

概要

近年、電子テキストの増加により、膨大なテキストデータから情報を得る機会が増えている。その際に、膨大なテキストデータから有用な情報を取り出す技術が必要とされる。

大竹ら [1] は、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び、情報抽出の技術を用いて社会構造モデルを自動で構築した。窪 [2] は、新聞文書群において単語間の文字列を用い、単語間の関係を示す文字列を抽出した。そして、単語ネットワークの作成後、ネットワークのリンクにノード同士の関係性を示す文字列の付与を行った。土遠 [3] は、テキストデータを用いたネットワークの作成の際の、ノード(単語)の選定処理を改良した。作成されたネットワークに無関係なノードを含むことを問題点として、無関係なノードの削除を行った。しかし、先行研究では新聞の情報不足により、情報量が不十分なネットワークが作成される場合があった。

そこで本研究では、検索エンジンを用いて単語ネットワークの作成を行う。ネットワークの主となる概念をテーマキーワードとして設定し、テーマキーワードに関する情報を検索エンジンで収集する。そして、収集したテーマキーワードに関する情報をもとにネットワークを作成する。これにより、幅広いテーマキーワードで情報量が十分なネットワークを作成できる。

実際に 20 個のテーマキーワードでネットワークを作成し、新規で内容把握や発想支援に役立つとされるノードに着目して、評価を行った。評価の結果、作成したネットワークのうち、約 8 割のネットワークが情報取得に効果的であることが確認できた。また、新聞を用いた場合に十分なネットワークが作成できなかった 8 個のテーマキーワードのうち、6 個のテーマキーワードは検索エンジンを用いた提案手法で情報取得に十分とされるネットワークが作成できた。さらに、検索エンジンを用いた提案手法では、新聞を用いた場合には見られなかったノードが多く見られた。

しかし、作成したネットワークのノードが多すぎることで、重要な情報が把握しづらい場合も見られた。今後は、得られた情報を使用者に理解しやすい状態で提示することで、さらに役立つことができると考えている。

目次

| | | |
|-------|-----------------------|----|
| 第1章 | はじめに | 1 |
| 第2章 | 関連研究 | 3 |
| 2.1 | 単語ネットワーク作成の関連研究 | 3 |
| 2.2 | Web上のデータを用いた関連研究 | 4 |
| 2.3 | キーワードの抽出の関連研究 | 4 |
| 第3章 | 先行手法 | 6 |
| 3.1 | ネットワーク作成の概要 | 6 |
| 3.2 | キーワードを含む記事の抽出 | 7 |
| 3.3 | 記事の形態素解析 | 8 |
| 3.4 | ノード候補の抽出 | 9 |
| 3.5 | ノード候補の選定 | 9 |
| 3.6 | ネットワークの拡大 | 9 |
| 3.7 | 新聞を用いたネットワークの作成 | 11 |
| 第4章 | 提案手法 | 12 |
| 4.1 | 検索エンジンを用いたネットワークの作成 | 12 |
| 4.2 | 不要単語リストの作成 | 13 |
| 4.3 | アンドテーマキーワードによる情報の追加取得 | 13 |
| 4.4 | DFの置き換えによるネットワークの変化 | 14 |
| 第5章 | 実験 | 15 |
| 5.1 | 主観によるネットワークの評価 | 15 |
| 5.1.1 | 実験条件 | 15 |
| 5.1.2 | 評価方法 | 15 |
| 5.1.3 | スニペットを用いた評価方法 | 16 |

| | | |
|------------|----------------------|-----------|
| 5.1.4 | 実験結果 | 17 |
| 5.2 | 複数被験者によるネットワークの評価 | 25 |
| 5.2.1 | 実験条件 | 25 |
| 5.2.2 | 評価方法 | 25 |
| 5.2.3 | 実験結果 | 25 |
| 5.3 | アンドテーマ手法の評価 | 26 |
| 5.3.1 | 実験条件 | 26 |
| 5.3.2 | 評価方法 | 26 |
| 5.3.3 | 実験結果 | 26 |
| 5.4 | DFの置き換えによる変化の評価 | 40 |
| 5.4.1 | 実験条件 | 40 |
| 5.4.2 | 評価方法 | 40 |
| 5.4.3 | 実験結果 | 40 |
| 第6章 | 考察 | 45 |
| 6.1 | 主観によるネットワークの評価の考察 | 45 |
| 6.2 | 複数被験者によるネットワークの評価の考察 | 46 |
| 6.3 | アンドテーマ手法の評価の考察 | 46 |
| 6.4 | DFの置き換えによる変化の考察 | 47 |
| 6.5 | 今後の課題 | 47 |
| 第7章 | おわりに | 48 |

表 目 次

| | | |
|------|--|----|
| 5.1 | テーマキーワード「宇宙」のスニペット例 | 16 |
| 5.2 | 基本手法における一定個数以上の新規の情報取得に役立つノード（単語） を持つネットワーク及びスニペットの比率 | 18 |
| 5.3 | テーマキーワード「鳥取」のスニペット | 21 |
| 5.4 | テーマキーワード「石油」のスニペット | 24 |
| 5.5 | 被験者実験における一定個数以上の新規の情報取得に役立つノード（単語） を持つネットワーク及びスニペットの比率 | 25 |
| 5.6 | アンドテーマ手法における一定個数以上の新規の情報取得に役立つノード （単語）を持つネットワーク及びスニペットの比率 | 27 |
| 5.7 | テーマキーワード「京都」のスニペット | 29 |
| 5.8 | テーマキーワード「京都」, アンドテーマキーワード「寺」のスニペット | 31 |
| 5.9 | テーマキーワード「世界遺産」のスニペット | 33 |
| 5.10 | テーマキーワード「世界遺産」, アンドテーマキーワード「遺跡」のスニ ペット | 35 |
| 5.11 | テーマキーワード「宇宙」のスニペット | 37 |
| 5.12 | テーマキーワード「宇宙」, アンドテーマキーワード「銀河」のスニペット | 39 |
| 5.13 | DFの置き換えにおける一定個数以上の新規の情報取得に役立つノード（単語） を持つネットワーク及びスニペットの比率 | 41 |
| 5.14 | テーマキーワード「ギリシャ」のスニペット | 44 |

目 次

| | | |
|------|---|----|
| 3.1 | ネットワーク作成の流れ | 6 |
| 3.2 | 記事の抽出 | 7 |
| 3.3 | 形態素解析の出力例 | 8 |
| 3.4 | ネットワーク拡大の例 | 10 |
| 5.1 | 先行手法によるテーマキーワード「鳥取」のネットワーク | 19 |
| 5.2 | 基本手法によるテーマキーワード「鳥取」のネットワーク | 20 |
| 5.3 | 先行手法によるテーマキーワード「石油」のネットワーク | 22 |
| 5.4 | 基本手法によるテーマキーワード「石油」のネットワーク | 23 |
| 5.5 | 基本手法によるテーマキーワード「京都」のネットワーク | 28 |
| 5.6 | アンドテーマ手法によるテーマキーワード「京都」, アンドテーマキーワード「寺」のネットワーク | 30 |
| 5.7 | 基本手法によるテーマキーワード「世界遺産」のネットワーク | 32 |
| 5.8 | アンドテーマ手法によるテーマキーワード「世界遺産」, アンドテーマキーワード「遺跡」のネットワーク | 34 |
| 5.9 | 基本手法によるテーマキーワード「宇宙」のネットワーク | 36 |
| 5.10 | アンドテーマ手法によるテーマキーワード「宇宙」, アンドテーマキーワード「銀河」のネットワーク | 38 |
| 5.11 | IDF=1としたテーマキーワード「ギリシャ」のネットワーク | 42 |
| 5.12 | DF=Wikiとしたテーマキーワード「ギリシャ」のネットワーク | 43 |

