概要

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おり，それらの中から有益な情報を取り出すことが望まれている．また，リーマンショックや東日本大震災など，社会を揺るがす出来事も多くなり，社会構造を的確に把握する技術が望まれている．

松尾ら [1] は，Web 上の情報から人間関係のネットワークを抽出しているが，社会構造に着目したネットワークを抽出している研究はない．

そこで本研究では，社会的な事物に着目し，Web 上の情報から社会構造を対象としたネットワークの抽出手法を提案した．また Web 上の電子テキストからキーワードに基づき抽出されるネットワークを社会構造モデルと呼ぶ．

実験対象の電子テキストとして，Wikipedia と新聞の比較を行い，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抽出手法として，TF-IDF，条件付き確率を用いた実験を行い，TF-IDF 法が有効であることを示した．

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．

目 次

第 1 章　はじめに **1**

第 2 章　関連研究 **2**

第 3 章　提案手法 **3**

3.1　社会構造モデルの構築 3

3.2　活性伝搬 6

第 4 章　実験 **8**

4.1　実験データの選定 8

4.2　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 10

4.2.1　条件付き確率と TF-IDF によるノードの抽出 10

4.2.2　被験者による評価 12

4.2.3　以降の実験に用いるノードの抽出手法 12

4.3　TF-IDF を用いた社会構造モデルの構築 13

4.3.1　ノードの抽出結果 13

4.3.2　エッジの重みの計算結果 14

4.3.3　抽出例 17

4.4　活性伝搬を用いた実験 18

4.4.1　活性伝搬 18

4.4.2　伝搬行列 18

4.4.3　活性伝搬式の比較実験 19

4.4.4　活性伝搬結果 22

第 5 章　考察 **25**

5.1　実験データの選定についての考察 25

5.2　ノードの抽出における条件付き確率と TF-IDF の比較についての考察 . . 25

5.3 抽出された社会構造モデルについての考察 25
5.4 活性伝搬についての考察 26

第 6 章 今後の課題 **27**

第 7 章 おわりに **28**

表 目 次

4.1 新聞データにおける単語の抽出 9
4.2 Wikipedia における単語の抽出 9
4.3 条件付き確率による単語の抽出 10
4.4 TF-IDF による単語の抽出 10
4.5 抽出方法の評価 12
4.6 第二単語群 13
4.7 第三単語群 13
4.8 第四単語群 13
4.9 第一単語からの抽出結果 14
4.10 第二単語群からの抽出結果 15
4.11 第三単語群からの抽出結果 16
4.12 $\gamma = 0$ のときの各単語の活性値の変化 19
4.13 $\gamma = 0.5$ のときの各単語の活性値の変化 20
4.14 $\gamma = 1$ のときの各単語の活性値の変化 21
4.15 $\gamma = 0$ のときの活性値の上位の単語 22
4.16 $\gamma = 0.5$ のときの活性値の上位の単語 22
4.17 $\gamma = 1$ のときの活性値の上位の単語 23

図 目 次

3.1 社会構造モデルのノード抽出の例 5
3.2 活性伝搬の例 7

4.1 抽出された社会構造モデルの一部 17
4.2 活性伝搬の結果例 24

第1章　はじめに

近年，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．またリーマンショックや地震など，社会を揺るがす出来事も多くなり，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，電子テキストから特定のキーワードに基づく関係情報 (ネットワーク) を抽出する方法を提案する [2]．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.「地震」というキーワードに基づいて社会構造モデルの抽出を行った.

また，抽出された単語の関係間のエッジに重みを持たせることで活性伝搬 [3] を用い，モデルにおいてどういった概念が特に重要であるかの分析も行った [4].

本研究の主な特徴をあらかじめ整理すると以下のようになる.

- テキストから社会構造の把握に役立つ社会構造モデルの情報を取り出すという特色のある研究対象を扱った.
- 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は新聞の方が役立つことを確認した.
- 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.
- 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，地震が起きた際に特に重要となる可能性のある概念を抜き出した.

第 2 章では，本研究の関連研究を述べ，第 3 章では，提案手法の手順，説明を行い，活性伝搬の方法を説明する．第 4 章で実験データの選定，比較手法との比較を行う．第 5 章では，結果の考察を行う．第 6 章では，今後の課題の説明を行う.

