2020年度（令和2年度）　卒業論文

Word2vecで作成した単語集合を用いた概念ネットワークの改良

電気情報系学科 卒業論文検印	
学科長	

指導教員

村田真樹

村上仁一

鳥取大学工学部　電気情報系学科

自然言語処理研究室

B17T2096A 本田　涼太

概要

近年，電子テキストが増加しており，大量の電子テキストの中から有用な情報を取り出す技術が求められている．

大竹ら [1] は，TF-IDF 法を用いて概念ネットワークの構築手法を提案した．土遠ら [2] は，概念ネットワークに出現した単語にテーマキーワードと無関係な単語があることに着目し，これら無関係な単語を出現させないために,「テーマ限定抽出法」を提案した．上東ら [3] は，検索エンジンを用いて概念ネットワークを構築することで，より多くのテーマキーワードで十分な情報量のネットワークを構築した．

しかし，これらは，単に TF-IDF 値が大きい単語を取り出してネットワークを構築しているため，よく似た内容の単語であっても離れて出現することがあった．

そこで本研究では，この概念ネットワークの構築において，Word2vec[4] を用いてある単語から出現する単語を同種の単語が出やすくなるようにする．そのようにすることでネットワークをより見やすくするように改良する．本研究の目的は，ネットワークの構築において出現する単語を同種の単語が出やすくなるようにし，より見やすいネットワークを構築することである．

実際にネットワークの構築において出現する単語を同種の単語が出やすくなるようにしたところ，1 ネットワークあたりの役に立つ単語の個数は，従来手法が 3.2 個に対して，TF-IDF 合計値法が 3.3 個，TF-IDF 最大値法が 3.1 個と，従来手法と比べても情報量が減少することを抑えた．また，1 ネットワークあたりの見やすい部分の個数は，従来手法が 1.2 個に対して，TF-IDF 合計値法が 3.1 個，TF-IDF 最大値法が 2.6 個と，似た意味の単語が並んで見やすくなっている部分は増えた．

目 次

表 目 次

図 目 次

第1章　はじめに

近年，インターネットの普及等により電子テキストが増加している．これら大量の電子テキストから有用な情報を効率的に取り出す技術が求められている．そこで言語テキスト処理技術を用いテーマキーワードとなる単語を入力することで，電子テキストや新聞データ等のメディアから入力単語の概念にかかわる概要情報を抜き出し概念ネットワークの研究が進められた．本研究で改良を行う概念ネットワークは，単語の上位下位関係を木構造で表示するシソーラスとは異なるものである．

これまでのネットワークの研究で，松尾ら [5][6] は Web 上の情報からどのような人間関係があるかを示した人間関係ネットワークの構築を行った．概念ネットワークの構築に際して，大竹ら [1] は，TF-IDF 法を用いて概念ネットワークの構築手法を提案した．また，土遠ら [2] は，概念ネットワークに出現した単語にテーマキーワードと無関係な単語があることに着目し，これら無関係な単語を出現させないために,「テーマ限定抽出法」を提案した．上東ら [3] は，検索エンジンを用いて概念ネットワークを構築することで，より多くのテーマキーワードで十分な情報量のネットワークを構築した．

しかし，これまでの研究では，関連する単語を概念ネットワークとして表示する際に，単に TF-IDF 値が大きい単語を取り出してネットワークを構築しているため，よく似た内容の単語であっても離れて出現することがあった．

そこで本研究では，この概念ネットワークの構築において，Word2vec[4] を用いてある単語から発展するネットワークの単語を同種の単語が出やすくなるようにする．そのようにすることでネットワークをより見やすくするように改良する．本研究の目的は，ネットワークの構築において出現する単語を同種の単語が出やすくなるようにし，より見やすいネットワークを構築することである．

本研究の主な主張点を以下に整理する．

- 概念ネットワークの構築に際して，Word2vec を用いて同種の単語が出やすくなるようにするという点が新規であり，特にネットワークの見やすさという点に着目した．

- 従来手法と 2 種類の提案手法を用いて構築したネットワークを比較した結果，見やすい箇所の平均が従来手法が 1.2 個であったのに対して，TF-IDF 合計値では 3.1 個，TF-IDF 最大値では 2.6 個といずれも従来手法を上回った．

本論文の構成は以下の通りである．第 2 章では，本研究に関連する研究としてどのような研究が行われてきたかを記述し，その研究と本研究との関連を説明する．第 3 章では，提案手法について説明を行う．第 4 章では，本研究が行った実験についての説明と，その結果について記述する．第 5 章では，第 4 章の結果について考察を行う．第 6 章では，今後の課題について記述する．第 7 章では，まとめを行う．

第2章　先行手法

本章では，これまでの概念ネットワークに関する研究について記述する．2.1 節では，松尾ら [5][6] が行った Web 上の情報から人間関係ネットワークを構築する研究と堀田ら [7] が行った人間関係ネットワークを用いた情報推薦システムの研究について記述する．2.2 節では，大竹ら [1] が行った概念ネットワーク構築手法の研究について記述する．2.3 節では，2.2 節のネットワーク構築手法に用いられた TF-IDF 法について記述する．2.4 節では，土遠ら [2] のネットワークに出現する無関係ノードの扱いの研究について記述する．2.5 節では，窪ら [8] のネットワークのリンクに文字列を付与した研究について記述する．2.6 節では，上東ら [3] のネットワークを構築する際に，検索エンジンを用いてネットワークの構築を行った研究について記述する．2.7 節では，南ら [9] の単一文書でネットワークを構築した研究について記述する．2.8 節では，畑山ら [10] の重要語句を抽出して新聞記事の自動要約を行う研究について記述する．2.9 節では，岡田ら [11] と Takama ら [12] の拡張したベクトル空間モデルを用いたオンライン上でのテキストマイニングの研究について記述する．2.10 節では，松本ら [13] の深層学習を用いた新聞記事分析による市場動向予測の研究について記述する．

2.1　人間関係ネットワークの構築とその利用

松尾ら [5][6] は，Web 上の同じページにどれだけ氏名が共起しているかということから人物同士がどのような関係であるかを視覚的に示す人間関係ネットワークを構築した．また，このネットワークには共著関係であるかや同じ研究室に所属しているかといった情報をラベルとして付与することで，より人物同士の関係性を明確に表すようにした．

さらに，堀田ら [7] は，人間関係ネットワークを用いて Web 上の広告配信システムとして情報推薦システムを実装し，ランダムな広告配信と比較し，1.9 倍の広告効果が得られたとしている．

2.2 概念ネットワークの構築方法

大竹ら [1] が提案したネットワークの構築手法を述べる．

手順 1 構築したいネットワークの主となる単語をテーマキーワードとして設定する．

手順 2 キーワードとなる単語を含んだ記事群を抽出し，その記事群から形態素解析を用いて名詞のみを抽出する．その際，一文字，ひらがなのみ，数字のみの単語を除外する．

手順 3 手順 2 で抽出された単語の出現頻度を調べ，上位 100 単語をノード候補とする．

手順 4 得られたノード候補の中から，TF-IDF 法を用い，値の大きな上位 5 単語をネットワークのノードとして選定する．TF-IDF 法については 2.3 節にて述べる．

手順 5 単語間の関係に重みを付与し，単語間の関連の強さに差をつける．

手順 6 手順 2 から手順 5 を繰り返して概念ネットワークを拡張する．

2.3 TF-IDF 法

ネットワークの構築において，ノードを選定する際に利用した TF-IDF 法について述べる．この節では入力データの電子テキストを新聞データとして説明する．TF とは単語頻度 (Term Frequency) のことであり，入力データにおいて, 単語 t が出現した頻度のことをいう．また，DF は文書頻度 (Document Frequency) のことであり，単語 t がある記事群 A において出現した記事の数のことをいう．N を記事群 A の総記事数として，TF-IDF 法を用いたノードの選定式を式 (2.1) に示す．

$$w = tf * log(\frac{N}{df}) \tag{2.1}$$

2.4 無関係ノードの扱い

土遠ら [2] は，大竹らの構築したネットワークにテーマ限定抽出法を導入した．テーマ限定抽出法とは，2.2 節の手順 2 において，記事を抽出する際に，テーマ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したものである．そうすることにより，テーマキーワードと関連性のない，または，関連性が薄いと考えられる単語が取り出されにくくなる．

2.5　リンクへの文字列の付与

窪ら [8] は，ネットワークのリンク，すなわち出現単語の間に文字列を付与した．従来ネットワークには，単語のみを表示していたが，ネットワークを構築する際に新聞記事群のデータから単語間の関係を表す文字列を抽出し，抽出した文字列を単語間に付与した．その結果，より詳細な単語間の関係を得られ，関係が分かりづらい単語同士についてもその関係性を確認することができた．

2.6　検索エンジンを利用したネットワーク構築

上東ら [3] は，ネットワークを構築する際に，検索エンジンを利用した．従来は，新聞データからネットワークを構築していたが，新聞から得られる情報が不足しているため，情報量が不十分なネットワークが構築されたが，検索エンジンを用いてネットワークを構築することで，以前より多くのテーマキーワードで十分な情報量のネットワークを得られた．

2.7　ネットワークを利用した単一文書の可視化

南ら [9] は，ネットワーク構築システムを利用して新聞記事や小説などの単一文書の可視化を行った．その結果，新聞記事においては，ネットワークを利用することが，記事の内容の把握に役立つことが分かった．また，小説では，登場人物や事柄などについてネットワークにノードとして出現している単語の出現段落の推定が可能になり，実験で用いた全ての小説で，登場人物の特徴や 2 人の登場人物の関係性といった物語にとって有益な情報を獲得することができた．

