概要

現在，インターネット上には様々な電子テキストがあり，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．

松尾ら [1] は，Web 上の情報からの人間関係ネットワークを抽出している．大竹ら [2] は，大量の新聞記事から TF-IDF を用いて，事物の関係情報をネットワークとしてまとめたものを構成した．しかし大竹らの構築方法では関連のない事物のノードを含むネットワークへ発展していく問題がある．本研究では，関連のない事物のノードを無関連ノードと呼ぶ．

そこで本研究では，2 つのノード選択方法を新たに提案した．これらの手法により無関連ノードが削除できるかを調査した．実際に「オリンピック」「オウム」「ギリシャショック」「地震」に関するネットワークを構築し，そのネットワークにおいて無関連ノードの削除を確認した．

目 次

表 目 次

図 目 次

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．そこで大竹ら [2] 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し，「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．しかし大竹らの構築方法では，キーワードをオリンピックとした場合を例とすると「オリンピック→メートル→津波」のような関連のない事物のノードを含むネットワークへ発展していく問題がある．本研究では，関連のない事物のノードを無関連ノードと呼ぶ．

そこで本研究では，ノード選択方法の変更により，無関連ノードが削除ができるかを調査した．実際に「オリンピック」「オウム」「ギリシャショック」「地震」に関するネットワークを構築した．結果として，無関連ノードの削除を確認できた．

本論文の構成は以下の通りである．第 2 章では，本研究の関連研究を述べ，第 3 章では，提案手法のベースとなる大竹らの電子テキストから特定のキーワードに基づく関係情報を抽出する方法について説明し，第 4 章では，提案手法の説明を行う．第 5 章では，先行手法と提案手法によるネットワークの構築とその比較を行い，第 6 章では，結果の考察を行う．第 7 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．

2.1　キーワード抽出の関連研究

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．

森ら [4] 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 [5] 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 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．その構造化に基づき必要な情報を抽出している．

小嶋ら [6] 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．

2.2　ネットワークの関連研究

内山ら [7] 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．

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．抽出手法として，氏名の関係性の強さを知るための様々な指標を用いている．

松尾ら [8] 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．

村田ら [9] は，英語品詞間の転換について調べ，自己組織化マップを利用し調査結果の可視化を行った．

鳥澤ら [10] は，Web 上より多様な意味的関係を抽出し ”鳥式改 ” と呼ばれる巨大な意味ネットワークの構築を行った．

2.3　無関連ノード，多義性解消の関連研究

単語の多義性により，無関連なノードが得られることがある．これの対処として以下の論文がある．

村田 [11] らは，多義性のある単語だけで検索すると複数の関連ある検索結果になる問題を，多義性のある単語に分野を足して検索することで検索結果を絞っている．

村田 [12] らは，教師あり機械学習を用いての多義性解消も行っている．

第3章　先行手法

3.1　ネットワーク構築の概要

比較となる先行手法について説明する．先行手法では，大量の新聞記事群のデータ（新聞データ）からネットワークを構築する．以下に手順を示す．

1. 構築したいネットワークの主となる概念を，キーワードとして設定する．
2. 新聞データからキーワードを含む記事を抽出する．
3. 抽出された記事に対し形態素解析を行う．
4. 抽出された記事において，キーワードと関係性の強い5単語を次のノードとする．
5. 次のノードとなった単語を新たなキーワードとして2にもどり，同様の処理を繰り返しネットワークを拡大していく．

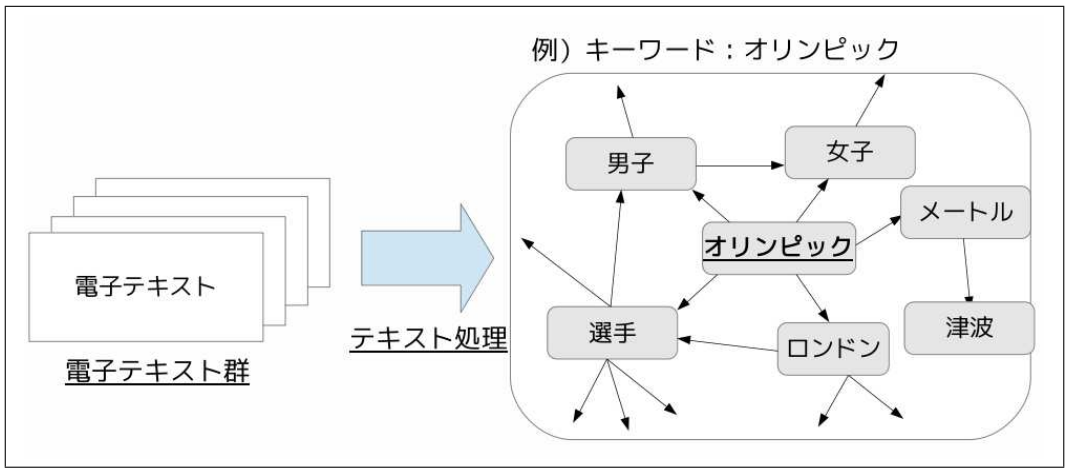

図 3.1: ネットワーク構築の概要

より詳細な構築方法を以下で説明する．

3.2　ノード候補の抽出

キーワードとする単語を単語 k とする．まず単語 k を含んだ記事群を抽出する．抽出された記事群を記事群 S とする．形態素解析を用い記事群 S 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．残った単語をネットワークのノード候補とする．

3.3　形態素解析

日本語の文は英語の文と違い，単語の明確な区切りがない．そのため，ノード候補となる単語を抽出するために，形態素解析と呼ばれる処理を行う必要がある．形態素解析とは，テキストを形態素と呼ばれる単位に分割することである．形態素というのは，厳密には単語と違った分割の単位であるが，おおよそ単語と同じようなものになる．形態素は品詞の情報を持つ．形態素解析結果の例を図 3.2 に示す．