第2章　関連研究

関連研究としては以下のものがある．

松尾ら [1] は，Web 上の情報から，人間関係のネットワークを抽出している．その際に，抽出手法として，氏名の関係性の強さを知るために様々な指標を用いて実験している．

松村ら [3] は，文書の主張をキーワードとし，文書の要約や文書検索のために，語の活性度に基づいたキーワード抽出法を提案している．

松尾ら [5] は，ノードが離れているにも関わらず，別のノードを介せば近いという Small World 構造を用いてネットワークを構築し，そのネットワークからキーワードを抽出する手法を提案した．

内山ら [6] は，大規模な出来事の要約，すなわち，複数のトピックに関する複数の文書の要約を目的としている．複数文書においてネットワークを構成し，ネットワークの各ノードの重要度を活性拡散を利用し求めている．それにより，複数文書の要約を行っている．

森ら [7] は，Web 上の情報を用いて，研究者の情報をキーワードとして自動的に抽出する手法を提案している．研究者の情報とは，例えば，所属組織，研究テーマ，共著者などである．それらの研究者の情報をキーワードから自動で抽出している．

岡崎ら [8] らは，Web 文書から人の安全，危険に関わる情報を抽出している．談話構造に基づく論述構造の分析を行い，Web からの文章に対して分類を行うことで情報の構造化を行っている．その情報に基づき必要な情報を抽出している．

小嶋ら [9] は，英語の物語における場面の境界を推定するための統計的な指標を提案している．場面ごとに現れる単語は互いに結束性によって結ばれる傾向をもつ．この単語列の結束度を用いてテキスト区画の境界を推定している．

第3章　提案手法

3.1　社会構造モデルの構築

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．

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．次に関係性が強いとされた単語とさらにその単語に関係性が強い単語も抽出する．繰り返し抽出を繰り返すことで社会構造モデルを拡大していく．

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．

ノード候補の抽出

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．

ノードの選定

得られた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．選定を行う際には，条件付き確率，または TF-IDF のスコアの上位 5 単語をキーワードと関係性の強い単語とする．

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 3.1 で表す．

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{n(A \cap B)}{n(A)} \tag{3.1}$$

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 3.2 で表す．

$$w = tf * \log \frac{N}{df} \tag{3.2}$$

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．上記の方法で選定した 5 単語を単語 a のノードから繋がるノード n とする．

エッジに重みの付与

単語間の関係 (エッジ) に重みを付与し，単語間の関連の強さに差をつける．エッジに付与する重みを式 3.3 に示す．

$$score = \frac{\text{単語 } n \text{ の } tfidf}{5 \text{ 単語の } tfidf \text{ の和}} \tag{3.3}$$

ここで，単語 n は単語 a から抽出された 5 単語のうちの 1 単語とする．単語 a 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 TF-IDF に基づく値を利用する．

社会構造モデルの拡大

単語 a から 5 つの単語が抽出される流れを上記で説明した．これによって得られた単語 n を新たに単語 a' と設定し同様の手順で単語 a' から 5 つの単語を抽出する．これにより単語 a から抽出された 5 つの単語にさらに単語 a' から抽出された単語 5 つが加る．同様に各単語からの抽出を繰り返すことで社会構造モデルを拡大していく．