2.8　新聞記事の自動要約

畑山ら [10] は，格フレーム辞書を用いて新聞記事から語句単位で重要語句を抽出し，それらを用いて新聞記事の要約の自動生成を行った．この際，文生成に必要な主語，述語，目的語を特定するために格フレーム辞書を用いた．その結果，必要最低限の重要語句を抽出することで，人手で作成された要約文に匹敵する要約文を自動生成することができた．

2.9　オンライン上でのテキストマイニングの研究

岡田ら [11] と Takama ら [12] はメタキーワードを用いた拡張ベクトル空間モデル (M2VSM) をテキストマイニングに利用した．従来のベクトル空間モデルでは，文書のクラスタリング，つまり，文書群の分類を行う際，用語のクラスタの粒度を調整することが困難であった．そこで，索引語となる名詞に加えて，名詞を修飾する形容詞や副詞といったメタキーワードを抽出し，類似度計算においてこれらを考慮するようにベクトル空間を拡張した．そのようにすることで，クラスタの粒度を調節でき，文書群をより詳細な内容ごとに分類することを可能にした．そのクラスタリングの結果を，オンラインのテキストマイニングツールに利用することで，大量のテキストを処理できるようにした．

2.10　深層学習を用いた新聞記事の分析

松本ら [13] は，新聞記事を分析し金融市場の動きを予測するために，新聞記事を時系列データととらえ，予測対象日の前営業日の夕刊と，予測対象日当日の朝刊のテキストの差分を分析し，深層学習を用いて株価が上昇したか下落したかを予測した．その結果，従来用いられていたサポートベクトルマシンによる分析よりも高い精度で株価の上昇，下落を予測することができた．

第3章　提案手法

本章では，提案手法の説明を記述する．3.1 節では，ネットワークの構築において出現する単語を同種の単語が出やすくなるようにするために用いた Word2vec[4] について記述する．3.2 節では，提案手法について記述する．

3.1　似た意味を持つ単語集合の作成

ネットワークの構築において出現する単語を同種の単語が出やすくなるようにするために，Google 社が開発した Word2vec[4] 内にある「単語のクラスタリング」を利用して，似た意味を持つ単語の集合 (クラスタ) を作成する．

単語のクラスタリングとは，Word2vec にテキストデータを学習させ，単語をベクトル化する．そのベクトルのコサイン類似度を求め類似度の高い単語をまとめて単語のクラスタを作り，各クラスタにクラスタ番号を割り当てるものである．このクラスタ番号が一致している単語群を似た意味を持つ単語とする．

単語クラスタリングの一例を図 3.1 に示す.「ビデオカメラ」と「レンズ」がクラスタ番号 2455,「パソコン」と「スマートフォン」がクラスタ番号 2423,「市場」がクラスタ番号 2703 でまとまっている．

図 3.1: クラスタリングの一例

3.2 概念ネットワーク構築の提案手法

先行研究では，単に TF-IDF 値の大きい順に単語をネットワークに表示させていたため，よく似た単語であっても離れて出現することがあった．この一例を図 3.2 に示す．図 3.2 では下線で印をつけた,「端末」「機器」「ディスプレー」「測定機」といった似た意味を持つ単語が離れて出現して，情報のまとまりがつかみづらいネットワークになっている．

そこで本研究では，2.2 節で述べた従来のネットワーク構築法に，2.4 節で述べたテーマ限定抽出法を導入したものに，同種の単語が出やすくなるようにした概念ネットワークの構築方法を提案する．以下にその手順を示す．

手順 1 Word2vec の単語のクラスタリング機能を用いて，単語をクラスタ番号ごとにまとめる．

手順 2 2.2 節の手順 2 を行った後，すでにネットワークに出現している単語を除外する．

手順 3 2.2 節の手順 3 と同様の作業を行い，得られたノード候補の単語の TF-IDF 値を計算する．

手順 4 TF-IDF 値を計算した後に TF-IDF 値の大きい順に単語を並べ，各単語のクラスタ番号を取り出し，クラスタ番号ごとに単語をまとめる．

手順 5 クラスタ番号が同じ単語ごとに TF-IDF 値を計算し，TF-IDF 値が大きい順にクラスタ番号を並べる．

手順 6 手順 5 で求めた TF-IDF 値が上位 5 位までのクラスタ番号を持つ単語を抜き出し，上位のクラスタ番号に所属する単語から順に 5 個までネットワークに表示させる．

このうち，手順 5 で述べた TF-IDF 値の計算方法として，TF-IDF 合計値法と TF-IDF 最大値法の 2 通りを提案する．TF-IDF 合計値法は，クラスタ番号が同じ単語ごとにそれらの単語の TF-IDF 値を足し，その合計値上位 5 位までのクラスタ番号に所属する単語をネットワークに表示させる方法である．TF-IDF 最大値法は，クラスタ番号が同じ単語ごとにそれらの単語の TF-IDF 値の最大の値を探し，その最大値上位 5 位までのクラスタ番号に所属する単語をネットワークに表示させる方法である．

この 2 つの手法の計算例をノード候補の単語とその TF-IDF 値が図 3.3 のとおりであると仮定して説明する．

このとき，TF-IDF 合計値法の場合では，クラスタ番号 2455 が 22.1，クラスタ番号 2423 が 21.3，クラスタ番号 2703 が 9.7 となり，「ビデオカメラ」「レンズ」「スマートフォン」「パソコン」「市場」の順番に表示される．一方，TF-IDF 最大値法の場合では，クラスタ番号 2455 が 13.5，クラスタ番号 2423 が 17.7，クラスタ番号 2703 が 9.7 となり，「スマートフォン」「パソコン」「ビデオカメラ」「レンズ」「市場」の順番に表示される．

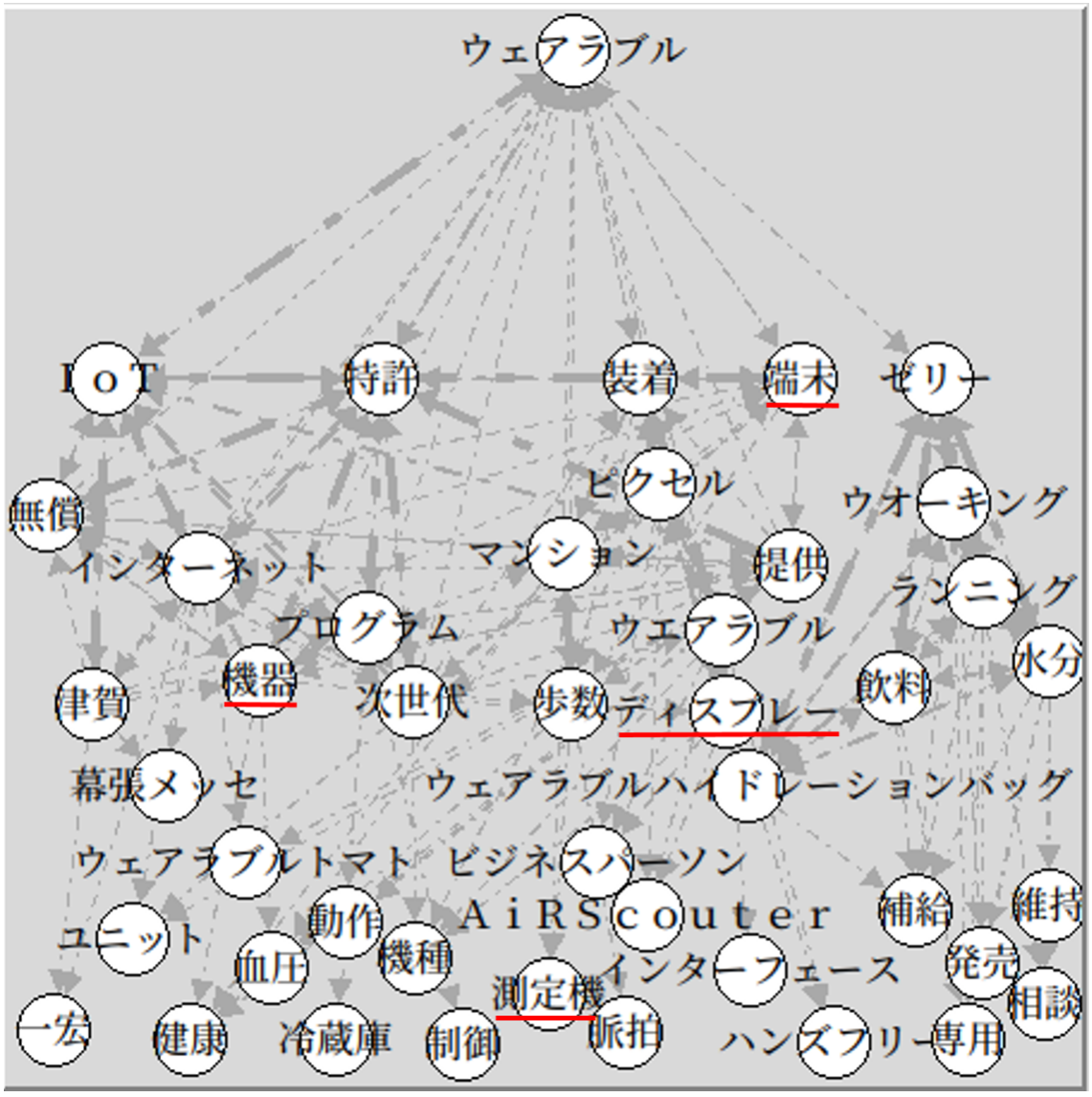


図 3.2: よく似た単語が離れて出現しているネットワークの例

図 3.3: ノード候補の一例

第4章　実験

4.1　著者によるネットワークの評価

3.2 節で述べた提案手法が有用性，見やすさの双方の観点で有効性があるかを確認する．2.2 節で述べた構築手法に，2.4 節で述べたテーマ限定抽出法を加えた従来手法と，従来手法に 3.2 節で述べた手法を加えた 2 通りの提案手法を比較する．それぞれの手法で 20 個のテーマキーワードでネットワークを構築し，後述する評価方法に従って評価する．