第1章 はじめに

近年，電子テキストの増加により，膨大なテキストデータから情報を得る機会が増えている．その際に，膨大なテキストデータから有用な情報を取り出す技術が必要とされる．

大竹ら [1] は，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び，情報抽出の技術を用いて社会構造モデルを自動で構築した．窪 [2] は，新聞文書群において単語間の文字列を用い，単語間の関係を示す文字列を抽出した．そして，単語ネットワークの作成後，ネットワークのリンクにノード同士の関係性を示す文字列の付与を行った．土遠 [3] は，テキストデータを用いたネットワークの作成の際の，ノード(単語)の選定処理を改良した．作成されたネットワークに無関係なノードを含むことを問題点として，無関係なノードの削除を行った．しかし，先行研究では新聞の情報不足により，情報量が不十分なネットワークが作成される場合があった．

そこで本研究では，検索エンジンを用いて単語ネットワークの作成を行う．ネットワークの主となる概念をテーマキーワードとして設定し，テーマキーワードに関する情報を検索エンジンで収集する．そして，収集したテーマキーワードに関する情報をもとにネットワークを作成する．これにより，幅広いテーマキーワードで情報量が十分なネットワークを作成できる．

本研究の主張点を以下に示す．

- 新聞を用いて単語ネットワークを作成した場合，新聞の情報不足によって情報量が不十分なネットワークが作成される場合がある．そのため，検索エンジンを用いてネットワークを作成する．
- 新規で内容把握や発想支援に役立つとされるノードに着目して，評価を行う．また，Web の要約文であるスニペットも評価に用いる．
- 検索エンジンを用いた提案手法により 20 個のテーマキーワードでネットワークを作成した結果，作成したネットワークのうち，約 8 割のネットワークが情報取得に効果的であることが確認できた．また，新聞を用いた場合に十分なネットワークが作成できなかった 8 個のテーマキーワードのうち，6 個のテーマキーワードで検索

エンジンを用いた提案手法で情報取得に十分とされるネットワークが作成できた。さらに、検索エンジンを用いた提案手法では、新聞を用いた場合には見られなかったノードが多く見られた。

- DFの計算に使用するデータを変更し、ネットワークに出現するノードの調査を行った。また、どのデータをDFの計算に使用することがネットワークの作成に適しているかを確認する評価も行った。調査の結果、2016年10月までのWikipediaのデータがネットワークの作成に適していることが確認できた。
- ネットワークの作成後、使用者が興味をもったノードや興味のある単語をアンドテーマキーワードとする。そして、テーマキーワードとアンドテーマキーワードでアンド検索を行い、ネットワークを再度作成した。これにより、使用者にとって興味深いネットワークやアンドテーマキーワードに沿った内容のネットワークを作成できる。作成したネットワークを評価したところ、6割のネットワークが情報取得に効果的であることが確認できた。また、アンドテーマキーワードに関わるノードが多く出現し、テーマキーワードのみを用いたネットワークでは得られない情報が得られた。

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、先行手法を述べ、第4章では、提案手法を述べる。第5章では、実験条件と実験結果や評価方法と評価結果を述べる。第6章では、実験の結果の考察と、研究の今後の課題を述べる。第7章では、本研究のまとめを述べる。

第2章 関連研究

2.1 単語ネットワーク作成の関連研究

大竹ら [1] は、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び、情報抽出の技術を用いて社会構造モデルを自動で構築した。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った。

窪 [2] は、新聞文書群において単語間の文字列を用い、単語間の関係を示す文字列を抽出した。そして、単語ネットワークの構築後、ネットワークのリンクにノード同士の関係性を示す文字列の付与を行った。これにより構築した単語ネットワークから得られる情報がより詳細なものとなり、単語ネットワークの利便性を向上することができる。以下に文字列の抽出例を示す。

「宇宙」、「飛行」の2単語

- 宇宙飛行士の星出彰彦さんらが搭乗したロシアの宇宙船が

「改革」、「一体」の2単語

- 政治・行政改革とも一体で
- 定数削減と公務員給与削減は「一体改革と一体で理解する人が多い」と述べ
- 参院の改革も一体で実施すべきだった

土遠 [3] は、テキストデータを用いたネットワークの構築の際の、ノード(単語)の選定処理を改良した。構築されたネットワークに無関係なノードを含むことを問題点として、無関係なノードの削除を行った。無関係なノードを削除するために、最初に設定したキーワードを含む記事を抽出するテーマ限定法と、キーワードとの関連性が低い単語を消去するテーマ関連抽出法の2つの手法を用いた。これにより、関連のないノードを含みにくいネットワークを構築した。

2.2 Web上のデータを用いた関連研究

松尾 [4] [5] らは、Web マイニングの技術を用いて、Web 上の情報だけから特定のコミュニティ（人工知能学会）の人間関係を自動的に抽出した。検索エンジンを利用して、特定の2者に関係する文書を検索し、そのヒット件数や文書の内容から、関係の強さおよび関係の種類を判断した。その後、人をノード、関係をエッジ、関係の種類をエッジのラベルとしたネットワークを作成し、人間関係を可視化した。

友部 [6] らは、Web から抽出した情報を用い、人の関係性をセマンティック Web の枠組みの中で記述できること、得られたネットワークの中心性の解析をすることで、コミュニティにおける著名な人物が分かること、それが信頼性につながる別の尺度と相関していることを示した。セマンティック Web とは、Web ページおよびその中に記述された内容について、それが何を意味するかを表す情報を一定の規則に従って付加することで、コンピュータが効率よく情報を収集・解釈できるようにする構想である。セマンティック Web では、計算機に理解できる RDF などの形式的な記述の方法を用い、他との関係性を明示的に記述していくことで推論が可能になり、高度な情報検索や問題解決を期待することができる。

Harada [7] らは、トピックのために検索された Web ページから適切な名前を抽出することを提案した。そして、効果的な人の発見方法を探るために約 5,000 万の日本の Web ページを分析して、日本の個人名の分布を調べた。また、Web ページの数ではなく、名前を含む Web サーバーの数を利用し、幅広いテーマをカバーする実験により、その有効性を示した。

2.3 キーワードの抽出の関連研究

松村 [8] らは、著者の想定する読者が文書を読むときの記憶の活性状態に着目して、文書を読んだ後に読者に強い印象を残す語をキーワードとして取り出すことを提案した。人が文書を読んでその内容を理解できるのは、文書中の語が何らかの形で脳に記憶され、脳がその記憶を解釈するからであり、記憶には、ある語が想起（活性化）されるとその語に関連する語も想起（活性化）されるというプライミング効果がある。そこで、プライミング効果によって読者の記憶に強く残る語、すなわち強く活性化される語を著者の主張を表すキーワードとして取り出した。

大沢 [9] らは、文書の主張の内容を表すキーワードの抽出を行った。そのようなキー

ワードはユーザの検索語に近い主張をもつ文書を検索するのに役立ち、ユーザは、自分と同じ考えの著者や、内容は見聞しているが出典が不明な文書を探ることができる。

第3章 先行手法

本研究では，土遠 [3] による単語ネットワークを作成する手法を，先行手法とした．本章では，先行手法について説明する．

3.1 ネットワーク作成の概要

作成したいネットワークの主となる概念を，テーマキーワードとして設定する．例として「宇宙」「ギリシャ」等の単語が挙げられる．そして，設定したテーマキーワードでネットワークの作成を行い，関係のある事物を可視化する．ネットワーク作成の流れを図 3.1 に示す．