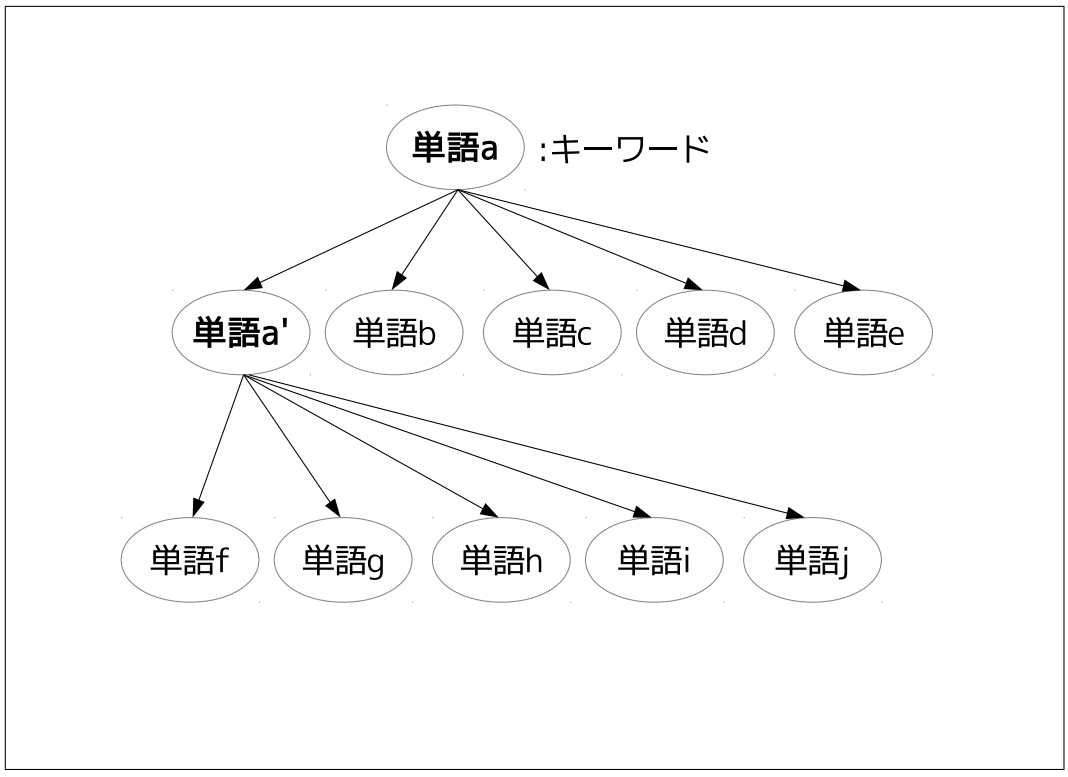

図 3.1: 社会構造モデルのノード抽出の例

3.2 活性伝搬

人間の記憶のメカニズムを近似したものに活性伝搬モデルというものがある．活性伝搬は，エッジで結ばれたネットワーク構造において，活性を伝搬させ，その活性度の変化を調べることでネットワークのノードの重要度を計るという考えである．

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．伝搬した際の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 3.4 により行う．

$$A(t) = C + ((1-\gamma)I + \alpha R(t))A(t-1) \tag{3.4}$$

ここで，t はモデルを活性させる活性回数であり，$A(t)$ は活性回数 t のときの各ノードの活性値を表すベクトル，C はモデルに外部から注入される刺激を表すベクトル，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のエッジの重みに基づき表される伝搬行列である．$R(t)$ の i 行 j 列の要素 Rij は単語 Wi と単語 Wj の関連の強さを表している．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．

本研究では，社会構造モデルはそのモデルだけで完結しており，外部からの刺激はないものとする．よって式 4.1 の外部から注入される活性値を表すベクトル C は $C = 0$ とする．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 α は，活性の伝搬はモデルの構造を表すベクトル $R(t)$ によってのみ行われるため，$\alpha = 1$ とする．また, 減衰率を表す減衰パラメータ γ は，適応する文書により異なるため，減衰パラメータは $0 < \gamma < 1$ において比較実験を行う．

$$A(t) = ((1-\gamma)I + \alpha R(t))A(t-1) \tag{3.5}$$

よって，本研究の活性伝搬式には式 3.5 を用いることとする．

活性値が伝搬していく流れ表したものを図 3.2 に示す.

図において，単語 a が活性した際に，その活性がエッジの重みにより単語 a' に伝わることで単語 a' が活性し，単語 a' が活性することで，単語 j が活性している．このように，起点となる単語 (社会構造モデルのキーワード) が刺激され，活性し，その活性値がエッジの重みに基づき分散され各単語に伝わる.