4.1.1　実験方法

まず，ネットワークの構築において出現する単語を同種の単語が出やすくなるようにするために，Word2vec に学習させるデータとして，毎日新聞 12 年分 (2007〜2018) のデータ (1,166,761 記事) を使用した．単語のクラスタリングを行う際，クラスタ数は 5,000 とした．ネットワークを構築する際のデータとして 2018 年の毎日新聞の記事 (88,032 記事) を使用した．また，ネットワーク構築に使用した 20 個のテーマキーワードを以下に示す．

- 5G，がん，イギリス，オリンピック，パソコン，ロボット，安全保障，遺跡，宇宙，映画，感染症，京都，銀河，産業構造，寺院，世界遺産，石油，台風，独立，廃線

4.1.2　評価方法

ネットワークについて次の 2 つの観点から評価した．

有用性

ネットワークに出現した単語がテーマキーワードの概念を理解するのに役に立つかという観点で評価した．具体的な評価点を以下に示す．

- 出現した単語について，あまり知らなかった事柄を知れた場合や，キーワードの概念を知るうえで役に立つと判断した場合．
- それぞれが知っている単語であっても，ノードのリンクによって新たな情報が得られる場合，意外な関係性である場合．

見やすさ

ネットワークの見やすい部分があるかという観点で評価した．具体的な評価点を以下に示す．

- 似た意味の単語が並んで出現していることにより，見やすくなっている場合．情報がまとまっていると考えられる場合．
- 似た意味の単語が並んでいることにより，知らなかった単語でも，web 検索等をして並んでいる単語が同じような意味であると判断した場合．

4.1.3 実験結果

従来手法と提案手法で構築したネットワークの評価について，有用性の観点から評価した結果を表 4.1 に示す．また，見やすさの観点から評価した結果を表 4.2 に示す．表 4.3 に本実験で用いたネットワークのテーマキーワードを除いた単語数を示す．

また，各手法ごとに有意差を調べるため両側検定の t 検定を 20 対のデータで行った．20 個のテーマキーワードにおいて，各手法で役に立つと判断した単語の個数と，見やすくなっていると判断した部分の個数を比較した．ここで，有意水準は 5%である．p 値を表 4.4 に示す．

表 4.1: 各テーマキーワードと役に立つ単語の個数

テーマキーワード	従来手法	TF-IDF 合計値法	TF-IDF 最大値法
5G	6	5	1
がん	2	1	4
イギリス	1	0	1
オリンピック	1	2	1
パソコン	3	1	3
ロボット	3	4	4
安全保障	3	2	2
遺跡	6	5	5
宇宙	4	9	9
映画	4	9	6
感染症	2	3	5
京都	0	0	0
銀河	5	5	5
産業構造	5	3	3
寺院	2	3	2
世界遺産	8	6	4
石油	1	1	1
台風	1	2	2
独立	5	2	2
廃線	1	2	1
平均値	3.2	3.3	3.1

表 4.2: 各テーマキーワードと似た意味の単語が並んで、見やすくなっている部分の数

テーマキーワード	従来手法	TF-IDF 合計値法	TF-IDF 最大値法
5G	1	2	0
がん	0	0	0
イギリス	1	5	5
オリンピック	1	3	2
パソコン	0	1	1
ロボット	0	3	2
安全保障	1	4	4
遺跡	1	5	5
宇宙	2	4	2
映画	0	1	1
感染症	1	3	3
京都	0	0	2
銀河	1	1	1
産業構造	1	4	1
寺院	3	6	4
世界遺産	1	3	3
石油	1	6	7
台風	1	2	1
独立	4	5	6
廃線	3	3	1
平均値	1.2	3.1	2.6

表 4.3: ネットワークの単語数

テーマキーワード	従来手法	TF-IDF 合計値法	TF-IDF 最大値法
5G	32	40	20
がん	75	42	52
イギリス	56	64	60
オリンピック	31	42	34
パソコン	48	44	59
ロボット	56	34	48
安全保障	28	35	25
遺跡	39	46	48
宇宙	61	77	58
映画	39	48	49
感染症	43	48	53
京都	65	18	55
銀河	55	29	29
産業構造	56	48	40
寺院	65	89	63
世界遺産	54	48	44
石油	42	46	50
台風	30	44	40
独立	67	69	57
廃線	30	36	31
平均値	49	47	46

表 4.4: 著者による評価における有意差検定

	従来手法と合計値法	従来手法と最大値法	合計値法と最大値法
役に立つ単語の個数	0.83	0.85	0.58
見やすい部分の個数	0.00002	0.003	0.09

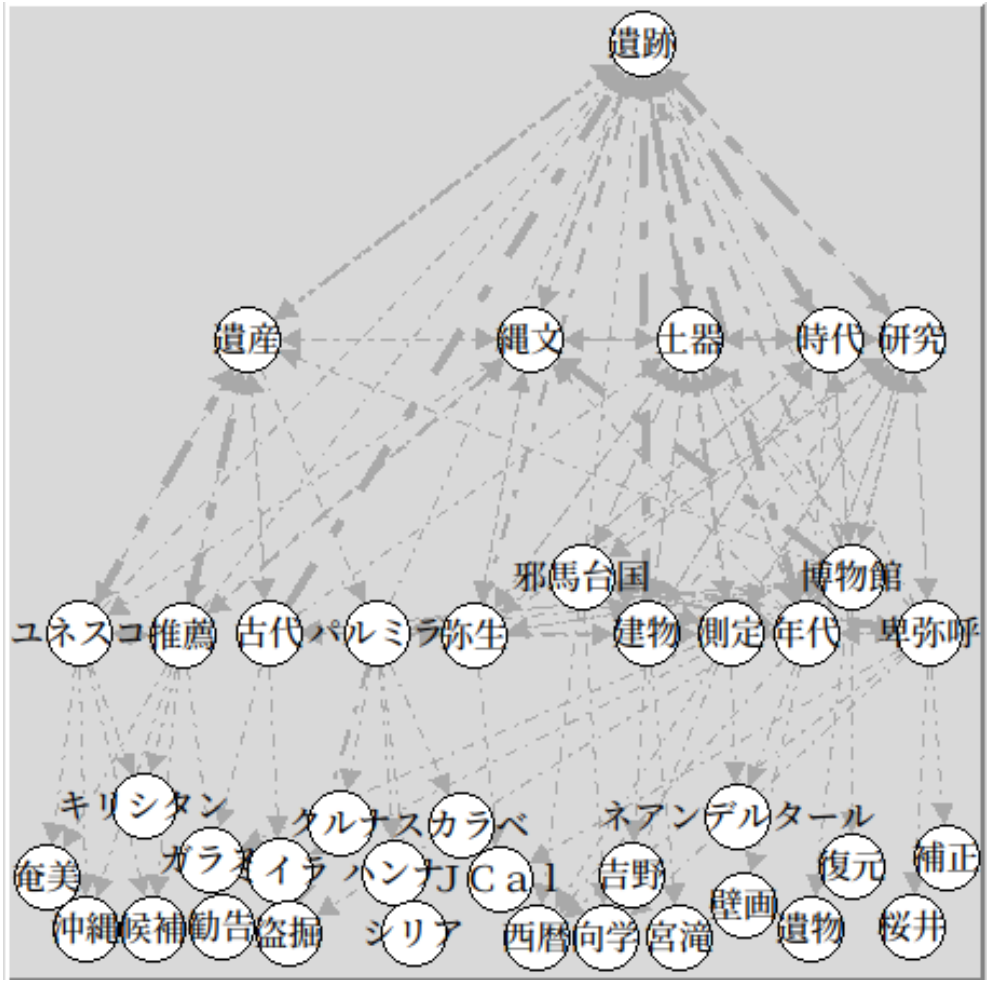
表 4.1 より，役に立つ単語の個数の平均は，従来手法の 3.2 個に対して，TF-IDF 合計値法が 3.3 個，TF-IDF 最大値法は 3.1 個とほぼ同数であった．また，表 4.4 からも役に立つ単語の個数については各手法間で有意差がないことが分かる．表 4.2 より，見やすい部分の個数の平均は，従来手法が 1.2 個であるのに対して，TF-IDF 合計値法が 3.1 個，TF-IDF 最大値法が 2.6 個といずれも上回った．さらに，表 4.4 より，見やすい部分の個数については従来手法と TF-IDF 合計値法の間，従来手法と TF-IDF 最大値法の間で有意差があった．この結果より，TF-IDF 合計値法が見やすさの観点からもっともよい方法であると考える．

本実験の評価方法では役に立つ単語の個数と見やすくなった部分の個数を数えているため，ネットワークの出現単語数が多いほど有利になると考えた，そこで，各ネットワークに出現している単語数を数えた．表 4.3 より，各手法で構築したネットワークの出現単語の平均は，従来手法が 49 個，TF-IDF 合計値法が 47 個，TF-IDF 最大値法が 46 個と従来手法が最も多かったが，ほとんど同じ条件で実験が行えていると考える．