入力：「今年はロンドンで開催されるが２０２０年には東京でオリンピックがある」

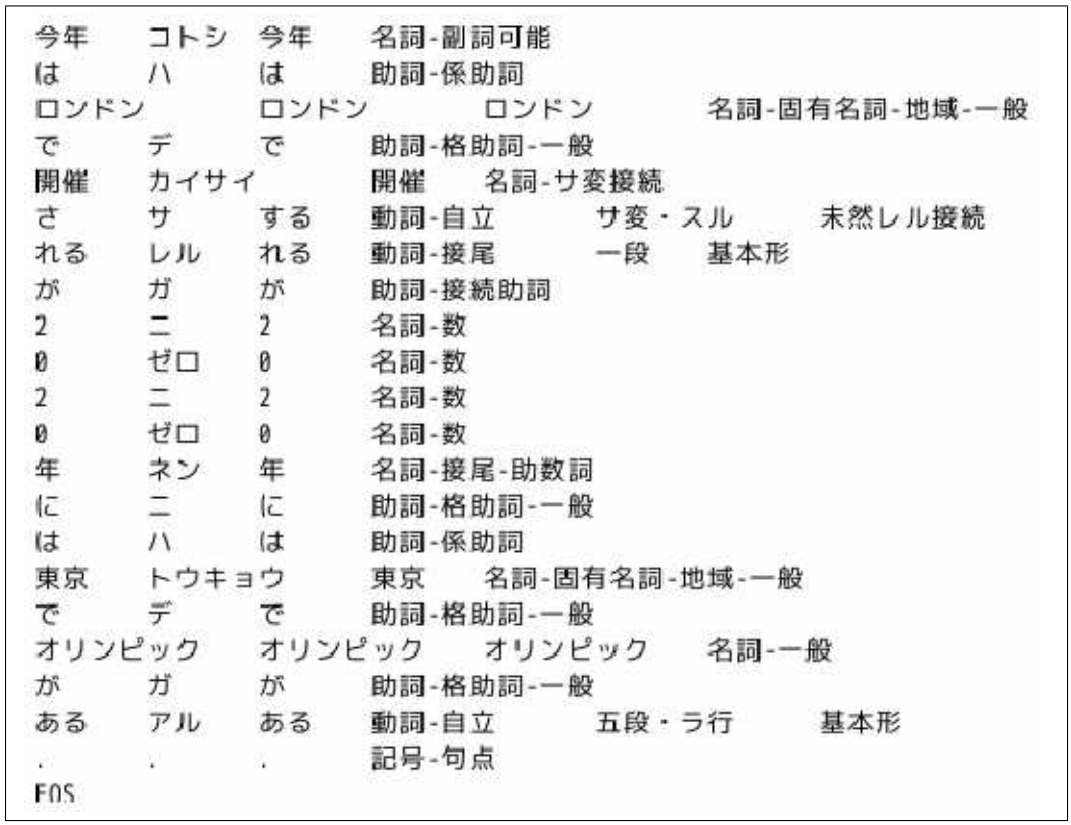

図 3.2: 形態素解析の出力例

このようにして形態素解析によりノード候補になる単語を取り出す．本研究では形態素解析に ChaSen を用いた．

3.4 TF-IDF によるノードの選定

得られたノードの候補の中から，TF-IDF を用いて，実際にノードに用いる単語を選定する．TF-IDF 値の上位 5 単語をキーワードと関係性の強い単語とする．

TF-IDF について説明する．TF-IDF は抽出した記事内におけるノード候補となっている単語の重要度を表す．TF-IDF は以下の式で算出される．

$$TF\text{-}IDF = tf_t * \log \frac{N}{df_t} \tag{3.1}$$

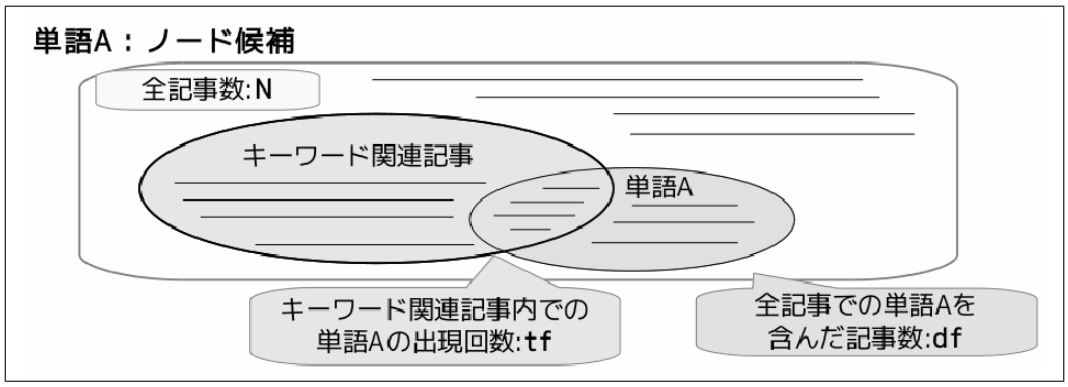

図 3.3: tf と df の関係図

tf_t は抽出された対象テキスト内でのノード候補の単語 t の出現回数，df_t は新聞データ内でのノード候補の単語 t の出現記事数とし，N は新聞データの総記事数とする．この式からどの記事にも現れるような重要度の低い単語については低い重みを，他の記事にあまり現れないような貴重な単語には高い重みを与えることになる．TF-IDF の値が大きいノード候補の単語をネットワーク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とする．

3.5　ネットワークの拡大

単語 k 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 a を新たなキーワードとして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 k 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 つが加わる．同様に各単語からの抽出を繰り返すことでネットワークを拡大していく．ノード抽出の例を図 3.4 に示す.