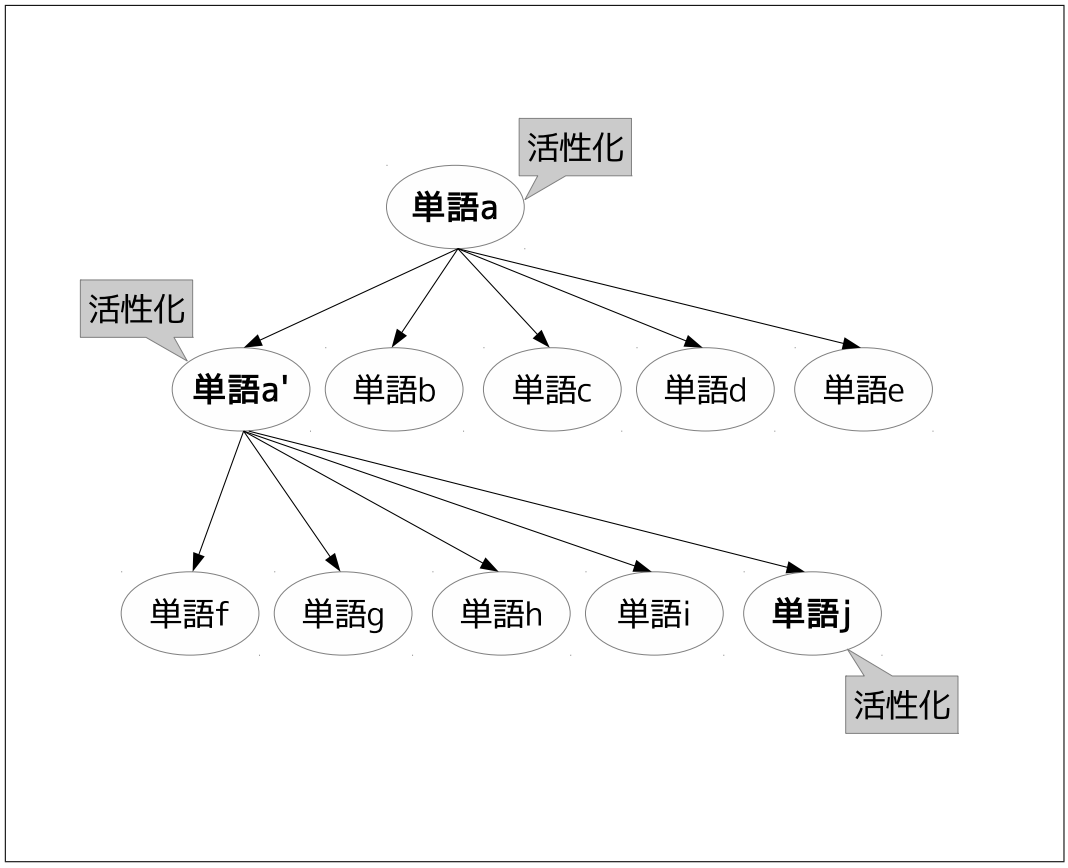

図 3.2: 活性伝搬の例

第4章　実験

4.1　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 4.1, 表 4.2 に示す．

表 4.1: 新聞データにおける単語の抽出

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 4.2: Wikipedia における単語の抽出

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

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 では多くの事柄の説明を簡潔に記載しているだけであり，ある重要な事柄が頻度が高く繰り返し記載されるということはないため，そのような文章の頻度では，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．

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．

4.2 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」をキーワードとし提案手法を行い，地震につながるノードに利用する単語を取得する．

4.2.1 条件付き確率とTF-IDFによるノードの抽出

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 4.3 に示す．また TF-IDF を用いる方法で取得した結果を表 4.4 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．

表 4.3: 条件付き確率による単語の抽出

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 4.4: TF-IDF による単語の抽出

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

TF-IDF を用いた場合には,「津波」「原発」「避難」な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 を利用した方が良いことがわかった．

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．

4.2.2 被験者による評価

前節で得られた結果を元に，TF-IDF 法と条件付き確率法どちらが社会構造モデルを抽出するのに適しているかを判断するために人手評価を被験者 8 人に対して行う．前節の表 4.4 と表 4.3 の抽出結果の一部を示し，どちらの手法が適しているかを判断してもらう．

結果を表 4.5 に示す．表に示された数字は，その手法を良いとした人数である．

表 4.5: 抽出方法の評価

TF-IDF 法○	条件付き確率法○
7	1

4.2.3 以降の実験に用いるノードの抽出手法

前節での人手評価で得られた結果により，TF-IDF を用いる方が適していることがわかる．また，抽出結果の考察によっても，TF-IDF を用いる方が適している．

よって，本研究のネットワークの抽出の手法には TF-IDF を用いることとする．また，エッジに付与する重みにも TF-IDF のスコアを用いる．

4.3 TF-IDF を用いた社会構造モデルの構築

4.3.1 ノードの抽出結果

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 4.1 で表される．

$$score = \frac{\text{単語}\, n \,\text{の}\, tfidf}{\text{5 単語の}\, tfidf \,\text{の和}} \tag{4.1}$$

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 4.6，表 4.7, 表 4.8 に示す.