また，評価の一例としてテーマキーワードを「遺跡」として構築したネットワークを図 4.1 から図 4.3 に示す．従来手法によるネットワークを図 4.1 に，TF-IDF 合計値法を用いたネットワークを図 4.2 に，TF-IDF 最大値法を用いたネットワークを図 4.3 にそれぞれ示す．また，それぞれの図の下にそのネットワークにおいて役に立つ単語と，似た意味の単語が並び見やすくなっている部分を列挙している．

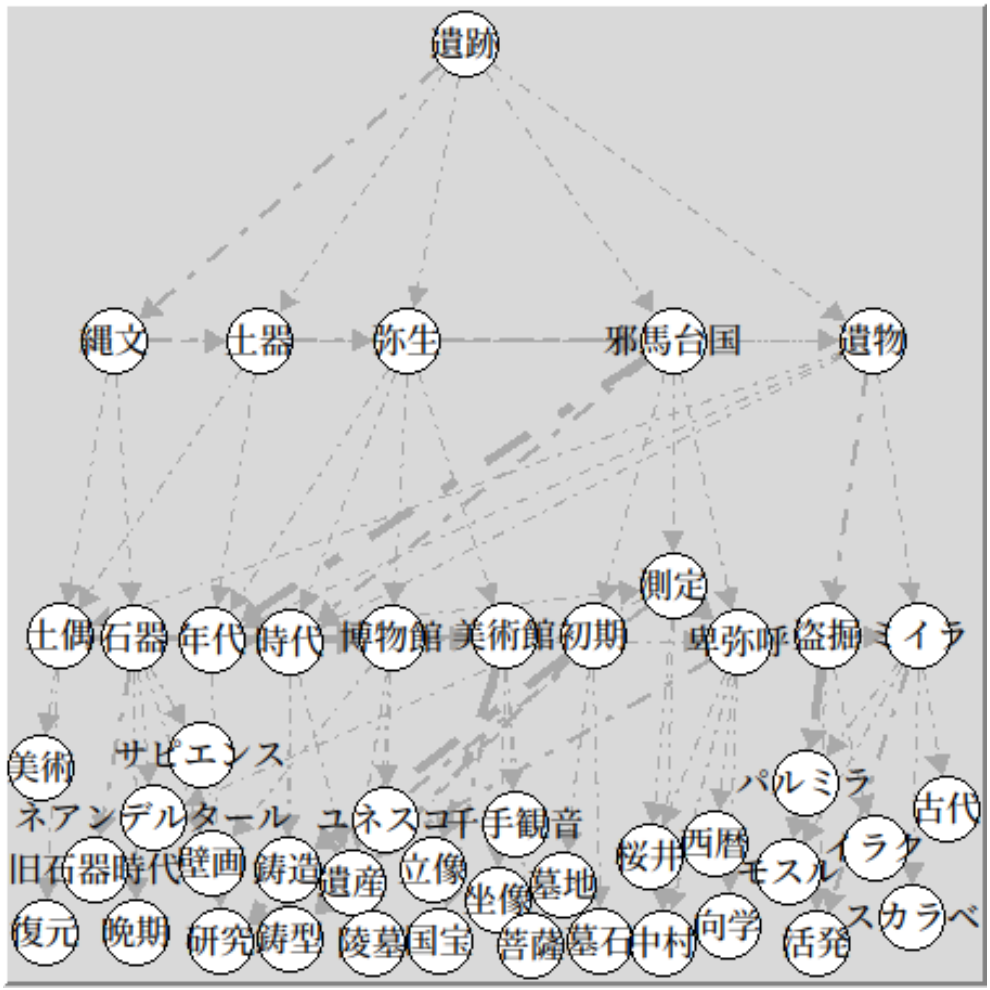
図 4.1，図 4.2，図 4.3 より，役に立つと判断した単語は「パルミラ」や「モスル」といった遺跡のある町や,「ネアンデルタール」や「邪馬台国」といったテーマキーワードである遺跡に関連する語句がどの手法でも表示されていることが分かる．

一方，見やすくなったと判断した部分は，従来手法が「奄美と沖縄」という地名の情報のみであったのに対して，TF-IDF 合計値法と TF-IDF 最大値法では,「博物館と美術館」や「ネアンデルタールとサピエンス」といった地名以外の情報のまとまりを得ることができた．



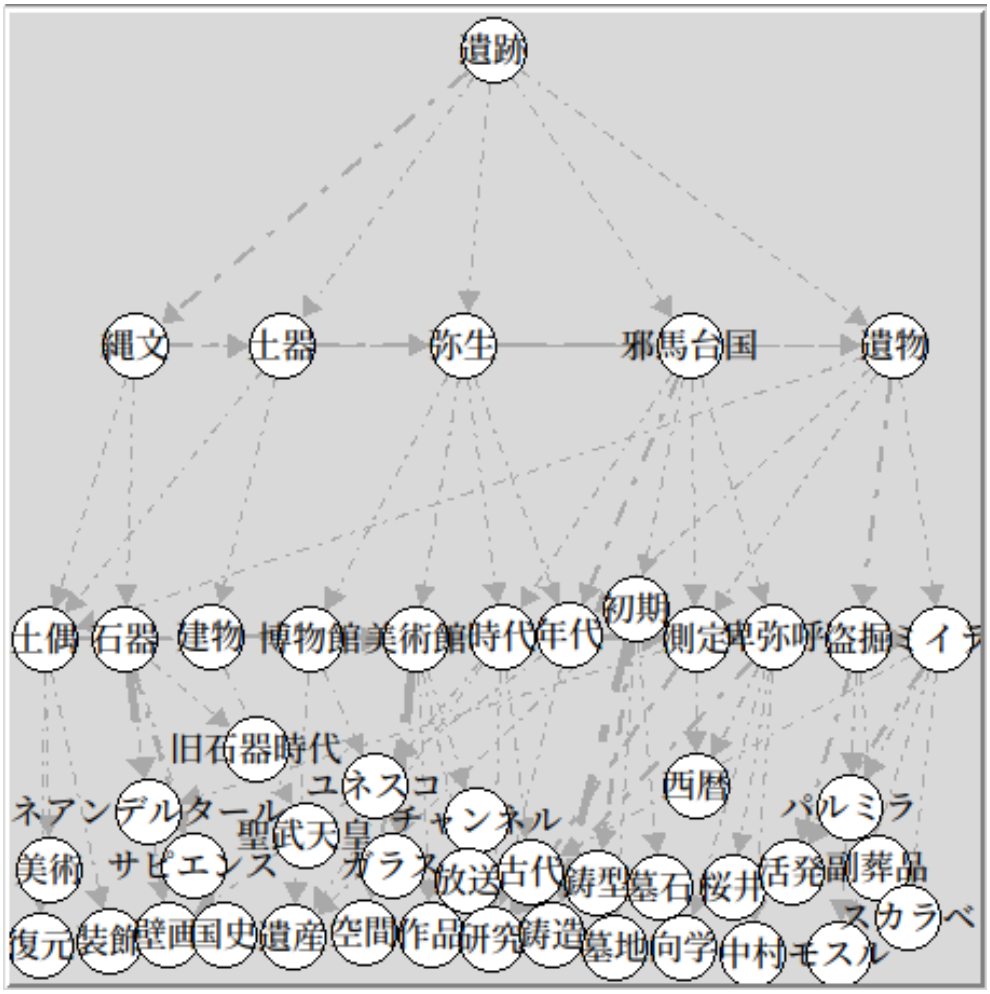
役に立つ単語:ユネスコ，パルミラ，邪馬台国，キリシタン，クルナ，ネアンデルタール
見やすくなっている部分:奄美と沖縄

図 4.1: テーマキーワードを「遺跡」として従来手法で構築したネットワーク



役に立つ単語:邪馬台国，パルミラ，ネアンデルタール，ユネスコ，モスル
見やすくなっている部分:年代と時代，博物館と美術館，ネアンデルタールとサピエンス，鋳造と鋳型，パルミラとモスル

図 4.2: テーマキーワードを「遺跡」として TF-IDF 合計値法で構築したネットワーク



役に立つ単語:ユネスコ，パルミラ，モスル，ネアンデルタール，邪馬台国
見やすくなっている部分:博物館と美術館，時代と年代，ネアンデルタールとサピエンス，鋳造と鋳型，パルミラとモスル

図 4.3: テーマキーワードを「遺跡」として TF-IDF 最大値法で構築したネットワーク

4.2　複数の被験者によるネットワークの評価

作成したネットワークが，本当に有用性，見やすさの観点から有効であるかを確認するため，4.1 節と同様の評価実験を複数の被験者に対して行う．

4.2.1　実験方法

被験者実験は 5 名に対して行った．各テーマキーワードごとに従来手法と 2 通りの提案手法で構築したネットワークを見せ，評価させる．

4.1 節の実験と同様に，ネットワークの構築において出現する単語を同種の単語が出やすくなるようにするために，Word2vec に学習させるデータとして，毎日新聞 12 年分 (2007〜2018) のデータ (1,166,761 記事) を使用した．単語のクラスタリングを行う際，クラスタ数は 5,000 とした．ネットワークを構築する際のデータとして 2018 年の毎日新聞の記事 (88,032 記事) を使用した．また，テーマキーワードは，4.1.1 節で示した 20 個のうち，以下の 5 個を抽出して行った．その 5 個のテーマキーワードを以下に示す．

- 産業構造，安全保障，5G，独立，遺跡

4.2.2　評価方法

4.1.2 節と同様の基準で評価を行う．

4.2.3　実験結果

従来手法と提案手法で構築したネットワークの被験者による評価について，有用性の観点から評価した結果を表 4.5 に示す．また，見やすさの観点から評価した結果を表 4.6 に示す．また，被験者実験に使用したネットワークの出現単語数を表 4.7 に示す．表 4.5 と表 4.6 のいずれも 5 人の被験者の平均を記載している．