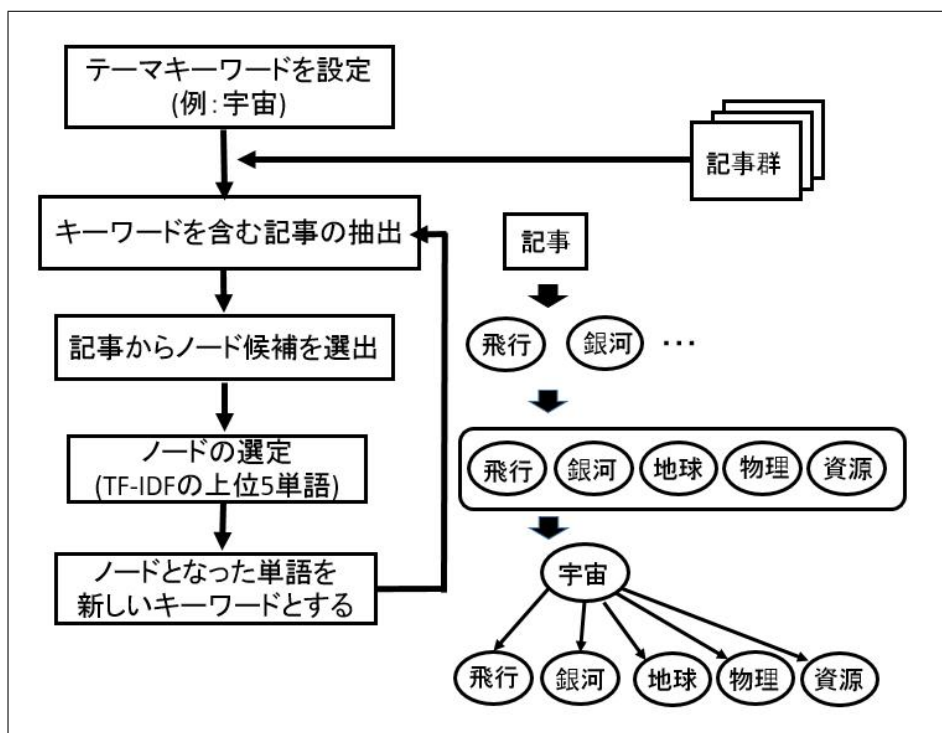


図 3.1: ネットワーク作成の流れ

3.2 キーワードを含む記事の抽出

本研究では，土遠 [3] が用いたテーマ限定抽出法を用いる．ネットワークを作成する際に，単語ネットワークを作成する元となる記事群から，キーワードを含む記事のみを抽出する．そして，抽出した記事で新たに記事群を作成する．これにより，キーワードに関係する内容を含む記事群を利用することができ，ネットワークにキーワードと無関係なノードが含まれにくくなる．抽出方法を図 3.2 に示す．

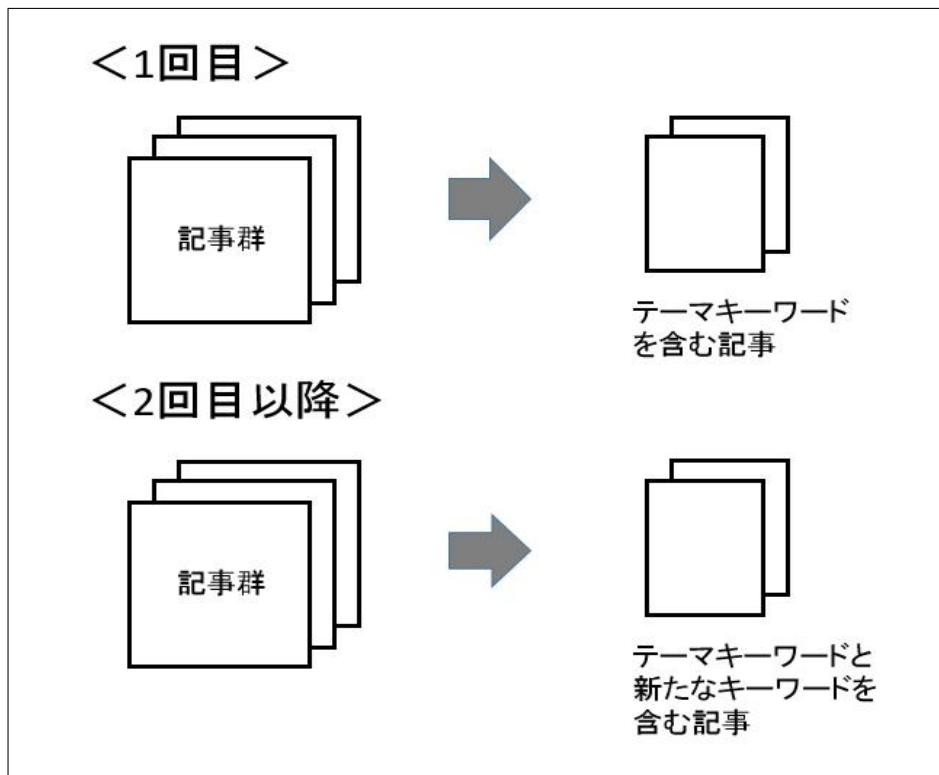


図 3.2: 記事の抽出

1 回目は，テーマキーワードを含む記事を抽出する．2 回目以降は，テーマキーワードと，次のノードとなった新たなキーワードの，2つのキーワードを含む記事を抽出する．

3.3 記事の形態素解析

3.2 節で作成された記事群に対して形態素解析を用い、名詞を取り出す。

形態素解析とは、テキストを形態素と呼ばれる言語で意味を持つ最小単位に分割することである。形態素は、厳密には単語とは違った分割の単位だが、おおよそ単語と同じようなものになり、品詞の情報を持つものである。形態素解析結果の例を図 3.3 に示す。

入力:「宇宙飛行士の若田光一さんが国際宇宙ステーションの第 39 代船長に就任した」

| | |
|---------------------------------------|-----------------------------------|
| --- | |
| 宇宙飛行士の若田光一さんが国際宇宙ステーションの第 39 代船長に就任した | |
| 宇宙 | 名詞,一般,*,*,*,*宇宙,ウチュウ,ウチュー |
| 飛行 | 名詞,サ変接続,*,*,*,*飛行,ヒコウ,ヒコー |
| 士 | 名詞,接尾,一般,*,*,*,*士,シ,シ |
| の | 助詞,連体化,*,*,*,*の,ノ,ノ |
| 若田 | 名詞,固有名詞,人名,姓,*,*,*若田,ワカタ,ワカタ |
| 光一 | 名詞,固有名詞,人名,名,*,*,*光一,コウイチ,コーイチ |
| さん | 名詞,接尾,人名,*,*,*,*さん,サン,サン |
| が | 助詞,格助詞,一般,*,*,*,*が,ガ,ガ |
| 国際 | 名詞,一般,*,*,*,*国際,コクサイ,コクサイ |
| 宇宙 | 名詞,一般,*,*,*,*宇宙,ウチュウ,ウチュー |
| ステーション | 名詞,一般,*,*,*,*ステーション,ステーション,ステーション |
| の | 助詞,連体化,*,*,*,*の,ノ,ノ |
| 第 | 接頭詞,数接続,*,*,*,*第,ダイ,ダイ |
| 3 | 名詞,数,*,*,*,*3,サン,サン |
| 9 | 名詞,数,*,*,*,*9,キュウ,キュー |
| 代 | 名詞,接尾,助数詞,*,*,*,*代,ダイ,ダイ |
| 船長 | 名詞,一般,*,*,*,*船長,センチョウ,センチャー |
| に | 助詞,格助詞,一般,*,*,*,*に,ニ,ニ |
| 就任 | 名詞,サ変接続,*,*,*,*就任,シュウニン,シューニン |
| し | 動詞,自立,*,*,*,*サ変・スル,連用形,する,シ,シ |
| た | 助動詞,*,*,*,*特殊・タ,基本形,た,タ,タ |
| EOS | |