表 4.6: 第二単語群

第二単語群	津波, 原発, 避難, 被災, 福島

表 4.7: 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 4.8: 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

4.3.2　エッジの重みの計算結果

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 4.9, 表 4.10，表 4.11 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である．

第一単語からの抽出結果

表 4.9 に第一単語である地震からの抽出結果と各単語へのエッジの重みを示す．

表 4.9: 第一単語からの抽出結果

元の単語	抽出された単語	各単語へのエッジの重み
地震	津波	0.254
	原発	0.226
	避難	0.201
	被災	0.169
	福島	0.150

第二単語群からの抽出結果

表 4.10 に第二単語群からの抽出結果と各単語へのエッジの重みを示す.

表 4.10: 第二単語群からの抽出結果

元の単語	抽出された単語	各単語へのエッジの重み
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

第二単語群からの抽出結果

表 4.11 に第三単語群からの抽出結果と各単語へのエッジの重みを示す.

表 4.11: 第三単語群からの抽出結果

元の単語	抽出された単語	各単語へのエッジの重み
宮城	電話	0.281
	被災	0.226
	福島	0.175
	岩手	0.1656
	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	0.19
	福島	0.168
	安全	0.151
	号機	0.126
東電	原発	0.301
	号機	0.195
	事故	0.78
	福島	0.177
	東京電力	0.147
電話	相談	0.523
	被災	0.127
	携帯	0.12
	ボランティア	0.12
	東京	0.108
大震災	被災	0.224
	福島	0.21
	東日本	0.207
	原発	0.184
	避難	0.175
復興	被災	0.174
	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	0.21
	原発	0.178
	避難	0.173

同様にして，第三単語群からも各単語につき 5 つの単語が抽出され，さらに 5 つの単語それぞれに TF-IDF を用いた重みが付与されている．以上の結果より,「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．

4.3.3 抽出例

抽出された社会構造モデルの一部を図 4.1 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．

各エッジには TF-IDF を用いた重みが付与されている．

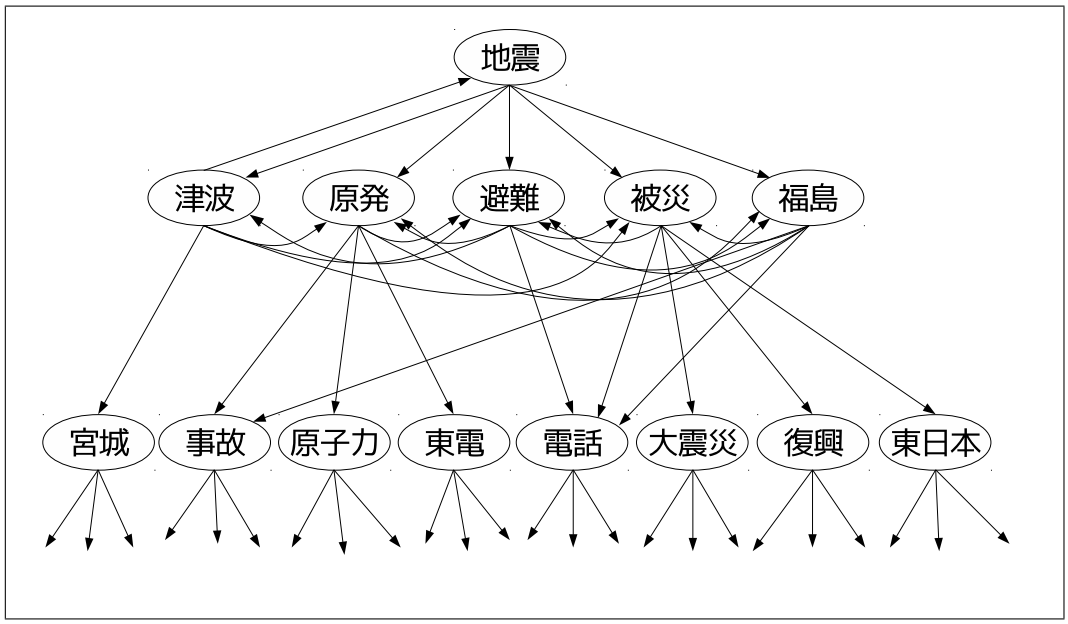

図 4.1: 抽出された社会構造モデルの一部

4.4 活性伝搬を用いた実験

4.4.1 活性伝搬

4.3 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う．

式 3.4 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 $A(t)$ の変化を求める．活性回数とは，モデルのキーワードである地震に 1 を入力し，入力された 1 がエッジの重みによって分散され，各ノードの活性値として蓄積されていく回数である．一定の活性回数で，各ノードの活性値を比べることで単語の重要度を調査する．