また，各手法で有意差を調べるために両側検定の t 検定を行った．5 個のテーマキーワードにおいて，各被験者が各手法で役に立つと判断した単語の個数と，見やすくなっていると判断した部分の個数を比較した．5 人の被験者が 5 つのネットワークを評価した結果で検定を行っているため 25 対のデータで有意差検定を行った．ここで，有意水準は 5%である．結果を表 4.8 に示す．

表 4.5: 被験者実験における各テーマキーワードと役に立つ単語の個数

テーマキーワード	従来手法	TF-IDF 合計値法	TF-IDF 最大値法
産業構造	5.8	7.8	4.6
安全保障	9	10.8	6.2
5G	6	4.6	4
独立	7.8	7.6	7.2
遺跡	12.4	12.6	12.2
平均値	8.2	8.7	6.8

表 4.6: 被験者実験における各テーマキーワードと似た意味の単語が並んで、見やすくなっている部分の数

テーマキーワード	従来手法	TF-IDF 合計値法	TF-IDF 最大値法
産業構造	2.2	3.6	2.8
安全保障	1.2	3.2	2.8
5G	1.2	1.8	1
独立	5.4	5.4	5.6
遺跡	2.4	5.2	5
平均値	2.5	3.8	3.4

表 4.7: 被験者実験で使用したネットワークの単語数

テーマキーワード	従来手法	TF-IDF 合計値法	TF-IDF 最大値法
産業構造	56	48	40
安全保障	28	35	25
5G	32	40	20
独立	67	69	57
遺跡	39	46	48
平均値	44	48	38

表 4.8: 被験者実験における有意差検定

	従来手法と合計値法	従来手法と最大値法	合計値法と最大値法
役に立つ単語の個数	0.40	0.01	0.01
見やすい部分の個数	0.002	0.02	0.07

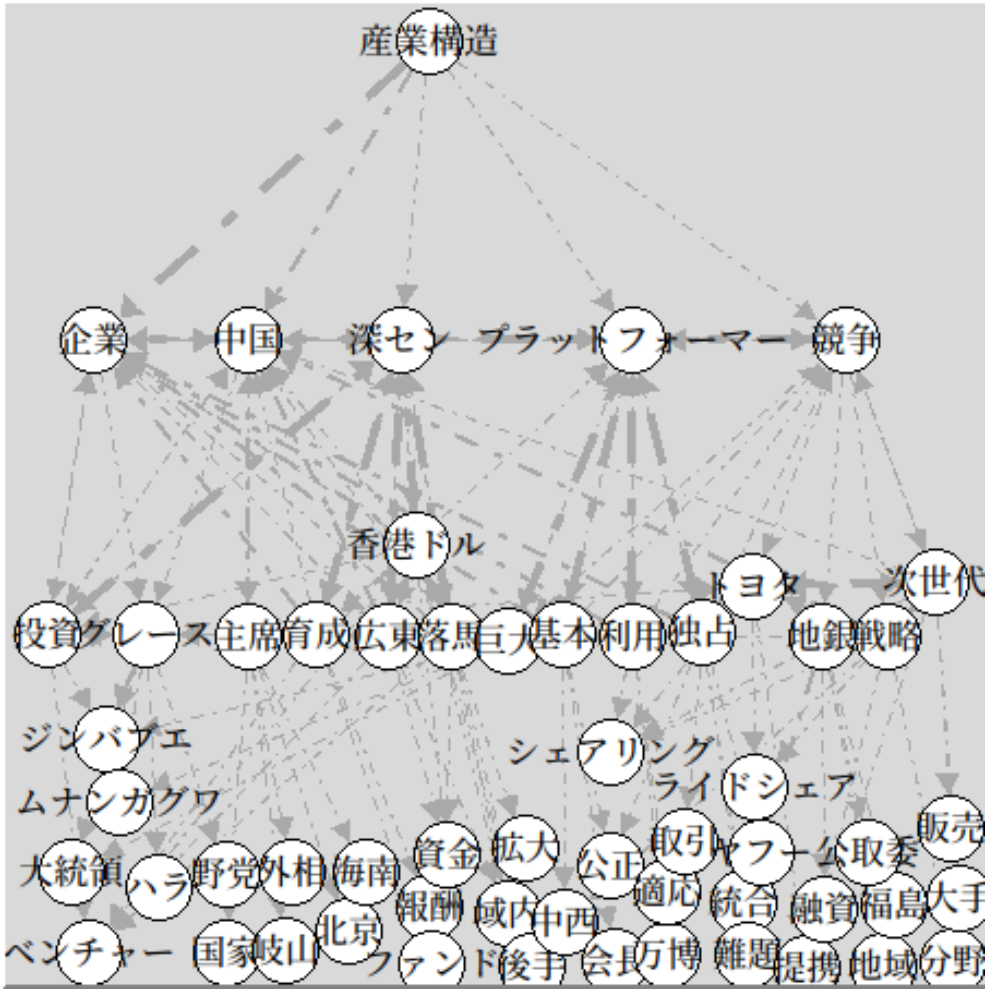
表 4.5 より，役に立つ単語の個数の平均は，従来手法の 8.2 個に対して，TF-IDF 合計値法が 8.7 個とほぼ同数であった．しかし，TF-IDF 最大値法は 6.8 個と下回った．また，表 4.8 より，役に立つ単語の個数では，従来手法と TF-IDF 最大値法の間，TF-IDF 合計値法と TF-IDF 最大値法の間で有意差があった．表 4.6 より，見やすい部分の個数の平均は，従来手法が 2.5 個であるのに対して，TF-IDF 合計値法が 3.8 個，TF-IDF 最大値法が 3.4 個といずれも上回った．さらに，表 4.8 より，見やすい部分の個数では，従来手法と TF-IDF 合計値法の間，従来手法と TF-IDF 最大値法の間で有意差があった．この結果より，4.1 節の実験同様，TF-IDF 合計値法が見やすさの観点からもっともよい方法であると考える．

4.1 節の実験同様，本実験の評価方法では，ネットワークの出現単語数が多いほど有利になると考えたので，各ネットワークに出現している単語数を数えた．表 4.7 より，被験者実験で使用したネットワークの出現単語の平均は，従来手法が 44 個，TF-IDF 合計値法が 48 個，TF-IDF 最大値法が 38 個と TF-IDF 合計値法が最も多く，TF-IDF 最大値法は他の 2 つの手法と比べて少し単語数が少なかった．

また，評価の一例としてテーマキーワードを「産業構造」として構築したネットワークを図 4.4 から図 4.6 に示す．従来手法によるネットワークを図 4.4 に，TF-IDF 合計値法を用いたネットワークを図 4.5 に，TF-IDF 最大値法を用いたネットワークを図 4.6 にそれぞれ示す．また，それぞれの図の下にそのネットワークにおいて被験者が役に立つと判断した単語と，似た意味の単語が並び見やすくなっていると判断した部分の一部を列挙している．

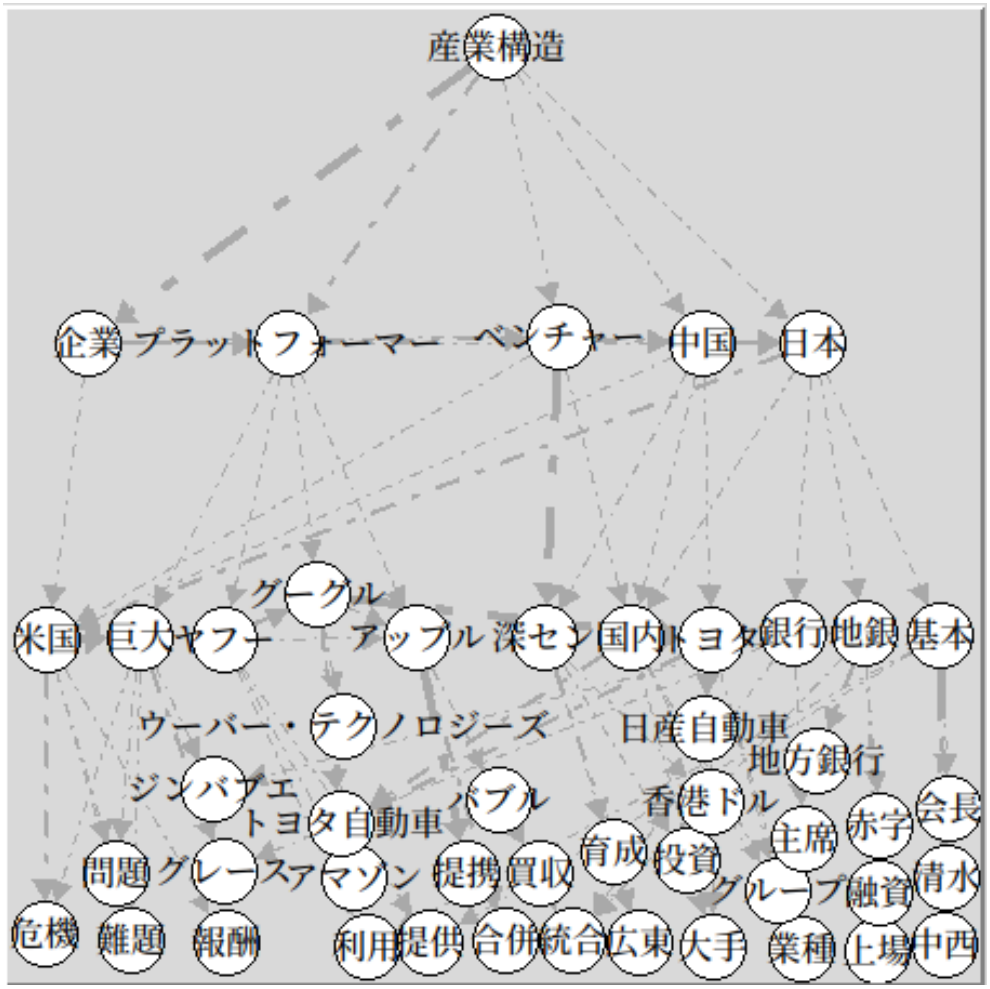
役に立つと判断された単語は，いずれの手法でも企業名が多く，産業構造の根幹をなしている企業がネットワークに出現していることが分かった．また，図 4.4 の従来手法では，独占といった社会的にも重大な問題がネットワークに出現していた．図 4.6 の TF-IDF 最大値法では，国保 (国民健康保険) やデフレという単語が役に立つ単語として挙げられており，日常生活に関する情報もネットワークから読み取ることができた．