図 3.3: 形態素解析の出力例

図 3.3 のように形態素解析を行うことで、品詞の情報を持った単語に分割する。本研究では、形態素解析に Mecab を用いる。また、形態素解析を用いて名詞を取り出す際に、一文字、数字のみの単語を除外する。また、4.2 節に後述する不要単語リストに含まれる単語も除外する。

3.4 ノード候補の抽出

形態素解析を行った後，ノード候補となる単語の抽出を行う．

3.3 節の図 3.3 を例とすると，「宇宙」「飛行」「若田」「光一」「国際」「ステーション」「船長」「就任」といった単語がノード候補として抽出される．

3.5 ノード候補の選定

TF-IDF とは，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* を用いて，抽出されたノード候補の中から，実際にノードに用いる単語を選定する．本研究では，*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 は以下の式 3.1 で算出される．

$$TF-IDF = TF * \log\left(\frac{N}{DF}\right) \quad (3.1)$$

TF はキーワードを含む記事における単語の出現回数，*DF* は全記事のうちの単語を含む記事数，*N* は全記事数を示す．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単語には高い重みを与えている．

3.6 ネットワークの拡大

3.5 節で得た *TF-IDF* 値の上位 5 単語を，キーワードから繋がる次のノードとする．次のノードを新たなキーワードとして設定し，3.2 節の 2 回目以降の処理に戻る．そして，3.2 節から本節までの処理を繰り返すことにより，単語ネットワークを拡大していく．ネットワーク拡大の例を図 3.4 に示す．

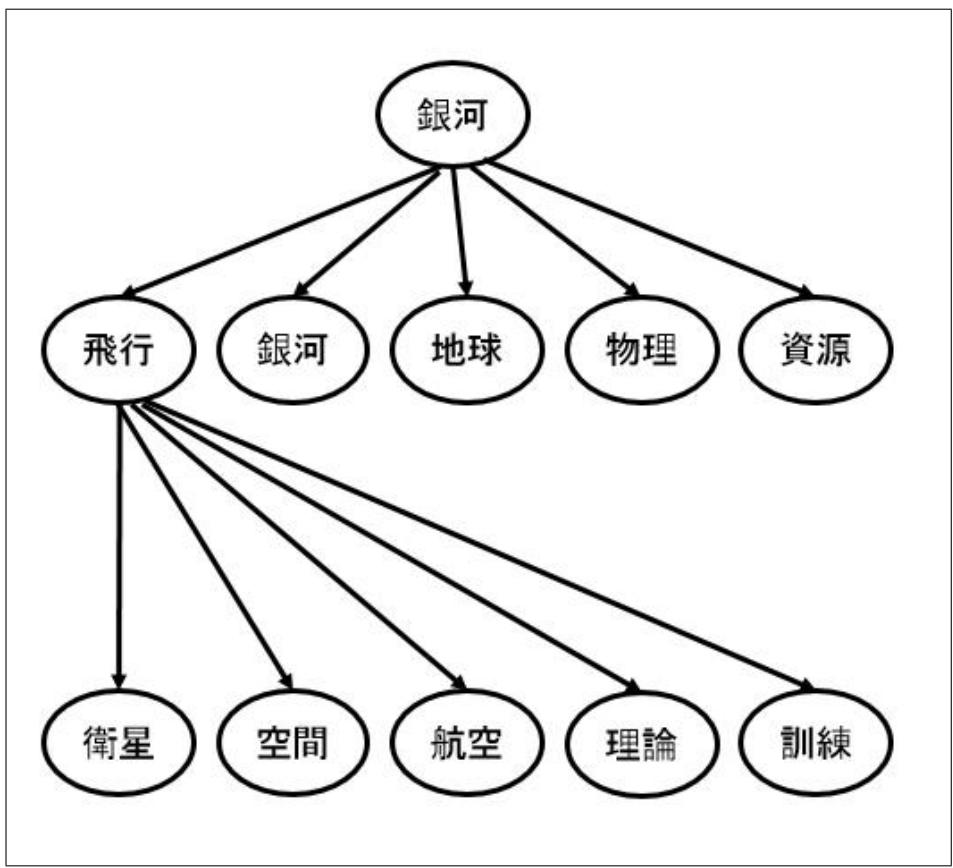


図 3.4: ネットワーク拡大の例

3.7 新聞を用いたネットワークの作成

先行手法では，新聞記事群のデータ（以降は新聞データとする）を元に，単語ネットワークを作成する．入力及びDFの計算にはそれぞれ，毎日新聞（2012年1年分）を用いる．先行手法の手順を以下に示す．

手順1 作成するネットワークの主となる概念を，テーマキーワードとして設定する．

手順2 新聞データのテーマキーワードが含まれている記事のみを利用するために，新聞データからテーマキーワードを含む記事を抽出し，記事群Aを作成する．2回目以降は，テーマキーワードと新たなキーワードをそれぞれ含む記事を抽出する【3.2節の処理に相当】

手順3 抽出した記事群Aに形態素解析を用い，ノード候補になる単語を抽出する．その際に1文字，数字のみの単語を除外する．また，4.2節の不要単語リストに含まれる単語も除外する【3.3，3.4節の処理に相当】

手順4 得られたノード候補の中から， $TF-IDF$ の値を用いて，その値が最も大きい5単語を次のノードとする【3.5節の処理に相当】

手順5 次のノードとなった単語を新たなキーワードとして手順2に戻り，同様の処理を繰り返すことでネットワークを拡大していく【3.6節の処理に相当】

第4章 提案手法

本章では，本研究の提案手法について説明する．

4.1 検索エンジンを用いたネットワークの作成

テーマキーワードに関する情報が使用する新聞データに少ない場合，情報量が不十分なネットワークが作成されてしまう．そのため，幅広いテーマキーワードでネットワークを作成するには不適切である．そこで，幅広いテーマキーワードでネットワークの作成を行うため，検索エンジンを用いてネットワークの作成を行う．本研究では，4.3節と4.4節のベースとなるこの節で述べる手法を，基本手法とする．検索エンジンには Bing Search API [10] を用い，検索結果の上位 50 件を利用する．また，DF の計算には Wikipedia (2016 年 10 月まで) を用いる．基本手法によるネットワークの作成手順を以下に示す．

- 手順 1 構築したいネットワークの主となる概念を，テーマキーワードとして設定する．
- 手順 2 検索エンジンを用い，テーマキーワードの検索結果を記事群 A として作成する．
- 手順 3 記事群 A のテーマキーワードが含まれている記事のみを利用するために，記事群 A からテーマキーワードを含む記事を抽出し，記事群 B を作成する．2 回目以降は，テーマキーワードと新たなキーワードをそれぞれ含む記事を抽出する．
- 手順 4 記事群 B に形態素解析を用い，ノード候補となる単語を抽出する．その際に 1 文字，数字のみの単語を除外する．また，4.2 節の不要単語リストに含まれる単語も除外する．
- 手順 5 得られたノード候補の中から， $TF-IDF$ の値を用いて，その値が最も大きい 5 単語を次のノードとする．
- 手順 6 テーマキーワードと次のノードとなった単語で検索エンジンを用いてアンド検索を行い，新しい検索結果で記事群 A を再度作成する．その後，次のノードの単語

を新たなキーワードとし、手順3より同様の処理を繰り返すことでネットワークを拡大していく。例として、テーマキーワードが「宇宙」、新たなキーワードが「飛行」の場合、「宇宙 飛行」で検索を行い、その検索結果を記事群 A として使用する。

4.2 不要単語リストの作成

本研究の提案手法では、検索エンジンによって収集した検索結果を用いてネットワークを作成する。そのため、ネットワーク上に頻出する「編集」「メニュー」といった単語が、ノードとしてネットワークに出現してしまう場合がある。他にも「よう」「さん」といった頻出する平仮名も当てはまる。これらの単語はノードとして出現しても得られる情報がほぼ無いため、不要と判断できる。そこで、人手で不要単語リストを作成し、不要ノードの削除を行った。不要単語の候補の選定には、5章の実験では使用しないテーマキーワードで作成したネットワークを利用した。また、不要単語リストの単語数は約270である。以下にその一部を示す。