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無いものとして式 4.1 を用いて実験を行う．

また, 減衰率を表す減衰パラメータ γ を変えることで活性値の変化を比較する．影響力を表す伝搬パラメータ α は，活性の伝搬はモデルの構造を表すベクトル $R(t)$ によってのみ行われるため，$\alpha = 1$ とする．減衰パラメータは $0 < \gamma < 1$ において比較実験を行う．

具体的には，$\gamma = 0$，$\gamma = 0.5$，$\gamma = 1$，に分けて実験を行い，比較する．

4.4.2 伝搬行列

式 3.4 における，伝搬行列 $R(t)$ を説明する．伝搬行列 $R(t)$ はモデルの構造に基づき活性をノードからノードへ伝搬させる行列である．$R(t)$ の要素 Rij はノード間を繋ぐエッジに付与された重みである．つまり単語 Wi と単語 Wj を繋ぐエッジの重みが $R(t)$ の要素 Rij となる．

前節の表 4.9，表 4.10，表 4.11 に示した重みが伝搬行列 $R(t)$ の要素となる．

4.4.3 活性伝搬式の比較実験

式 3.4 において，γ の値による比較実験の結果を示す．活性回数 t は 10 までとし，表には $t = 1, t = 2, t = 3, t = 10$ のときの活性値を示す．

$\gamma = 0$ のとき

式 3.4 において $\gamma = 0$ のときの各単語の活性値を表 4.12 に示す．

表 4.12: $\gamma = 0$ のときの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				
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	9.318
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

$\gamma = 0.5$ のとき

式 3.4 において $\gamma = 0.5$ のときの各単語の活性値を表 4.13 に示す.

表 4.13: $\gamma = 0.5$ のときの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				
地震	0.500	0.298	0.204	0.201
第二単語群				
津波	0.254	0.288	0.280	0.850
原発	0.226	0.365	0.529	5.896
避難	0.201	0.355	0.498	3.929
被災	0.169	0.286	0.407	3.245
福島	0.150	0.264	0.426	5.220
第三単語群				
宮城	0.000	0.040	0.066	0.167
事故	0.000	0.087	0.202	3.408
原子力	0.000	0.036	0.085	1.472
東電	0.000	0.032	0.078	1.369
電話	0.000	0.105	0.253	3.137
大震災	0.000	0.033	0.078	0.895
復興	0.000	0.030	0.066	0.676
東日本	0.000	0.030	0.077	1.024
第四単語群				
岩手	0.000	0.000	0.007	0.038
号機	0.000	0.000	0.005	0.243
安全	0.000	0.000	0.011	0.494
東京電力	0.000	0.000	0.005	0.220
相談	0.000	0.000	0.055	1.899
携帯	0.000	0.000	0.013	0.439
ボランティア	0.000	0.000	0.013	0.436
東京	0.000	0.000	0.011	0.392
首相	0.000	0.000	0.006	0.161
支援	0.000	0.000	0.005	0.143

$\gamma = 1$ のとき

式 3.4 において $\gamma = 1$ のときの各単語の活性値を表 4.14 に示す.

表 4.14: $\gamma = 1$ のときの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				
地震	0.000	0.048	0.007	0.002
第二単語群				
津波	0.254	0.034	0.038	0.009
原発	0.226	0.139	0.151	0.067
避難	0.201	0.154	0.116	0.043
被災	0.169	0.117	0.105	0.034
福島	0.150	0.114	0.143	0.059
第三単語群				
宮城	0.000	0.040	0.005	0.002
事故	0.000	0.087	0.071	0.039
原子力	0.000	0.036	0.032	0.017
東電	0.000	0.032	0.030	0.016
電話	0.000	0.105	0.095	0.033
大震災	0.000	0.033	0.028	0.009
復興	0.000	0.030	0.021	0.007
東日本	0.000	0.030	0.032	0.010
第四単語群				
岩手	0.000	0.000	0.007	0.001
号機	0.000	0.000	0.005	0.003
安全	0.000	0.000	0.011	0.006
東京電力	0.000	0.000	0.005	0.003
相談	0.000	0.000	0.055	0.020
携帯	0.000	0.000	0.013	0.005
ボランティア	0.000	0.000	0.013	0.005
東京	0.000	0.000	0.011	0.004
首相	0.000	0.000	0.006	0.002
支援	0.000	0.000	0.005	0.001