一方，見やすくなったと判断された部分は，いずれの手法でも，国名が隣り合って出現している箇所を，見やすくなっていると判断している場合が多かった．また，提案手法の中でも図 4.5 の TF-IDF 合計値法は具体的な企業名がまとまっているところが見やすくなっていると判断された．



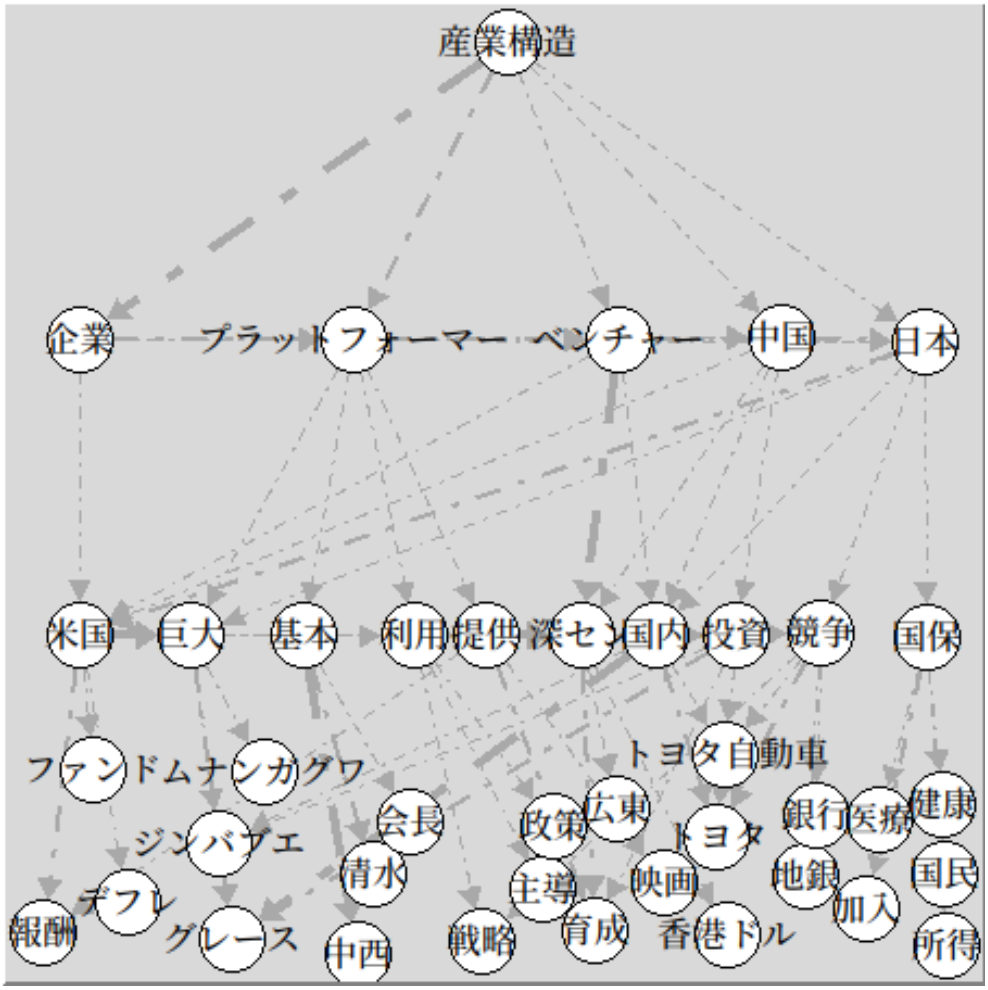
役に立つ単語:ヤフー，トヨタ，ベンチャー，プラットフォーマー，独占
見やすくなっている部分:中国と深セン，北京と海南

図 4.4: テーマキーワードを「産業構造」として従来手法で構築したネットワーク



役に立つ単語:ヤフー，グーグル，トヨタ，アマゾン，日産自動車，ウーバーテクノロジーズ
見やすくなっている部分:中国と日本，ヤフーとグーグルとアップル，銀行と地銀

図 4.5: テーマキーワードを「産業構造」として TF-IDF 合計値法で構築したネットワーク



役に立つ単語:デフレ，プラットフォーマー，トヨタ，ファンド，国保
見やすくなっている部分:中国と日本，プラットフォーマーとベンチャー

図 4.6: テーマキーワードを「産業構造」として TF-IDF 最大値法で構築したネットワーク

第5章　考察

5.1　著者によるネットワークの評価の考察

4.1 節より，著者によるネットワークの評価では，役に立つ単語の個数は，3 つの手法で大きく差がないことが分かった．また，見やすい部分の個数は，TF-IDF 合計値法が最も多いという結果となった．

5.1.1　役に立つ単語の個数に関する考察

提案手法の方が従来手法に比べて，役に立つ単語の個数が従来手法より少ないのは，TF-IDF 合計値法と TF-IDF 最大値法を用いた際に，従来手法で得られていた情報，つまり単語を表示させられないために，役に立つ単語が減ってしまった場合があると考えられる．一例として，4.1.3 節で示した「遺跡」をテーマキーワードとして構築したネットワークでは，図 4.1 の従来手法では出現していた「クルナ」という単語が，図 4.2 の TF-IDF 合計値法で構築したネットワークと図 4.3 の TF-IDF 最大値法で構築したネットワークでは出現していない．ここで例として挙げた「クルナ」はバングラデシュの遺跡が多い地域のことである．

一方，提案手法の方が従来手法に比べて，役に立つ単語の個数が従来手法より多いのは，従来手法で得ることができなかった単語が，同じ単語集合に所属する，より頻繁に新聞で取り上げられていた単語につられる形で，出現するようになったからだと考えられる．一例として，テーマキーワードを「宇宙」として従来手法で構築したネットワークを図 5.1 に示し，TF-IDF 合計値法で構築したネットワークを図 5.2 に示し，TF-IDF 最大値法で構築したネットワークを図 5.3 に示す．また，それぞれの図の下にそのネットワークにおいて役に立つと判断した単語を列挙している．役に立つと判断した単語の個数は図 5.1 の従来手法の 4 個に比べ，図 5.2 の TF-IDF 合計値法と図 5.3 の TF-IDF 最大値法の方がそれぞれ 9 個と多い．その一因として，図 5.1 では小惑星が「イトカワ」しか出ていないが，図 5.2 と図 5.3 では,「イトカワ」とともに「ベンヌ」

が出現している．このように，提案手法では，同じ単語集合に所属している単語につられる形で，従来手法では現れなかった単語が出現することがある．このことによって，従来手法より提案手法の方が，役に立つと判断した単語の個数が多くなったと考える．

また，TF-IDF 合計値法は，テーマキーワードによっては，人名や日付の情報を多く表示させ，役に立つと判断できる単語が少ないネットワークもあった．そのようなネットワークの一例を図 5.4 に示す．図 5.4 は「京都」をテーマキーワードとして，TF-IDF 合計値法でネットワークを構築したものである．図 5.4 のとおり，4 月や 3 月という月の情報のみが出ており，それらの月で何があったのかなどの情報を得られないネットワークとなった．この原因は,「3 月」「4 月」といった単語 1 つ 1 つの TF-IDF 値は小さいが，それらの合計を計算すると他の単語集合の TF-IDF 値の合計値よりも大きくなってしまうからだと考えられる．

5.1.2　見やすくなった部分に関する考察

見やすい部分の個数については，提案手法が従来手法を上回った．

しかし，従来手法でも見やすいと判断した部分は存在した．見やすいと判断した部分には，図 4.1 と図 4.4 で示したように，企業名や地名といった固有名詞が多く，これらの情報は新聞記事において共起しやすいと考えられる．