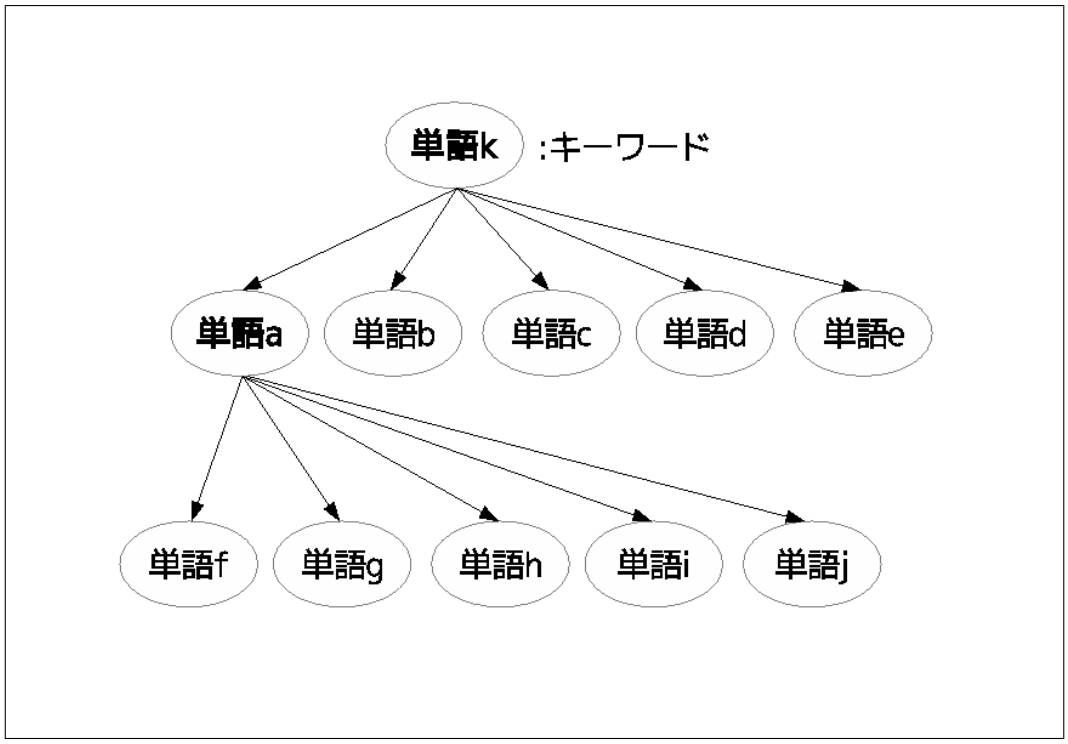

図 3.4: ネットワークのノード抽出の例

第4章　提案手法

4.1　テーマ限定抽出法

先行手法の 3.1 節の 4 の繰り返しにおいては記事を抽出する際に，最初に設定したキーワードと現在のキーワードの両方を含む記事を抽出するようにする．この手法により，最初に設定したキーワードを含む記事から単語を取り出すことになるため，最初に設定したキーワードに関連する単語が得られやすくなり，関連しない単語が取り出されにくくなる．テーマ限定抽出法によるネットワークの構築手順を図 4.1 に示す．

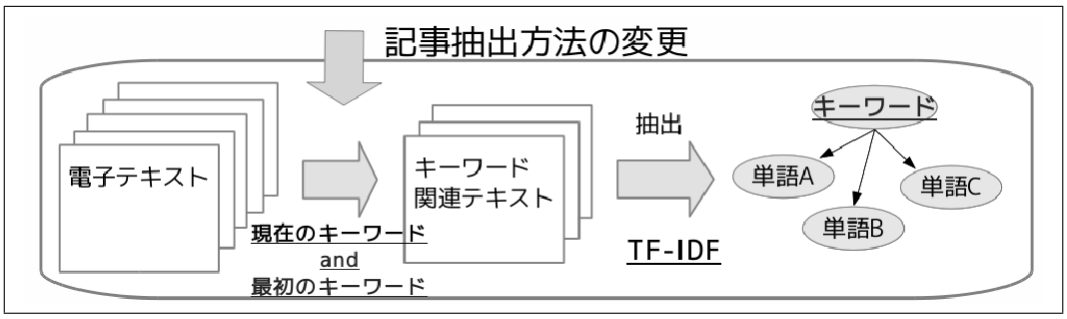

図 4.1: テーマ限定抽出法の概要

例えば図 4.2 の場合，最初に設定したキーワードが「オリンピック」で現在のキーワードが「選手」なので,「オリンピック」と「選手」の両方を含む記事を抽出する．

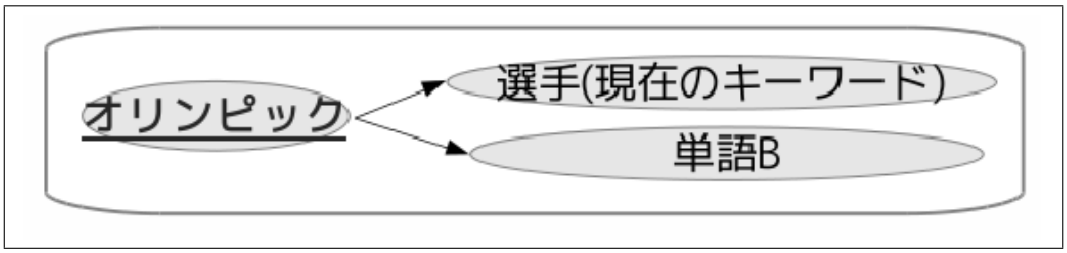

図 4.2: テーマ限定抽出法の例

4.2 テーマ関連抽出法

先行手法の 3.1 節の 3 の手順においてノードを選択する際に，キーワードとの関連度が閾値以下の単語を消去して残ったものから上位 5 単語を選ぶ．この手法により，最初に設定したキーワードと関連度の低い単語を消去できる．テーマ関連抽出法によるネットワークの構築手順を図 4.3 に示す．