- 編集, 新着, 方法, 市場, 用語, 解説, 販売, 数件, 案件, 発売, 全員, 利用, 広告, 条件, 検索, 関連, 案内, 移動, 定期, メニュー, サイト, プラン, ポイント, メール, リスト, ユーザー, ボタン, コメント, ニュース, よう, さん

4.3 アンドテーマキーワードによる情報の追加取得

ネットワークの作成後、使用者が興味をもったノードや興味のある単語をアンドテーマキーワードとする。そして、テーマキーワードとアンドテーマキーワードで検索エンジンを用いてアンド検索を行い、ネットワークを再度作成する。これにより、使用者にとって興味深いネットワークやアンドテーマキーワードに沿った内容のネットワークを作成することができる。この手法をアンドテーマ手法とする。以下にアンドテーマ手法の手順を示す。

手順1 作成したいネットワークの主となる概念を示す単語をテーマキーワードとして設定し、興味のある単語をアンドテーマキーワードとして設定する。

手順2 検索エンジンを用い、テーマキーワードとアンドテーマキーワードでのアンド検索の検索結果を記事群 A として作成する。

手順3 記事群 A のテーマキーワードが含まれている記事のみを利用するために、記事群 A からテーマキーワードとアンドテーマキーワードを含む記事を抽出し、記事群 B を作成する。2 回目以降は、テーマキーワード、新たなキーワード、アンドテーマキーワードの 3 つをそれぞれ含む記事を抽出する。

手順4 記事群 B に形態素解析を用い、ノード候補となる単語を抽出する。その際にアンドテーマキーワード、1 文字、数字のみの単語を除外する。また、4.2 節の不要単語リストに含まれる単語も除外する。

手順5 得られたノード候補の中から、 $TF-IDF$ の値を用いて、その値が最も大きい 5 単語を次のノードとする。

手順6 テーマキーワード、次のノードとなった単語、アンドテーマキーワードの 3 つで検索エンジンを用いてアンド検索を行い、新しい検索結果で記事群 A を再度作成する。その後、次のノードの単語を新たなキーワードとし、手順3 より同様の処理を繰り返すことでネットワークを拡大していく。例として、テーマキーワードが「宇宙」、新たなキーワードが「飛行」、アンドテーマキーワードが「ロケット」の場合「宇宙 飛行 ロケット」で検索を行い、その検索結果を記事群 A として使用する。

4.4 DF の置き換えによるネットワークの変化

本研究ではネットワークの作成に $TF-IDF$ の値を用いるため、DF の計算に使用するデータによって、ネットワークに出現する単語が変化する。先行手法では新聞データを入力としていたため、DF の計算にも新聞データを利用していた。しかし、提案手法ではテーマキーワードの検索結果を入力とするため、DF の計算に新聞データを用いた場合、重要となる単語がネットワークに出現しない可能性がある。そこで、DF の計算に使用するデータを置き換え、どのデータがネットワークを作成するのに適しているかを調査した。

第5章 実験

5.1 主観によるネットワークの評価

4.1 節の基本手法が情報取得に有効なネットワークを作成できるかどうかを確認するために、作成したネットワークを主観で評価する。まず、テーマキーワードを 20 個設定し、4.1 節の基本手法でネットワークを作成する。そして、作成したネットワークを評価方法に基づき評価する。また、3.7 節の先行手法でもネットワークを作成し、4.1 節の基本手法との比較を行う。

5.1.1 実験条件

先行手法の入力及び DF の計算にはそれぞれ、毎日新聞 (2012 年 1 年分) を用いる。基本手法の DF の計算には Wikipedia (2016 年 10 月まで) を用いる。また、実験には 20 個のテーマキーワードを使用する。使用するテーマキーワードを以下に示す。

- エネルギー, ビタミン, 飲酒, 温泉, 科学, 花見, 観光, 義務, 喫煙, 京都, 資源, 世界遺産, 石油, 伝統, 発電, 鳥取, 中国, 哲学, 紅葉, 戦争

5.1.2 評価方法

本研究の評価では、新規で情報取得が可能なノードやリンクに注目する。一般的な内容のノードのみで構成されたネットワークでは、新しい知識となり得る情報が少なく、情報取得に繋がらない場合があるためである。

まず、作成したネットワークを確認し、新規で内容把握や発想支援に役立つとされるノードと、それに関連する情報をもつノードの個数を数える。数えたノードの数が 1, 3, 5 個以上かそれぞれ確認し、情報取得に有効なネットワークが作成されたかを評価する。役立つとされるノードを判断する基準を以下に示す。この基準は、どちらかまたは両方満たせばよいものとする。

- ノードについて内容を他人に説明できる程は知らず，そのノードがテーマキーワードに関する有用な情報である．
- ノードのリンクにより，テーマキーワードに関する新たな情報が得ることができる．また，参照しているリンクの各ノードを，それぞれ役立つノードとして数えてもよい．

例として，テーマキーワード「宇宙」の場合，地球軌道上の宇宙望遠鏡であるハッブル望遠鏡の「ハッブル」が挙げられる．なぜなら「ハッブル」というノードは一般的によく知られる「ブラックホール」などといったノードとは異なり，ネットワークを見た人にとって新たな知識となる可能性が高いためである．

また，スニペットはネットワークの評価と類似した評価方法で評価する．スニペットを読んで，新規で内容把握や発想支援に役立つとされる単語を数える．単語の区切りは形態素解析ではなく，主観で判断する．

5.1.3 スニペットを用いた評価方法

基本手法の有用性についても確認するため，スニペットを評価に用いる．スニペットとは，検索結果の一部として表示される Web サイトの要約文である．スニペットは検索エンジンである Bing [10] を用い，テーマキーワードでの検索結果の上位 5 件分を使用する．スニペット上位 5 件分の合計文字数は，ネットワークのノードの合計文字数の約 2 倍ほどである．テーマキーワード「宇宙」の場合のスニペットの例を表 5.1 に示す．

表 5.1: テーマキーワード「宇宙」のスニペット例

| |
|---|
| 宇宙（うちゅう）とは、以下のように定義される。コスモス。時間・空間内に秩序をもって存在する「こと」や「もの」の総体。何らかの観点から見て、秩序をもつ完結した世界体系。全ての時間と空間、およびそこに含まれるエネルギーと物質。 |
| 銀河系（ぎんがけい、英: Milky Way）は、宇宙に数ある銀河の中でも、人類の住む地球・太陽系を含む銀河の名称である。局部銀河群に属する。以前は渦巻銀河の一種と考えられていたが、20 世紀末以降は棒渦巻銀河であるとする説が有力になりつつある。 |

また，作成したネットワークと比較を行うため，5.1.2 節と類似する評価方法でスニペットを評価する．

まず、スニペットを読んで、新規で内容把握や発想支援に役立つとされる単語と、それに関連する情報をもつ単語の個数を数える。なお、単語の区切りは主観で判断するものとする。数えた単語の数が1, 3, 5個以上かそれぞれ確認し、情報取得に有効なスニペットであるか評価する。役立つとされる単語を判断する基準を以下に示す。この基準は、どちらかまたは両方満たせばよいものとする。

- 単語についての内容を他人に説明できる程は知らず、その単語がテーマキーワードに関する有用な情報である。
- テーマキーワードに関する単語であり、その単語が含まれている文章を読むことで新規の情報を得ることができる。また、その単語に関連し、同じ文章に含まれる他の単語も役立つ単語として数えてもよい。

例として、上記のスニペットからは「局部銀河群」「渦巻銀河」が挙げられる。なぜなら「空間」「太陽系」といった単語より、「局部銀河群」「渦巻銀河」といった単語は、スニペットを見た人にとって新たな知識となる可能性が高いためである。