4.4.4 活性伝搬結果

活性回数 10 回のときの活性値を単語群ごとに調査し，活性値上位の単語の抽出を行う．式 3.4 において $\gamma = 0$ のときの活性値の上位の単語を表 4.15 に示す．

表 4.15: $\gamma = 0$ のときの活性値の上位の単語

第二単語群	活性値
原発	125.539
福島	111.128
避難	90.412
第三単語群	
電話	70.863
事故	69.285
原子力	29.285
第四単語群	
相談	10.274
携帯	9.318
ボランティア	9.241
安全	9.229

式 3.4 において $\gamma = 0.5$ のときの活性値の上位の単語を表 4.16 に示す．

表 4.16: $\gamma = 0.5$ のときの活性値の上位の単語

第二単語群	活性値
原発	5.896
福島	5.220
避難	3.929
第三単語群	
電話	3.137
事故	3.408
原子力	1.472
第四単語群	
相談	1.899
携帯	0.439
ボランティア	0.436
安全	0.494

式 3.4 において $\gamma = 1$ のときの活性値の上位の単語を表 4.17 に示す．以上のように，γ

表 4.17: $\gamma = 1$ のときの活性値の上位の単語

第二単語群	活性値
原発	0.067
福島	0.059
避難	0.043
第三単語群	
電話	0.033
事故	0.039
原子力	0.017
第四単語群	
相談	0.020
携帯	0.005
ボランティア	0.005
安全	0.006

の値が変化すると式 3.4 に基づき単語の活性値自体は変化するが，γ の値が変化しても単語群内で活性値が大きくなる単語に変化はほとんど見られなかった．よって，式 3.4 における γ の値は，重要な概念の抽出とはほとんど関係がないことがわかる．

活性伝播の結果例

図 4.2 に活性伝搬を行った結果の一部を示す.

キーワードである地震を活性させた際に，その活性が伝わることで活性値の大きくなった単語が太字になっている単語である.

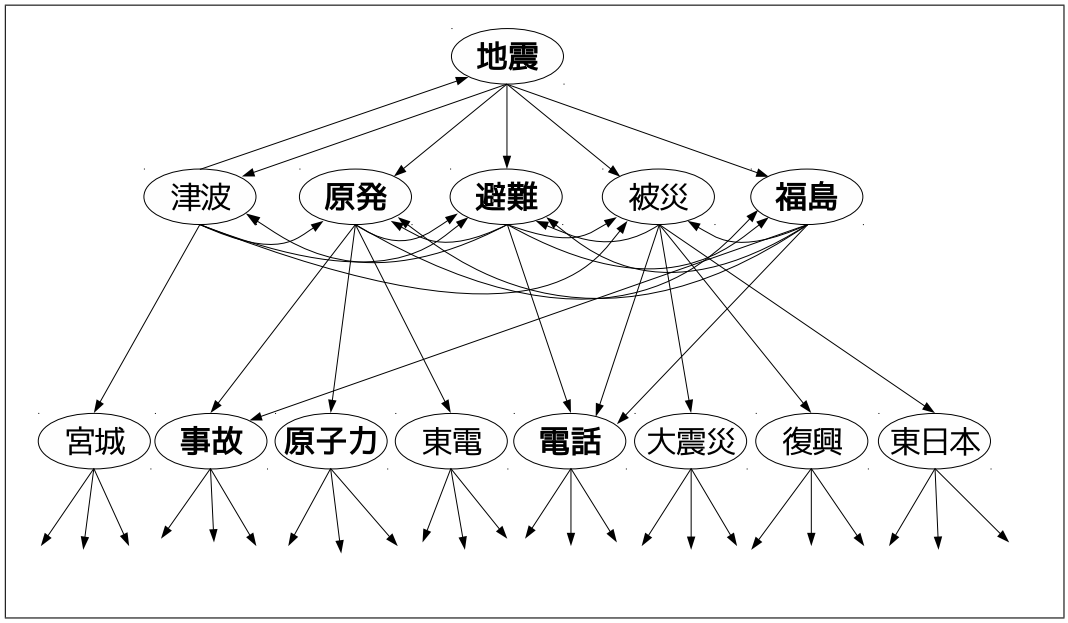

図 4.2: 活性伝搬の結果例

このように活性値がエッジの重みによって伝搬されることによって，単語の重要性を考察する．太字の単語のように，活性値の高くなった単語は，モデル内において重要な概念である可能性がある.