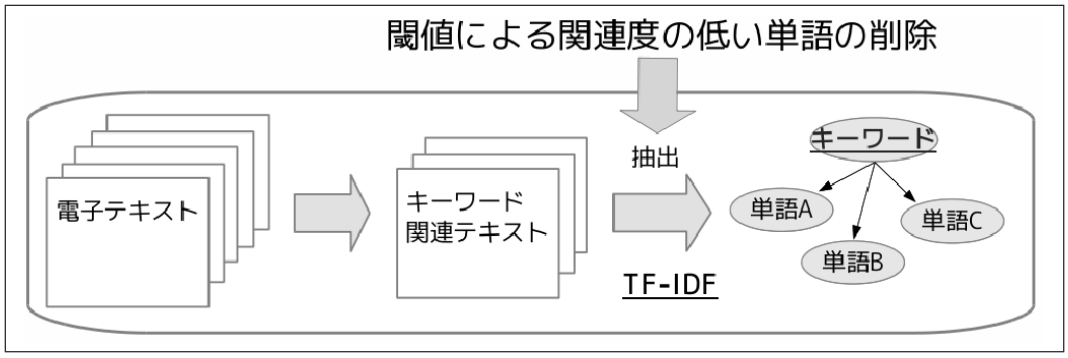

図 4.3: テーマ関連抽出法の概要

本研究で使用する単語 t の関連度 r の計算式は以下の通りである．

$$r(t;k) = \frac{df_{k,\,t}}{df_k} / \frac{df_t}{N} \tag{4.1}$$

k は最初に設定したキーワード, $df_{k,\,t}$は新聞データ内でのキーワード k と単語 t が共に出現した記事数, df_kは新聞データ内でのキーワード k が出現した記事数, df_tと N は 2 章の場合と同じである．$\frac{df_{k,\,t}}{df_k}$は k がある時の t の出現率, $\frac{df_t}{N}$は t の一般的出現率を表している．

本研究では閾値を 1.0 とする．この場合 k がある時の t の出現率が t の一般的出現率よりも小さい場合，最初に設定したキーワードと関連が低い単語とし削除する．

第5章　実験

5.1　先行手法の問題点の確認

本節では事前実験として，先行手法においてどのような箇所で無関連ノードへの発展が起こるかを調べる．

本実験では，キーワードは「オウム」と「オリンピック」とした．実験データには「オウム」の場合，毎日新聞 2011 年の 1 年分の記事 96,630 記事を用い,「オリンピック」の場合，毎日新聞 2012 年の 1 年分の記事 110,639 記事を用いる．出現した無関連ノードを表 5.1, 表 5.2 に示し，構築したネットワークを図 5.1，図 5.2 に示す．

表 5.1:「オウム」のネットワークで出現した無関連ノード

安打，右左，メートル，男子

表 5.2:「オリンピック」のネットワークで出現した無関連ノード

津波，事故，高校，衆院，首相，民主党

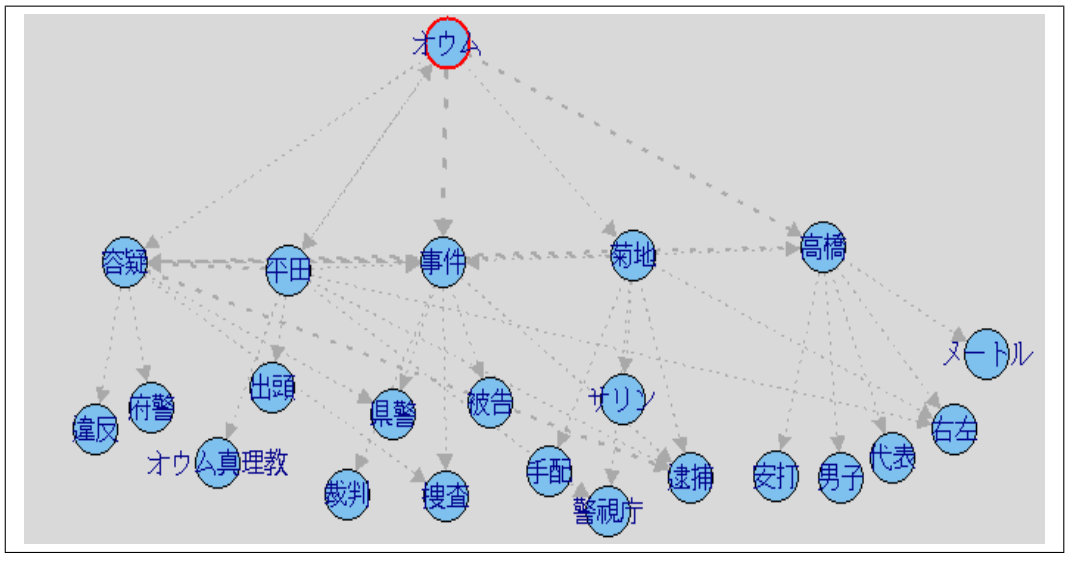

図 5.1: 基本手法で構築した「オウム」のネットワーク

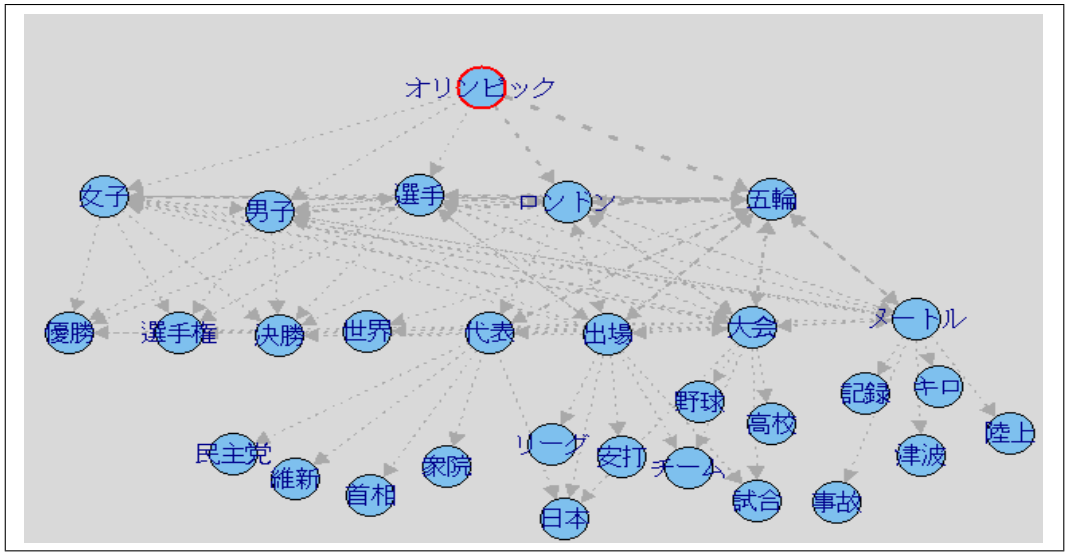

図 5.2: 基本手法で構築した「オリンピック」のネットワーク

「オウム」に関するネットワークにおいて「メートル」「右左」「安打」などの関連しない単語が得られた.「オリンピック」では,「津波」「事故」「維新」などの関連のない単語が得られた.