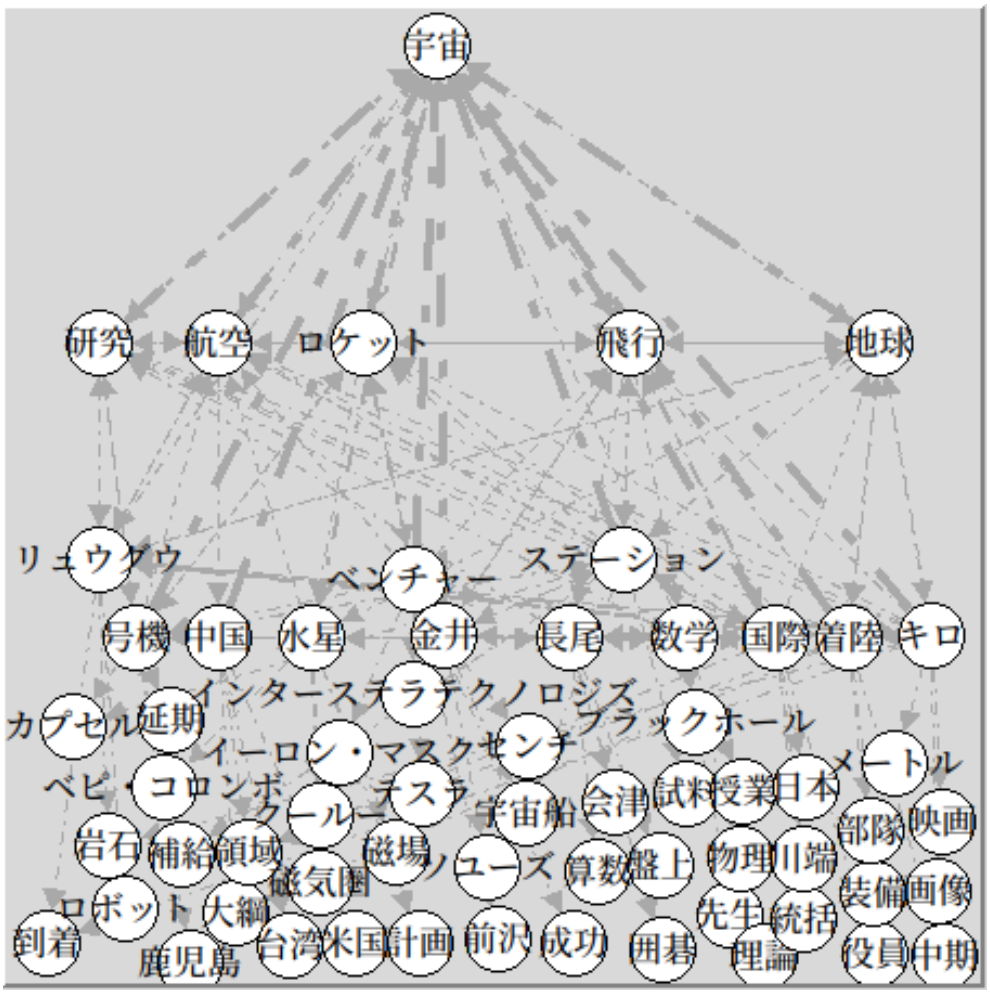
提案手法については，図 4.2 と図 4.3 で「遺跡」をテーマキーワードとしてネットワークを示したが，いずれも「縄文」と「弥生」という日本史上の歴史区分の間に,「土器」というものが出現していた．この原因は，Word2vec を利用して単語集合を作成した際,「縄文」「弥生」「土器」が同じ単語集合に所属しているからである．今回は Word2vec で 5000 個の単語集合を作成したが，より似た意味の単語を近くに配置し，見やすさを改善するために，単語集合作成の際に，単語集合の数の調整を行う必要があると考える．

5.2　複数の被験者によるネットワークの評価の考察

4.2 節より，複数の被験者によるネットワークの評価では，役に立つ単語の個数は，従来手法と TF-IDF 合計値法では差がなかったが，これら 2 つの手法と比べて TF-IDF 最大値法は少なかった．また，見やすい部分の個数は，4.1 節同様に TF-IDF 合計値法が最も多いという結果となった．

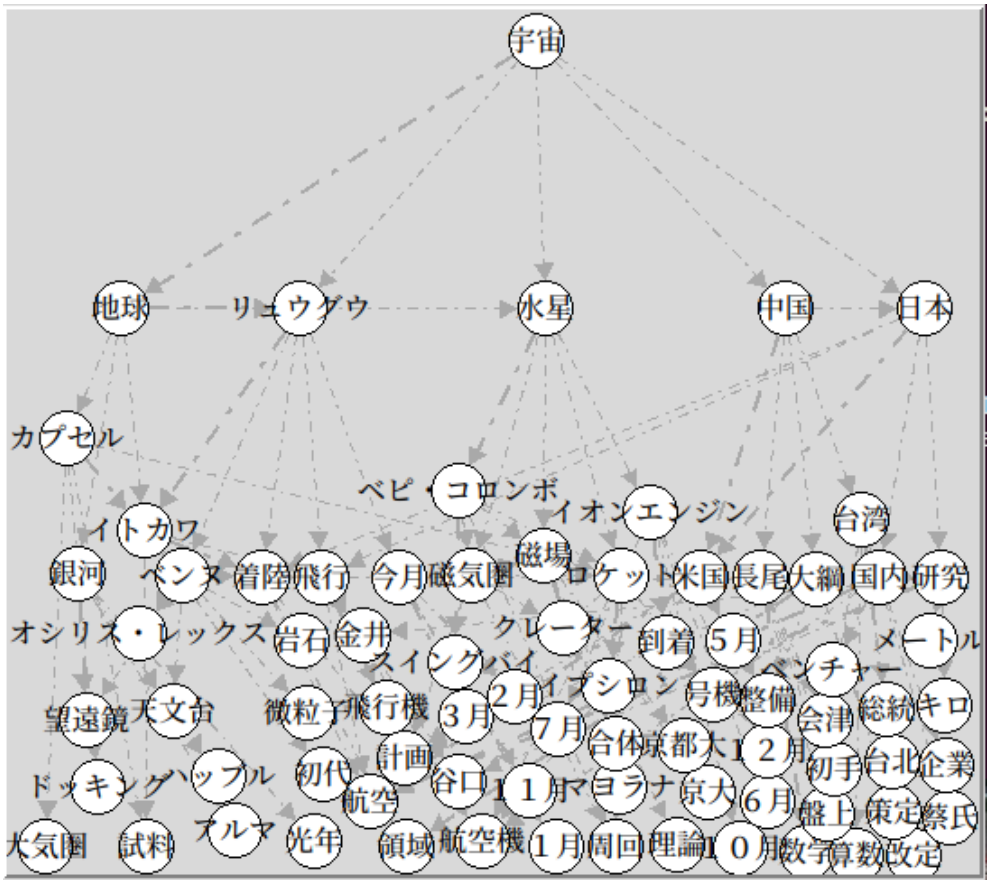
著者によるネットワークの評価に比べて，役に立つ単語の個数と見やすい部分の個数は増えた．特に役に立つ単語の個数は手法によっては 5 個ほど増えている．この原因は，被験者にとって普段なじみのない事柄をテーマキーワードとして設定したからと考える．また，被験者の持つ情報量によって役に立つと判断される単語は変化する．