5.1.4 実験結果

作成した先行手法と基本手法のネットワークの評価結果を、スニペット法の評価結果と合わせて表5.2に示す。表5.2には、手法ごとにテーマキーワード20個を用いて作成したネットワーク及びスニペットが、一定個数以上の新規の情報取得に役立つノード（単語）を持つ比率を個数別に示している。また、テーマキーワードを「鳥取」とした場合に、先行手法で作成したネットワークを図5.1に、基本手法で作成したネットワークを図5.2に示し、テーマキーワードを「石油」とした場合に、先行手法で作成したネットワークを図5.3に、基本手法で作成したネットワークを図5.4に示す。テーマキーワードを「鳥取」としたネットワークは、基本手法において比較的多くの新規の情報取得に役立つノードを取得できた例であり、テーマキーワードを「石油」としたネットワークは、基本手法においてあまり新規の情報取得に役立つノードを取得できなかった例である、スニペットは、テーマキーワード「鳥取」を表5.3に、テーマキーワード「石油」を表5.4に示す。

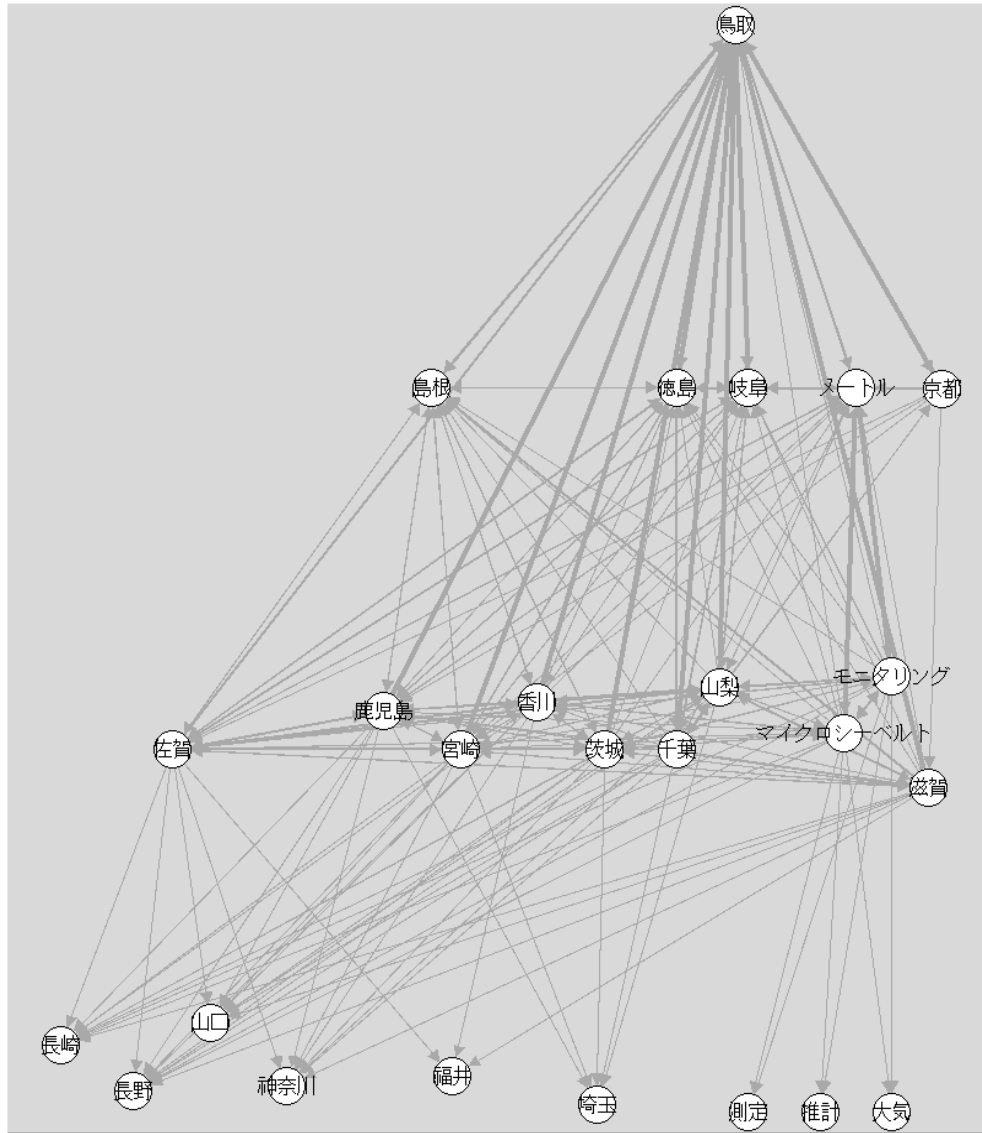
図5.2の基本手法で作成したネットワークは、新規の情報取得に役立つノードの個数が5個以上であるため、表5.2の役立つノード（単語）数が5個以上の部分に分類され

る．なお，図 5.1 の先行手法で作成したネットワークは，新規の情報取得に役立つノードの個数が 0 個となっている．

表 5.3 のスニペットは，新規の情報取得に役立つ単語の個数が 4 個であるため，表 5.2 の役立つノード（単語）数が 3 個以上の部分に分類され，表 5.4 のスニペットは，新規の情報取得に役立つ単語の個数が 2 個であるため，表 5.2 の役立つノード（単語）数が 1 個以上の部分に分類される．

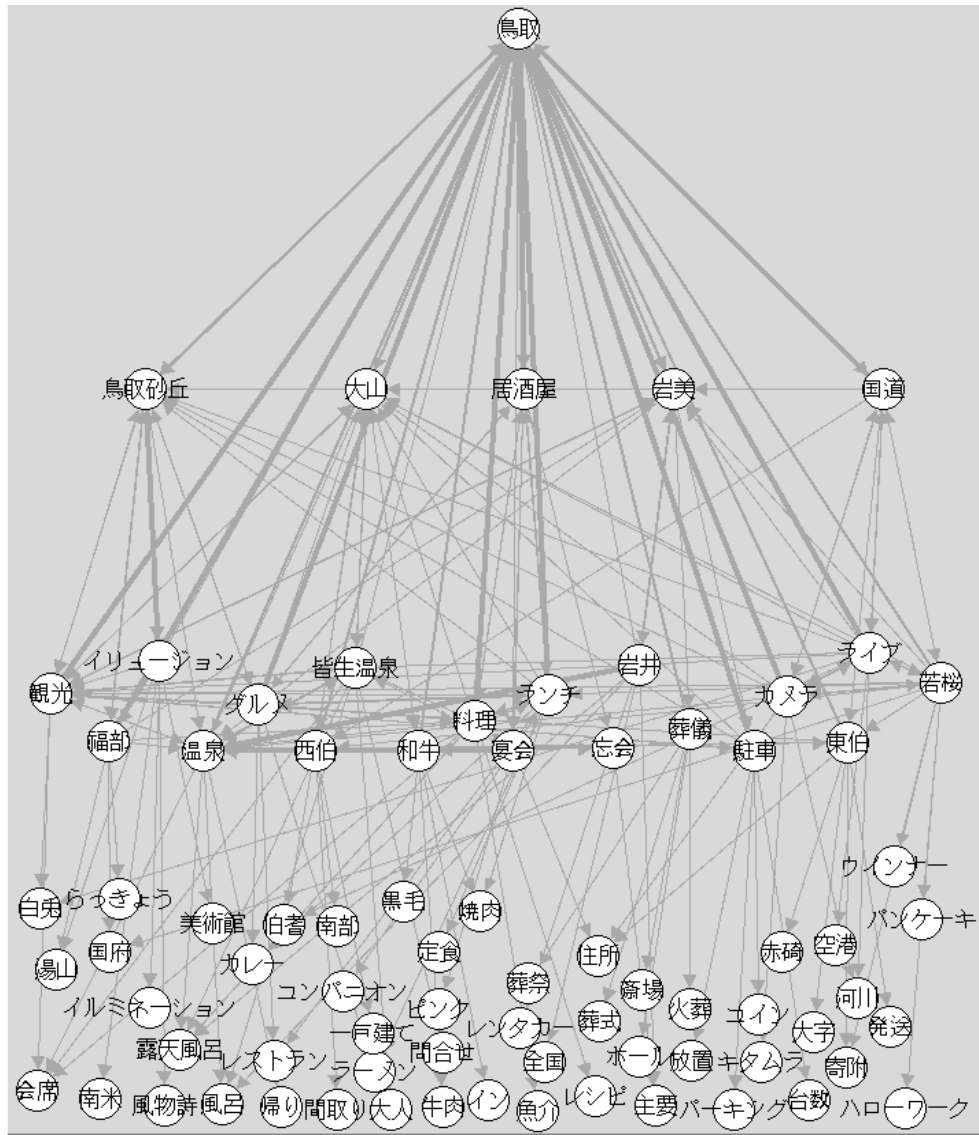
表 5.2: 基本手法における一定個数以上の新規の情報取得に役立つノード（単語）を持つネットワーク及びスニペットの比率