5.2　提案手法によるネットワークの構築

提案したそれぞれの手法でネットワークの構築を行う．本実験では，テーマ関連抽出法で用いる閾値を 1.0 とした．どのようなキーワードでも提案手法が有効であるかを確認するため，前節で行った「オウム」と「オリンピック」に加え「ギリシャ」と「地震」でも行った．

5.2.1　テーマ限定抽出法によるネットワークの構築

キーワードを「オウム」としてテーマ限定抽出法で構築したネットワーク図を図 5.3 に示す．

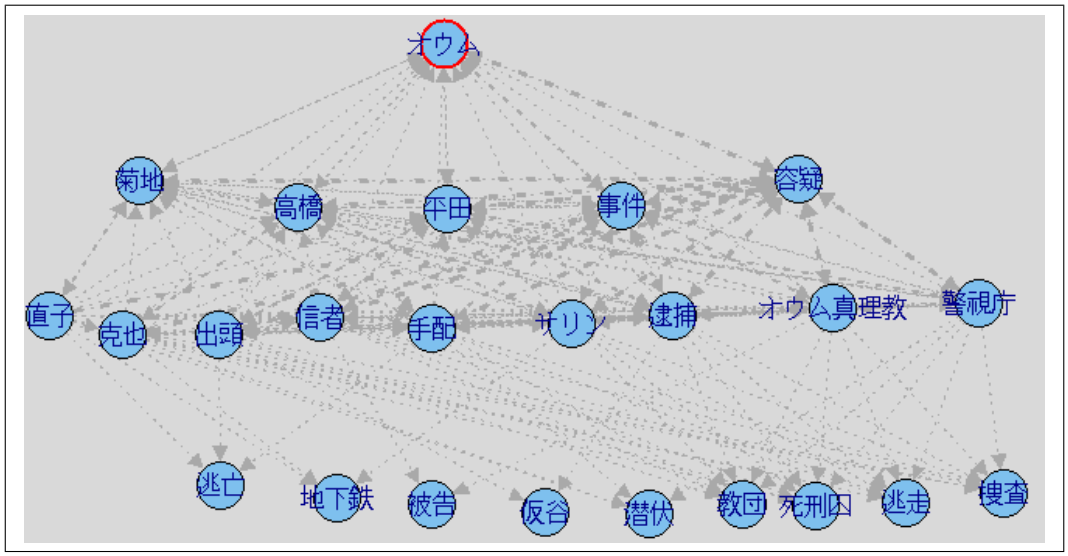

図 5.3: テーマ限定抽出法で構築した「オウム」のネットワーク

キーワードを「オリンピック」としてテーマ限定抽出法で構築したネットワーク図を図 5.4 に示す.

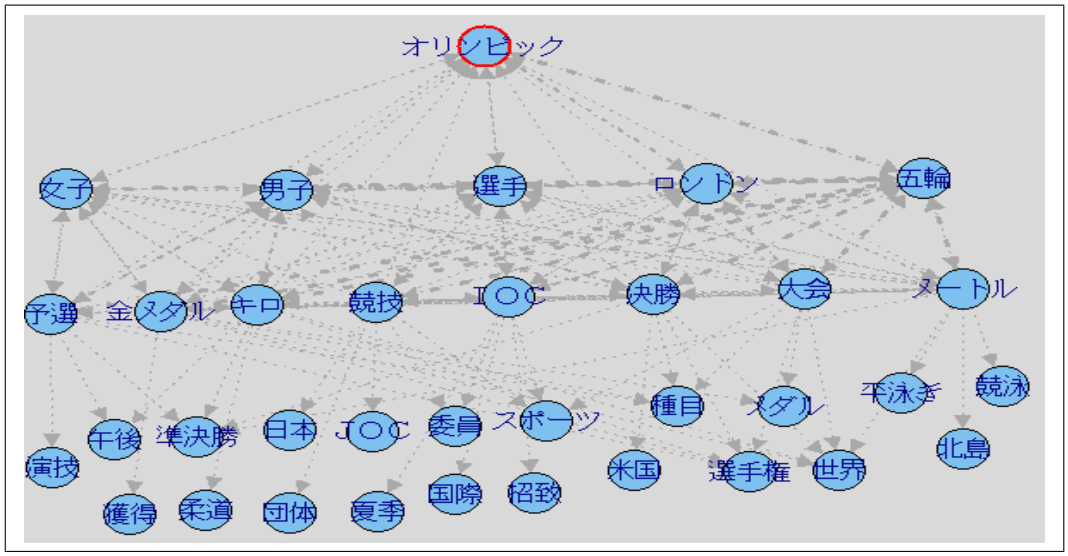

図 5.4: テーマ限定抽出法で構築した「オリンピック」のネットワーク

キーワードを「ギリシャ」としてテーマ限定抽出法で構築したネットワーク図を図 5.5 に示す.

