，支援

次に，単語 a とした単語と，その単語につながるノードとして得られた単語を，表 8，表 9，表 10 に示す．表中の単語の後ろの括弧内の数字はその単語へつながるエッジが持つ重み (確率) である．

表 8: 第一単語群からの抽出結果

元の単語	抽出された単語
地震	津波 (0.254), 原発 (0.226), 避難 (0.201), 被災 (0.169), 福島 (0.15)

表 9: 第二単語群からの抽出結果

元の単語	抽出された単語
津波	避難 (0.246), 被災 (0.212), 原発 (0.196) 地震 (0.189), 宮城 (0.158)
原発	福島 (0.293), 事故 (0.256), 原子力 (0.157) 避難 (0.151), 東電 (0.143)
避難	福島 (0.238), 被災 (0.214), 原発 (0.194) 電話 (0.184), 津波 (0.17)
被災	電話 (0.251), 避難 (0.199), 大震災 (0.196) 復興 (0.177), 東日本 (0.177)
福島	原発 (0.337), 事故 (0.196), 電話 (0.173) 避難 (0.161), 被災 (0.133)

「地震」を含んだ 24 個のノードが抽出された．それらのノードを TF-IDF を用いた確率値が繋いでいる．

抽出された社会構造モデルの一部を図 1 に示す．図では，ノードは第三単語群までのものを表示した．

表 10: 第三単語群からの抽出結果

元の単語	抽出された単語
宮城	電話 (0.281), 被災 (0.226), 福島 (0.175), 岩手 (0.166), 避難 (0.153)
事故	原発 (0.354), 福島 (0.297), 原子力 (0.117), 避難 (0.117), 東電 (0.114)
原子力	原発 (0.365), 事故 (0.190), 福島 (0.168), 安全 (0.151), 号機 (0.126)
東電	原発 (0.301), 号機 (0.195), 事故 (0.180), 福島 (0.177), 東京電力 (0.147)
電話	相談 (0.523), 被災 (0.127), 携帯 (0.120), ボランティア (0.120), 東京 (0.108)
大震災	被災 (0.224), 福島 (0.210), 東日本 (0.207), 原発 (0.184), 避難 (0.175)
復興	被災 (0.274), 首相 (0.201), 大震災 (0.183), 支援 (0.179), 東日本 (0.161)
東日本	被災 (0.221), 大震災 (0.219), 福島 (0.210), 原発 (0.178), 避難 (0.173)

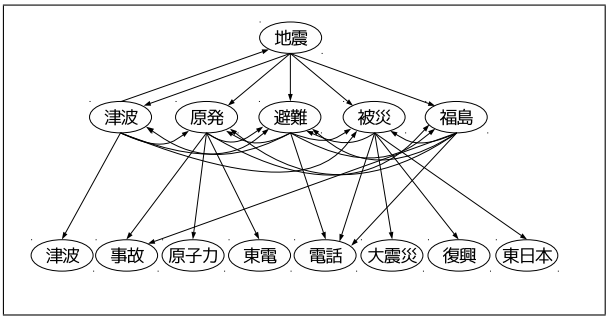

図 1: 抽出された社会構造モデルの一部

6 活性伝搬を用いた実験

前節で構築した社会構造モデルにおいて，実際に活性伝搬を行ってみる．

式 3 を用いて, 活性回数 t のときの単語の活性値を表すベクトル $A(t)$ の変化をもとめる．ここでは，地震が活性した場合の結果を調べることとして，初期値 A(0) には地震のみ 1 とし他を 0 としたベクトルを用いる．モデル外部からの刺激はないものとして $C = 0$ とする．また， $\gamma = 0, \alpha = 1$ とする．

A(t) の変化を表 11 に示す．表のように，活性回数 10 回 $(t = 10)$ のときに活性値が他の単語に比べて大きくなったのは，第二単語群では，原発，福島，避難，第三単語群では，電話，事故，原子力，第四単語群では，相談，安全，携帯，ボランティアであった．これらの単語の活性値が大きくなったのは，他の単語に比べ，これらの単語を指すエッジの数が多く，また，エッジに付与された確率も大きかったためである．

これらの活性値が大きくなった単語は，地震が起きた際に特に重要な概念となる可能性がある．

7 おわりに

テキストから社会構造の把握に役立つ社会構造モデル (ネットワーク) の情報を取り出す研究を行った．実験データとして新聞と Wikipedia を比較し，本研究の実験では社会構造モデルの構築に新聞の方が役立つことを確認した．

また，社会構造モデルのネットワークのノードの抽出には，条件付き確率よりも TF-IDF の方が役立つことを確認した．

実際に地震に関する社会構造モデルを抽出し，そのネットワークにおいて活性伝搬を行った．活性伝搬により，地震で重要となる可能性のある概念を抽出できた．

表 11: 各単語の活性値の変化

活性回数	t = 1	t = 2	t = 3	t = 10
第一単語群				
地震	1.000	1.048	1.151	7.697
第二単語群				
津波	0.254	0.542	0.903	23.263
原発	0.226	0.591	1.247	125.539
避難	0.201	0.556	1.182	90.412
被災	0.169	0.455	0.963	76.418
福島	0.150	0.414	0.935	111.128
第三単語群				
宮城	0.000	0.040	0.126	5.598
事故	0.000	0.087	0.332	69.285
原子力	0.000	0.036	0.139	29.755
東電	0.000	0.032	0.127	27.625
電話	0.000	0.105	0.411	70.863
大震災	0.000	0.033	0.128	20.908
復興	0.000	0.030	0.110	15.928
東日本	0.000	0.030	0.122	23.578
第四単語群				
岩手	0.000	0.000	0.007	1.440
号機	0.000	0.000	0.005	4.544
安全	0.000	0.000	0.011	9.229
東京電力	0.000	0.000	0.005	4.099
相談	0.000	0.000	0.055	40.274
携帯	0.000	0.000	0.013	9.241
ボランティア	0.000	0.000	0.013	9.241
東京	0.000	0.000	0.011	8.317
首相	0.000	0.000	0.006	3.591
支援	0.000	0.000	0.005	3.198

謝辞

この研究は栢森情報科学振興財団の助成を受けて遂行された．

参考文献

[1] 松尾 豊，友部 博教，橋田 浩一，中島 秀之，石塚 満 : “Web 上の情報からの人間関係ネットワークの抽出”, 人工知能学会論文誌 20 巻 1 号 E，pp.46-56，2005.

[2] 松尾 豊, 大澤 幸生, 石塚 満: “Small World 構造に基づく文書からのキーワード抽出”, 情報処理学会論文誌 43(6)，pp.1825-1833，2002.

[3] 松村 直宏，大澤 幸生，石塚 満 : “語の活性度に基づくキーワード抽出法”, 人工知能学会論文誌 17 巻 4 号 F, pp.398-406, 2002.