| 手法 | 役立つノード（単語）数 | | |
|-----------------|-------------|-------------|-------------|
| | 1 個以上 | 3 個以上 | 5 個以上 |
| 先行手法（新聞） | 0.85(17/20) | 0.60(12/20) | 0.35(7/20) |
| 基本手法（Web） | 1.00(20/20) | 0.85(17/20) | 0.80(16/20) |
| スニペット法（Web 要約文） | 0.60(12/20) | 0.25(5/20) | 0.05(1/20) |



- 役立つとしたノード：なし

図 5.1: 先行手法によるテーマキーワード「鳥取」のネットワーク

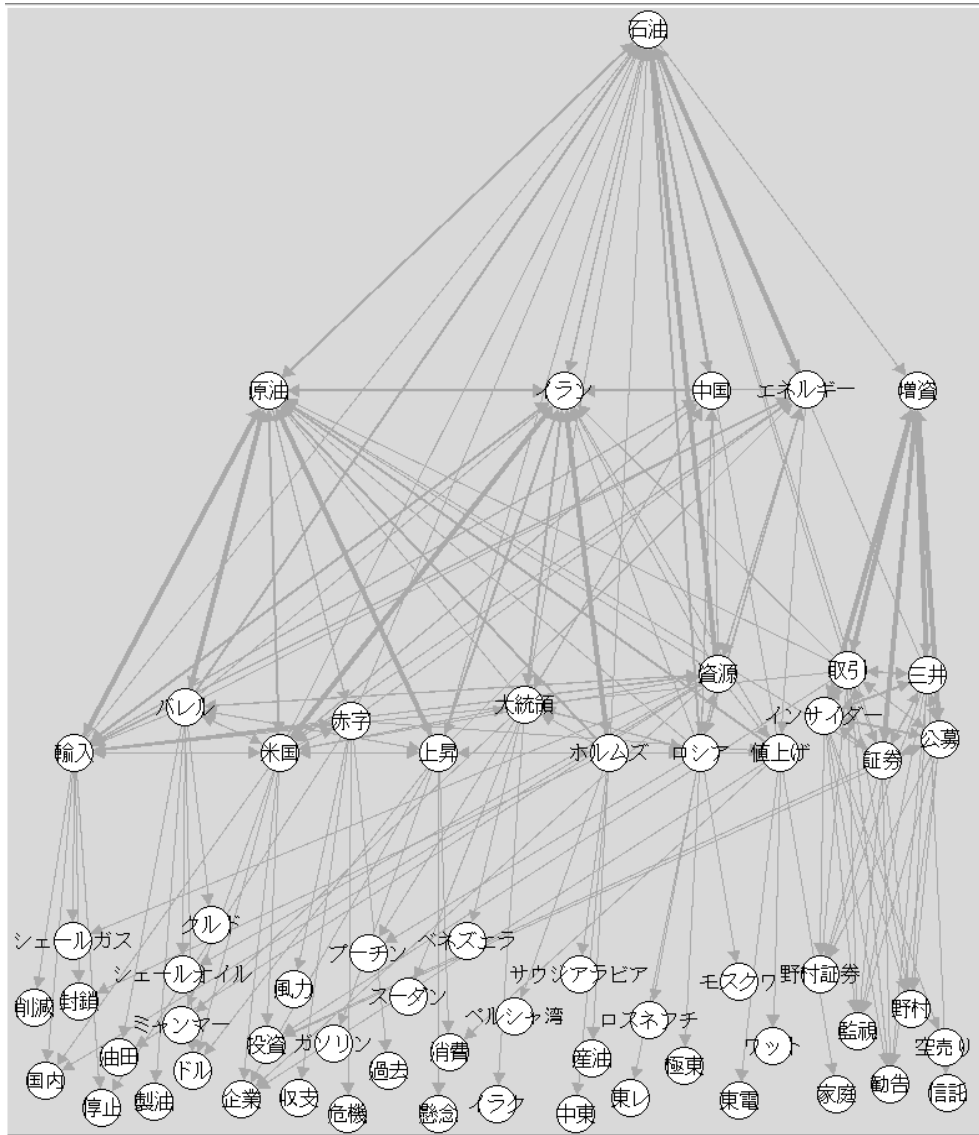


- 役立つとしたノード：鳥取砂丘，大山，福部，皆生温泉，らっきょう

図 5.2: 基本手法によるテーマキーワード「鳥取」のネットワーク

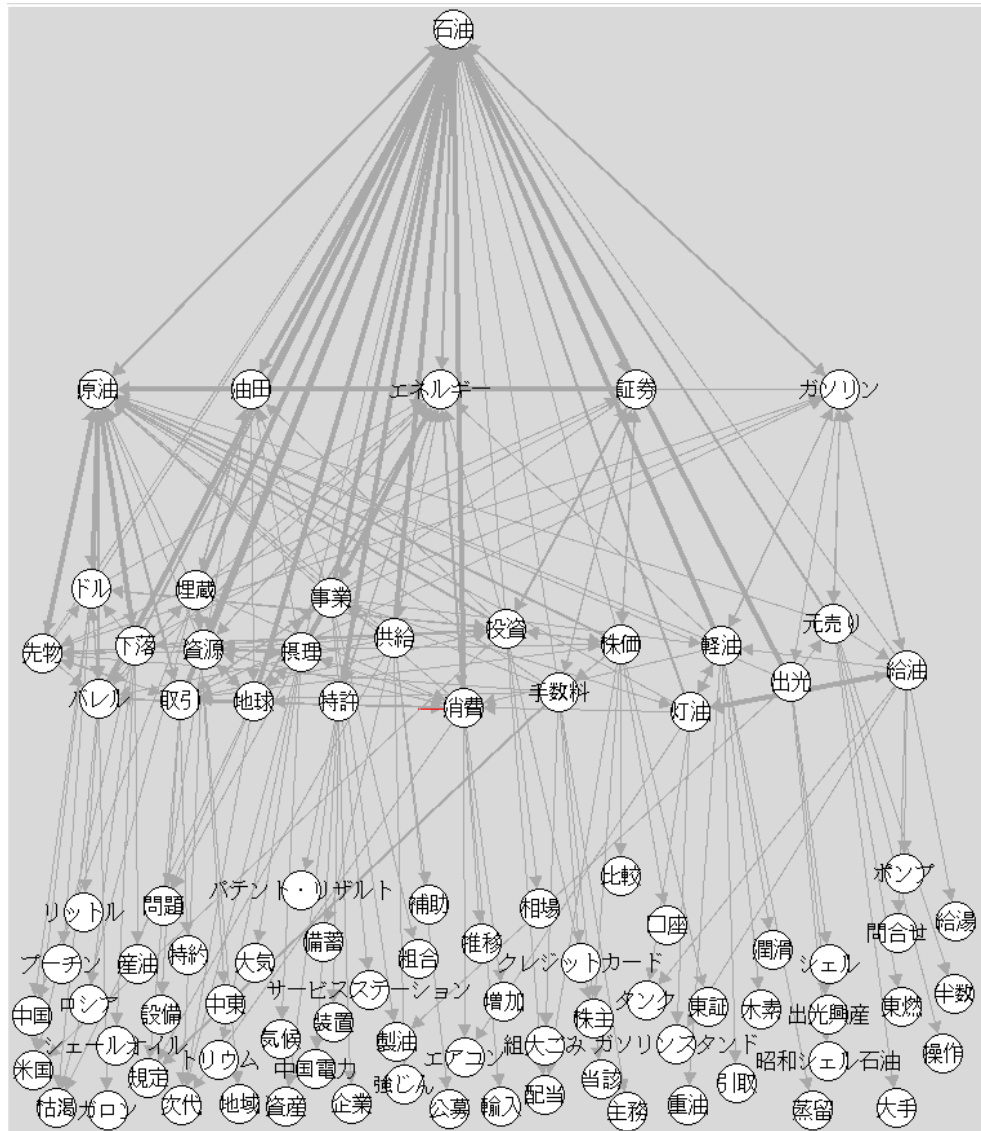
表 5.3: テーマキーワード「鳥取」のスニペット

| |
|---|
| 鳥取砂丘 鳥取の代名詞でもある日本最大の海岸砂丘、鳥取砂丘。風が吹くたびに変わる風紋や砂簾(されん)は、まさに自然が作りだす芸術です。 詳細ページへ 浦富海岸 山陰海岸国立公園内にあり、洞門や洞窟、白砂の浜など、変化 ... |
| 東京ビジネスセンターの利用者を募集します!「鳥取県知的財産マネジメント委員会委員」を公募します【参加募集】鳥取県ビジネスプランコンテスト表彰式 & とっとり起業フォーラム 平成 28 年度鳥取県立中央病院「第 4 回市民講座 ... |
| 鳥取県内各地の平年値 (統計期間: 1971 年 - 2000 年、出典: 気象庁・気象統計情報) 平年値 (月単位) 沿岸部 内陸部 岩美町 岩井 鳥取 鳥取市 湖山 鳥取市 青谷 倉吉 大山町 塩津 米子 境港 智頭 日南町 茶屋 平均 気温 (C) 最暖月 |
| ・【鳥取市雇用創造協議会】各種セミナー・人材育成研修のおしらせ (2017/1/12) ・第 10 回鳥取市地産地消フェアの開催について (2017/1/12) ・鳥取市中小企業・小規模企業振興条例 (案) に係る市民政策コメントの実施結果について |
| 「鳥取市は元気です」地震に負けず頑張っています! 10 月 21 日午後の「鳥取県中部地震」で、鳥取市は震度 5 弱を観測しましたが、当協会の被害はありませんでした。 鳥取砂丘周辺も被害はなく平常通り ... |
| 役立つとした単語: 鳥取砂丘, 砂簾, 山陰海岸国立公園, 浦富海岸 |



- 役立つとしたノード：ホルムズ，シェールガス，シェールオイル，ペルシャ湾，産油

図 5.3: 先行手法によるテーマキーワード「石油」のネットワーク



- 役立つとしたノード：シェールオイル

図 5.4: 基本手法によるテーマキーワード「石油」のネットワーク

表 5.4: テーマキーワード「石油」のスニペット

| |
|--|
| <p>概要 「石油」は沈括の『夢溪筆談』からの言葉 [1]。英語で石油は「Petroleum」という。これはラテン語の Petra (岩石) と Oleum (油) を語源とする。狭義には天然の原油 (crude oil) のことを指すが、より広い意味では天然ガスや固体の ...</p> |
| <p>石油は、英語で「Petroleum」と書きます。この語源は、ラテン語の「岩石の油」に由来すると言われ、石油の採掘過程をも表現しています。つまり、ラテン語が主言語として使用されていた時代から掘り当てられていたのです。</p> |