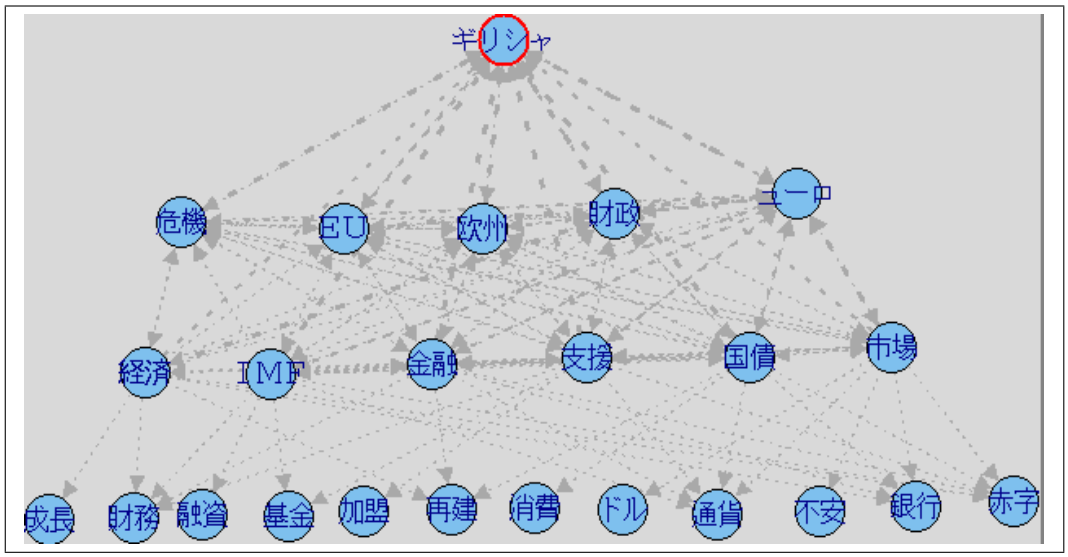

図 5.5: テーマ限定抽出法で構築した「ギリシャ」のネットワーク

キーワードを「地震」としてテーマ限定抽出法で構築したネットワーク図を図 5.6 に示す.

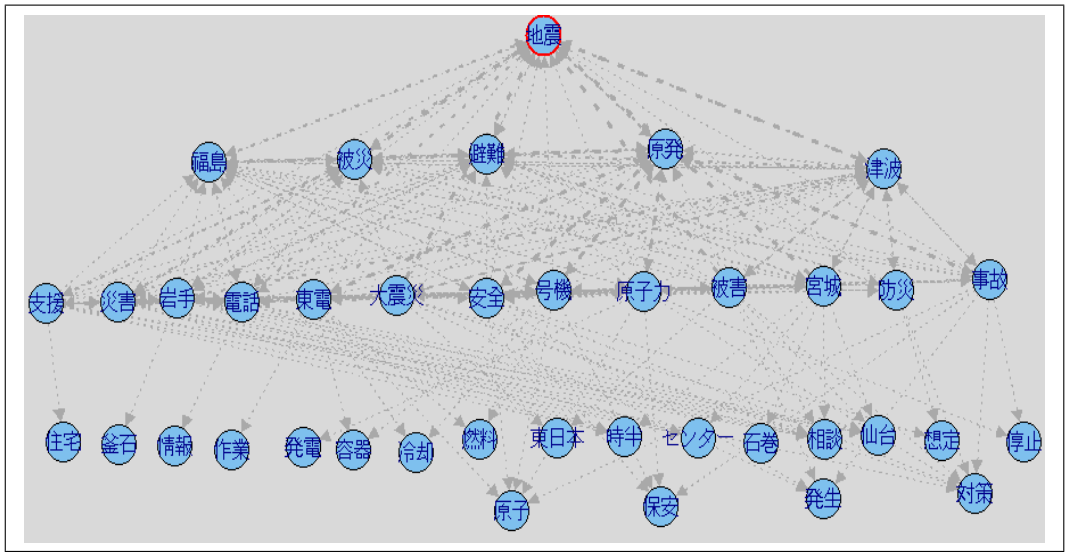

図 5.6: テーマ限定抽出法で構築した「地震」のネットワーク

どのネットワークにおいても無関連ノードの出現は見られなかった．しかしネットワークの広がりが先行手法よりも小さくなった．

5.2.2　テーマ関連抽出法によるネットワークの構築

キーワードを「オウム」としてテーマ関連抽出法で構築したネットワーク図を図 5.7 に示す.

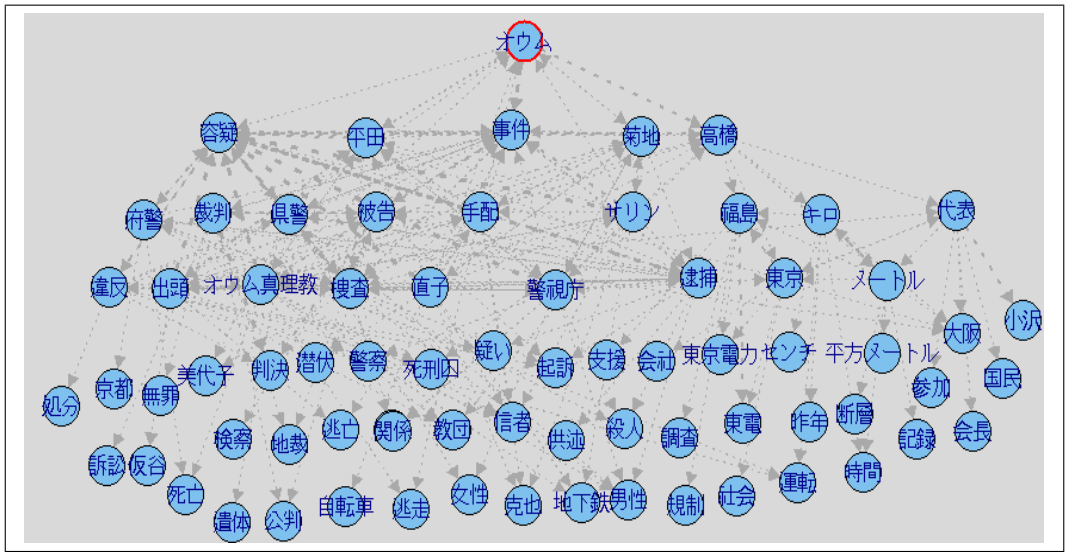

図 5.7: テーマ関連抽出法で構築した「オウム」のネットワーク

キーワードを「オリンピック」としてテーマ関連抽出法で構築したネットワーク図を図 5.8 に示す.