見やすい部分の個数は，著者によるネットワークの評価と同様に TF-IDF 合計値法が最も多く，ネットワークが見やすくなっていることが確認できた．



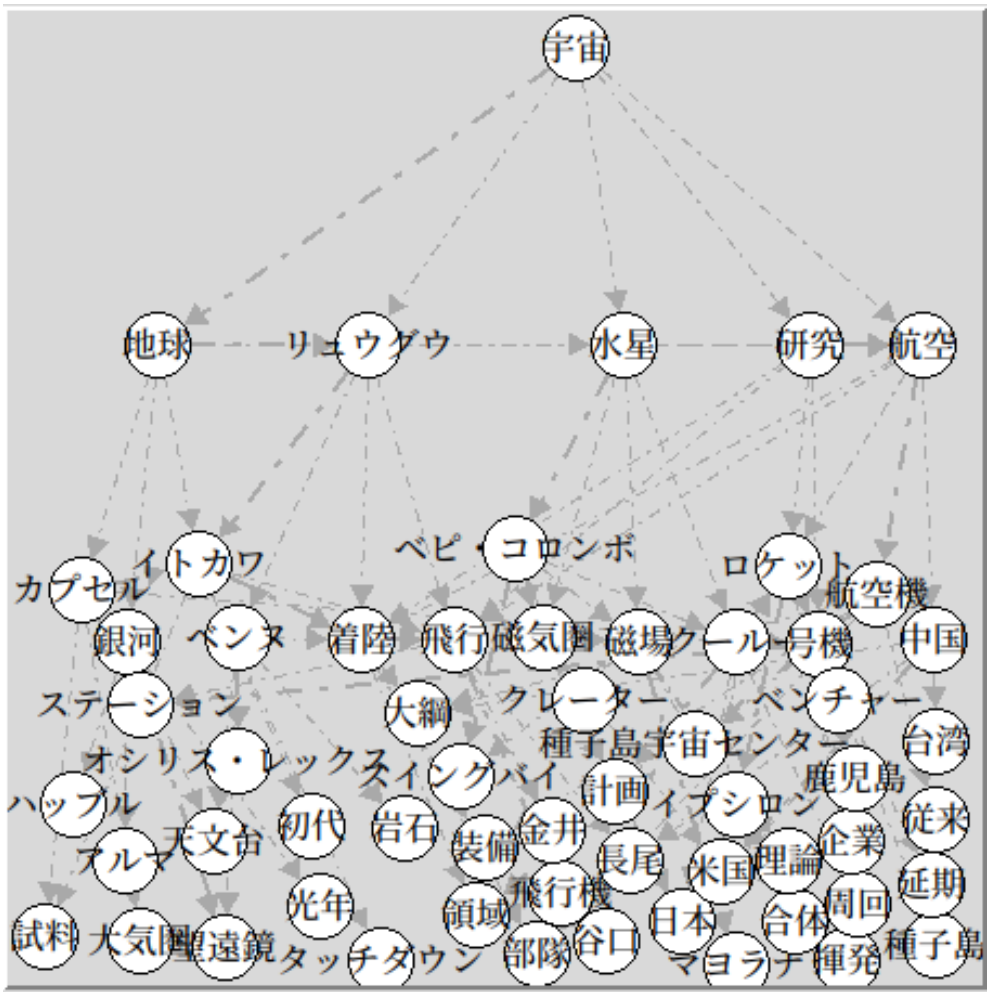
役に立つ単語:リュウグウ，ベピ・コロンボ，ベンチャー，ソユーズ

図 5.1: テーマキーワードを「宇宙」として従来手法で構築したネットワーク



役に立つ単語:リュウグウ，ベピ・コロンボ，ベンチャー，スイングバイ，ハッブル，ベンヌ，オシリス・レックス，クレーター，マヨラナ

図 5.2: テーマキーワードを「宇宙」として TF-IDF 合計値法で構築したネットワーク



役に立つ単語:リュウグウ，ベピ・コロンボ，ベンチャー，スイングバイ，ハッブル，マヨラナ，ベンヌ，オシリス・レックス，クレーター

図 5.3: テーマキーワードを「宇宙」として TF-IDF 最大値法で構築したネットワーク

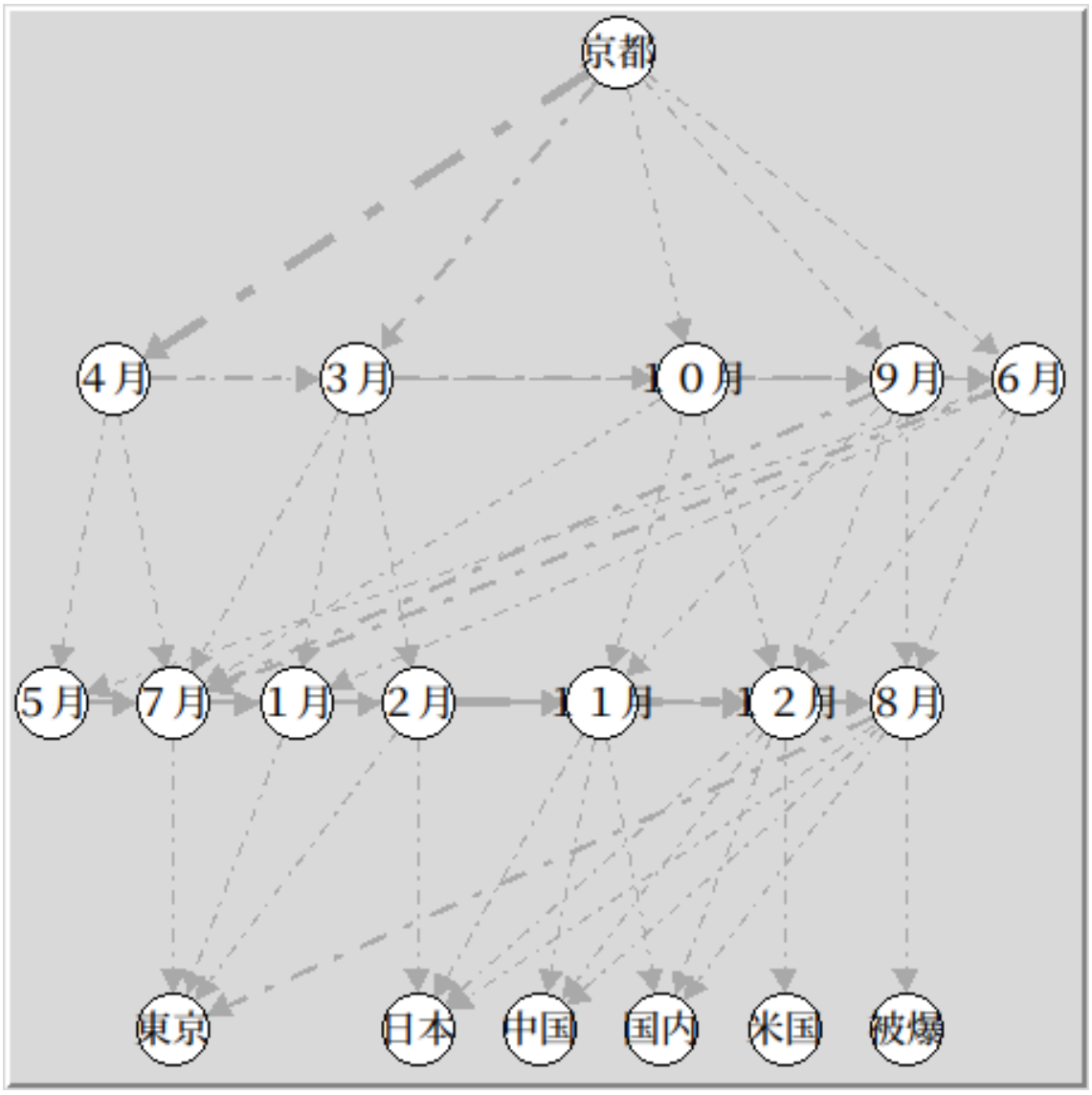


図 5.4: 日付の情報が多く表示されたネットワークの例

第6章　今後の課題

本研究では，概念ネットワーク構築の際に，Word2vec を用いて，出現する単語を同種の単語が出やすくなるようにし，似た意味を持つ単語を近くに配置することで見やすいネットワークを構築したが，いくつかの問題が残っている．本章では，残っている問題を今後の課題として以下にまとめる．

- ネットワークを構築した際に，単語のつながりしか表示されないため，単語間の関係性が分からない．単語に限らず 2 文節など短い文の形式でも表示できないかを検討する．
- 本研究の TF-IDF 合計値法は，有効な手法であるが，テーマキーワードによっては，人名や日付の情報を多く表示させるという問題があった．図 5.4 で示したように，出現単語がテーマキーワードとどのような関係性があるのかが分からないネットワークが構築される例もあった．この影響を低減する方策を検討したい．
- 1 年分の新聞記事を用いてネットワークを構築する際，これまでは本研究のように 1 年分の情報を一度にネットワークに表示させているが，例えば月ごとといったさらに細かい区分でネットワークを構築し，1 つのテーマキーワードから時期ごとの特徴を読み取ることができないかを検討したい．
- 5.1.2 節でも述べたように見やすさをさらに向上させるために，Word2vec で単語集合を作成する際に，最適な単語集合の数を検証する．

第7章　おわりに

本研究では，概念ネットワーク構築の際に Word2vec を利用して出現する単語を同種の単語が出やすくなるようにし，似た意味の単語を近くに配置し，ネットワークを見やすくすることを目的とした．著者による実験結果より，役に立つ単語の個数の平均は，従来手法が 3.2 個に対して，TF-IDF 合計値法が 3.3 個，TF-IDF 最大値法が 3.1 個と，従来手法と比べても情報量が減少することを抑え，見やすい部分の個数の平均は，従来手法が 1.2 個に対して，TF-IDF 合計値法が 3.1 個，TF-IDF 最大値法が 2.6 個と，似た意味の単語が並んで見やすくなっている部分は増えた．また，被験者実験では，役に立つ単語の個数の平均が，従来手法が 8.2 個に対して，TF-IDF 合計値法が 8.7 個，TF-IDF 最大値法が 6.8 個であり．見やすい部分の個数の平均は，従来手法が 2.5 個に対して，TF-IDF 合計値法が 3.8 個，TF-IDF 最大値法が 3.4 個であった．以上より，提案手法のなかでは，TF-IDF 合計値法を用いたネットワークのほうが，より見やすいネットワークを構築していることが分かった．

謝辞

最後に，この一年間研究を進めるにあたり，本研究のご指導をいただいた鳥取大学工学部電気情報系学科自然言語処理研究室の村田真樹教授，村上仁一准教授そして自然言語処理研究室の皆様に深く感謝するとともに，心から御礼申し上げます．また，ご多忙の中被験者実験にご協力いただいた皆様にも深く感謝申し上げます．そして，参考にさせていただいた論文の著者の方々に対して深く感謝申し上げます．

参考文献

[1] 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.

[2] 土遠雄大. テキスト処理に基づく概念ネットワークの構築における無関連ノードの扱い. 鳥取大学卒業研究発表会論文, 2013.

[3] 上東崇. 単語ネットワーク作成における検索エンジンの利用. 鳥取大学卒業研究発表会論文, 2016.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representation of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, 2013.

[5] 松尾豊, 友部博教, 橋田浩一, 石塚満. Web からの人間関係ネットワークの抽出と情報支援. 人工知能学会第 17 回全国大会講演論文, pp. 1–4, 2003.

[6] 松尾豊, 友部博教, 橋田浩一, 中島秀幸, 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.

[7] 堀田創, 萩原将文. 人間関係ネットワークに基づく情報推薦システムとその実装. 情報処理学会論文誌 データベース, Vol.2, No.1, pp. 46–56, 2009.

[8] 窪雄平. テキスト処理に基づく概念ネットワークの構築におけるリンクへの文字列付与. 鳥取大学卒業研究発表会論文, 2015.

[9] 南竣大. 単語ネットワーク構築システムを利用した単一文書の可視化. 鳥取大学卒業研究発表会論文, 2016.

[10] 畑山満美子, 松尾義博, 白井諭. 重要語句抽出による新聞記事自動要約. 自然言語処理 = Journal of Natural Language Processing, Vol.9, No.4, pp. 55–73, 2002.

[11] 岡田貴史, 石橋融, 高間康史. M2VSM を用いたテキストマイニングシステムの構築に関する考察. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 22, pp. 52–52, 2006.

[12] Yasufumi Takama, Takashi Okada, and Toru Ishibashi. Online Text Mining System based on M2VSM. SCIS & ISIS, Vol. 2008, pp. 739–743, 2008.

[13] 松本一樹, 松井藤五郎. 深層学習を用いた新聞記事分析による市場動向予測. 第 79 回全国大会講演論文集, pp. 353–354, 2017.