第5章　考察

5.1　実験データの選定についての考察

実験に扱うデータについての考察を行う．

4.1 節で新聞データと Wikipedia からの抽出結果を比較した．Wikipedia では頻度の高い単語においても重要でない単語が多く抽出された．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で記述されるため，抽出結果が良かった．

このような結果より，本研究では新聞データを使うのが適していると考える．

5.2　ノードの抽出における条件付き確率と TF-IDF の比較についての考察

条件付き確率を用いたノードの抽出方法と TF-IDF を用いたノードの抽出方法を比較した結果に対して考察を行う．人手評価では，TF-IDF の方が適しているという結果になり，抽出結果の考察においても TF-IDF の方が適しているという結果になった．これにより，ノードの抽出方法としては，TF-IDF を用いた手法の方が有効であることが確認できた．

5.3　抽出された社会構造モデルについての考察

抽出された社会構造モデルについて考察を行う．

抽出結果として，原発，避難，復興，事故など，地震に関連した単語が抽出できたと考えている．

5.4　活性伝搬についての考察

活性伝搬を用いて単語の重要度について考察を行う．

4.4 節に示したような結果が活性伝搬式により計算された．活性値の大きくなった単語として原発，福島，電話，事故，ボランティア，安全などがあげられる．これらの単語は地震により関係していると考える．これによって活性伝搬により重要な事物，概念を抽出できたと考える．

第6章　今後の課題

以下に，今後行うと良いと思われる実験を以下に示す.

実験 1

ネットワークを拡大する実験を行う．拡大したネットワークにおいて抽出された結果を考察することで提案手法を評価する.

実験 2

本論文では「地震」をキーワードとして用い社会構造モデルを構築した．しかし,「地震」に関わる事物は多くの人が知っていることであり，構築した社会構造モデルのネットワークにおいて意外性が少なかった．このため，キーワードを変更し，人が意外に思うようなノードを持つネットワークを構築する実験を行う．これにより抽出された結果の考察を行う.

今後,「経済」をキーワードとした抽出を行うことを考えている.「経済」は社会的に重要な概念であり,「経済」に関わる事物は複雑で得られるネットワークには，人が意外に思うノードが出現することが期待されるため，実験 2 のキーワードとして「経済」を用いるのがよいと考えている.

実験 3

関連研究との比較をより詳しく行う．関連研究における種々の手法を用いて抽出を行い評価する

第7章　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) 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，抽出されたネットワークが地震に関連していることを確認した．

抽出されたネットワークにおいて活性伝搬を行った．さらに活性伝搬により，地震で重要となる可能性のある概念を抽出できた．

謝辞

最後に，1 年間の間，研究を進めるに当たり，本研究のご指導を頂きました鳥取大学工学部知能情報工学科計算機Ｃ研究室の村田真樹教授，村上仁一准教授，徳久雅人講師に深く感謝するとともに心から御礼申し上げます．また，計算機Ｃ研究室の皆様に深く感謝いたします．

また本研究は栢森情報科学振興財団の助成を受けて遂行されました．ここに深く感謝いたします．

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, Vol. 20, No. 1, pp. 46–56, 2005.

[2] 石田基広, 金明哲. コーパスとテキストマイニング. 共立出版, 2012.

[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, Vol. 17, No. 4, pp. 398–406, 2002.

[4] 涌井良幸, 涌井貞美. 図解 ベイズ統計学. ナツメ社, 2012.

[5] 松尾豊, 大澤幸, 石塚満. Small world 構造に基づく文書からのキーワード抽出. 人工知能学会論文誌, Vol. 43, No. 6, pp. 1825–1833, 2002.

[6] 内山将夫, 橋田浩一. Gda タグを利用した複数文書の要約. 言語処理学会論文誌, Vol. 6, pp. 376–379, 2000.

[7] 森純一朗, 松尾豊, 石塚満. Web からの人物に関数キーワード抽出. 人工知能学会論文誌, Vol. 20, No. 5, pp. 337–345, 2005.

[8] 岡崎直観, 成澤克麻, 乾健太郎. Web 文書からの人の安全・危険に関わる情報の抽出. 言語処理学会論文誌, Vol. 18, pp. 896–898, 2012.

[9] 小嶋秀樹, 古郡廷治. 単語の結束性にもとづいてテキストを場面に分割する試み. 情報処理学会論文誌, Vol. 95, No. 7, pp. 49–56, 1993.