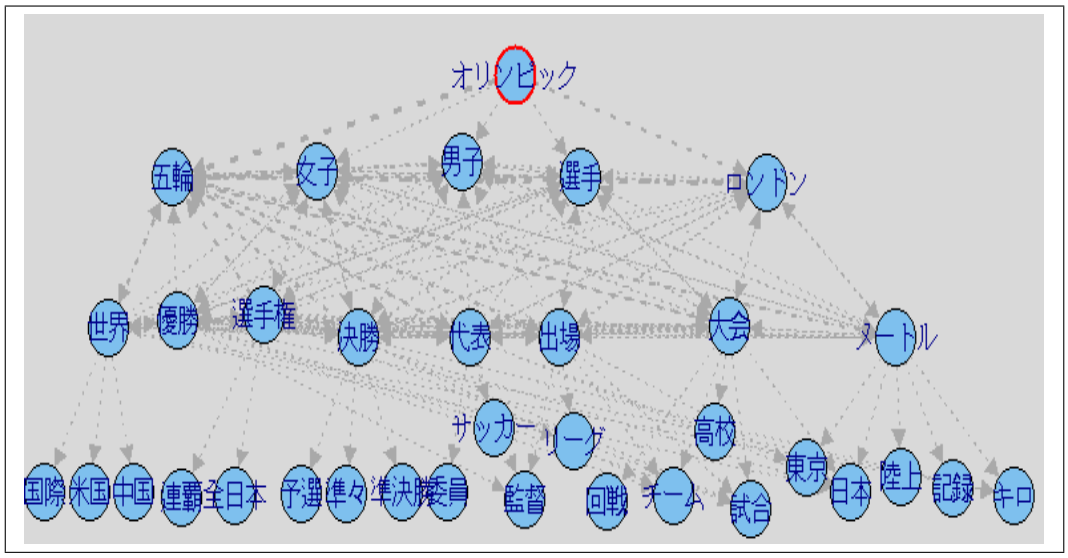

図 5.8: テーマ関連抽出法で構築した「オリンピック」のネットワーク

キーワードを「ギリシャ」としてテーマ関連抽出法で構築したネットワーク図を図 5.9 に示す.

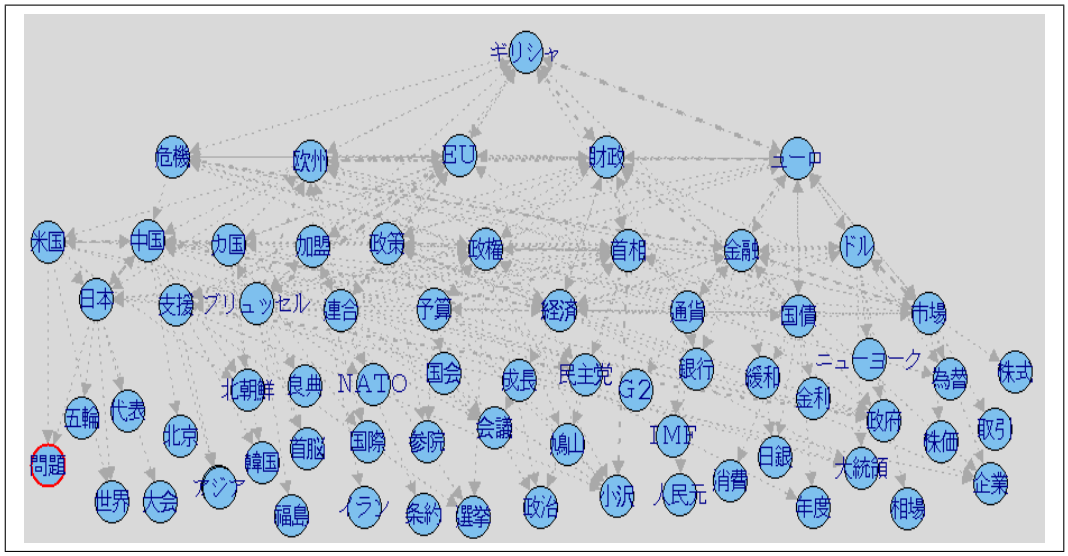

図 5.9: テーマ関連抽出法で構築した「ギリシャ」のネットワーク

キーワードを「地震」としてテーマ関連抽出法で構築したネットワーク図を図 5.10 に示す.

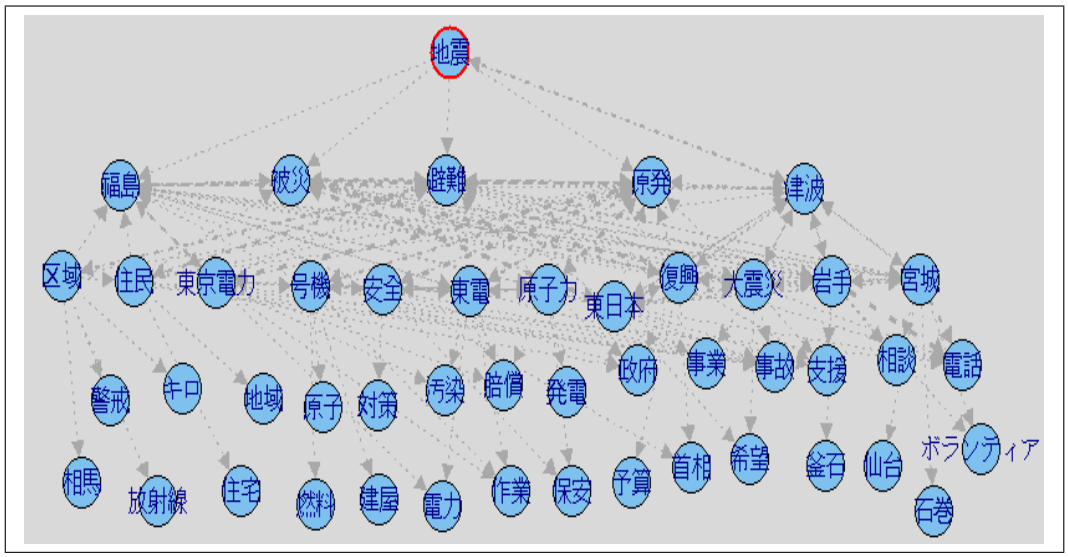

図 5.10: テーマ関連抽出法で構築した「地震」のネットワーク

基本手法で出現した無関連ノードを削除できた．ネットワークの広がり方も変わらなかった.

第6章　考察

6.1　無関連ノードの出現調査

基本手法の問題点について考察を行う．5.1 節にて基本手法によるネットワークの構築を行った.「オウム」「オリンピック」のどちらの場合でも明らかに関連のない単語が出現した.

6.2　テーマ限定抽出法についての考察

5.2.1 節においてテーマ限定抽出法によるネットワークの構築を行った．基本手法で出現した無関連ノードの削除を確認できた．ここでキーワード「オリンピック」の場合の，削除した単語と新たに出現した単語を表 6.1 に示す.

表 6.1: テーマ限定抽出法で削除された単語と新たに出現した単語（オリンピック）

削除された単語	出場，代表，優勝，津波，記録，事故，陸上，試合， 高校，チーム，野球，監督，安打，リーグ，衆院， 首相，維新，民主党，中国，経済，準々，全日本，連覇
新たに出現した単語	ＩＯＣ，競技，金メダル，競泳，北島，平泳ぎ，メダル， 種目，スポーツ，招致，委員，国際，夏季，ＪＯＣ， 団体，柔道，獲得，演技

表 6.1 の「陸上」や「連覇」などの最初に設定したキーワードと関連があるのではないかと考えられる単語も削除してしまっていた．新たに出現した単語は全て関連のある単語であり，ネットワークを見ると関連した単語だけで構成されていると考えられる．しかし抜き出す記事を限定したことでネットワークの広がりも他の手法と比べると小さくなったといえる.

6.3　テーマ関連抽出法についての考察

5.2.2 節においてテーマ関連抽出法によるネットワークの構築を行った．この手法でも基本手法で出現した無関連ノードの削除を確認できた．ここでキーワード「オリンピック」の場合の，削除した単語と新たに出現した単語を表 6.2 に示す．

表 6.2: テーマ関連抽出法で削除された単語と新たに出現した単語（オリンピック）

削除された単語	津波，事故，野球，安打，衆院，首相，維新，民主党，経済
新たに出現した単語	東京，回戦，サッカー，委員，国際

限定抽出法のように関連があるものまで削除する問題は発生しなかった．しかしテーマ関連抽出法では，全ての無関連ノードを削除できていなかった．この問題については閾値の調整で解決できるかと考えられる．テーマ限定抽出法のようにネットワークの広がりが小さくなるということはなかった．

6.4　その他のネットワークにおける考察

この節では先行手法での構築の時点で無関連ノードがなかった例として「地震」について考察する．提案手法により削除された単語と新たに出現した単語を表 6.3 と表 6.4 に示す．

表 6.3: テーマ限定抽出法で削除された単語と新たに出現した単語（地震）

削除された単語	ボランティア，汚染，活動，希望，区域，携帯，経済， 建屋，財源，事業，自治体，首相，住民，政府，増税， 地域，電力，東京，東京電力，賠償，復興，放射線，予算
新たに出現した単語	センター，災害，時半，情報，想定，停止，発生，被害，防災，容器

表 6.4: テーマ関連抽出法で削除された単語と新たに出現した単語（地震）

削除された単語	活動，携帯，経済，財源，自治体，増税，東京，冷却
新たに出現した単語	キロ，警戒，相馬

「増税」,「東京」などの関連がある単語を削除してしまった．しかしテーマ限定抽出法ではネットワークを更に広げていけば出現するのではないかと考えられる．テーマ関連抽出法でも閾値を調整すれば解決できると考えられる．

第7章　今後の課題

今後の課題として以下の２つの評価を行い，提案手法の有効性を確認する．

評価１

被験者に基本手法と提案手法のネットワークを見せ，どちらがよりキーワードに関連したネットワークを表しているかを判定してもらう．

評価２

基本手法において不要と考えられる単語をどのくらい削除できたかと関連があるネットワークを誤っていくつ削除したのかを単語の個数を用いて評価する．

その他の今後の課題として以下が考えられる．

実験１

リンクへの意味の付与を行う．具体的にはノードとなっている単語同士が記事内で同時に出現した際に，二つ単語の間に出現している単語を獲得し，その頻度が一番高いものを付与する．これによりリンクの繋がりがよりわかりやすくなる．

実験２

リンクへ重みを付与して，活性伝播を行い活性値の変化から重要な概念の抽出を行う．

第8章　おわりに

本研究では先行研究の問題点である無関連ノードの扱いについて研究を行った．新聞記事を抽出する方法の変更と閾値を設定し関連の低い単語を削除をすることで無関連ノードを削除するという手法を提案した．実際に大量の新聞データから事物の関係情報をネットワークとして構築した．今回は「オウム」「オリンピック」「ギリシャ」「地震」のどのキーワードにおいても提案手法による無関連ノードの削除を確認できた．テーマ限定抽出法では，関連のある単語のみでネットワークを構築できた．しかし基本手法で出現した関連がある単語まで削除してしまった．テーマ関連抽出法では，全ての無関連ノードを削除できていなかったがネットワークの広がりが小さくならずに無関連ノードを削除できた．

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.

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] 大竹竜太, 村田真樹, 徳久雅人. 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. 言語処理学会第 19 回年次大会発表論文集, pp. 798–801, 2013.

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 村田真樹, 進藤三佳, 馬青, 井佐原均. 単語辞書を用いた英語品詞間の転換に関する調査. 言語処理学会第 9 回年次大会発表論文集, pp. 478–481, 2003.

[10] 鳥澤健太郎, 隅田飛鳥, 野口大輔, 柿澤康範, 風間淳一, Stijn De Saeger, 村田真樹, 黒田航, 山田一郎, 塚脇幸代, 太田公子. ウェブ検索ディレクトリの自動構築とその改良 -鳥式改-. 言語処理学会第 15 回年次大会発表論文集, pp. 369–372, 2009.

[11] 村田真樹, 土井晃一, 三森智裕, 福田安志. 多義語による情報検索装置及びプログラム. 特許第 4857448 号, 2012.

[12] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均. SENSEVAL2J 辞書タスクでの CRL の取り組み-日本語単語の多義性解消における種々の機械学習手法と素性の比較-. 自然言語処理. Vol. 10, No. 3, pp. 115–134, 2003.

[13] 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.

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