| <p>石油情報センターは、公平かつ公正な立場で、一般消費者、産業界の石油製品ユーザー、石油関連事業者などに、石油に関する情報を様々な形で提供しています。</p> |
| <p>石油連盟は、わが国の石油精製・元売会社、すなわち原油の輸入、精製、石油製品の全国的な販売を行っている企業の団体として創立された基幹的産業団体です。</p> |
| <p>2014/10/10 なぜか増えていく石油「埋蔵量」の秘密 人類は地中の資源の全貌をまだ知らない ... 私たちの生活はエネルギーなしでは成り立たない。しかし私たちはエネルギーのことを知らなすぎる。</p> |
| <p>役立つとした単語：沈括，夢溪筆談</p> |

5.2 複数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情報取得に有効であるか確認するため，5.1 節で作成したネットワークとスニペットの評価実験を実施した．

5.2.1 実験条件

被験者実験は6名の被験者に実施した．被験者には先行手法と基本手法で作成したネットワークをテーマごとにシャッフルした状態で見せ，評価させる．また，スニペットもテーマキーワード5つ分を評価させる．

先行手法の入力及びDFの計算にはそれぞれ，毎日新聞（2012年1年分）を用いる．基本手法のDFの計算にはWikipedia（2016年10月まで）を用いる．また，テーマキーワードは5.1.1節の20個のうち，以下の5つを使用する．

- エネルギー，温泉，紅葉，世界遺産，科学

5.2.2 評価方法

5.1.2節と同様の評価方法を用いる．

5.2.3 実験結果

被験者実験の評価結果を，表5.5に示す．表5.5には，各手法ごとにテーマキーワード20個を用いて作成したネットワーク及びスニペットが，一定個数以上の新規の情報取得に役立つノード（単語）を持つ比率を個数別に示している．

表 5.5: 被験者実験における一定個数以上の新規の情報取得に役立つノード（単語）を持つネットワーク及びスニペットの比率

| 手法 | 役立つノード（単語）数 | | |
|-----------|-------------|-------------|-------------|
| | 1個以上 | 3個以上 | 5個以上 |
| 先行手法（新聞） | 0.93(28/30) | 0.50(15/30) | 0.27(8/30) |
| 基本手法（Web） | 1.00(30/30) | 0.67(20/30) | 0.40(12/30) |
| スニペット法 | 0.70(21/30) | 0.27(8/30) | 0.03(1/30) |

5.3 アンドテーマ手法の評価

テーマキーワードのみで作成されたネットワークを確認し、情報をより引き出すことができるアンドテーマキーワードを設定した。これにより、アンドテーマキーワードに沿った追加の情報を取得する。作成したネットワークは評価方法に基づき評価し、テーマキーワードのみで作成したネットワークと比較する。また、5.1 節と同様にスニペットとも比較する。スニペットはテーマキーワードとアンドテーマキーワードを Bing [10] でアンド検索し、検索結果の上位 5 件分を用いる。

5.3.1 実験条件

基本手法とアンドテーマ手法の DF の計算には Wikipedia (2016 年 10 月まで) を用いる。また、使用するテーマキーワードとアンドテーマキーワードを以下に示す。ここでは、テーマキーワードが「宇宙」、アンドテーマキーワードが「飛行」の場合、「宇宙 + 飛行」と表記する。

- 宇宙 + 銀河, 環境 + 地球, 世界遺産 + 遺跡, 京都 + 寺, 鳥取 + 鳥取砂丘, 観光 + 温泉, 資源 + 海底, 紅葉 + 名所, 伝統 + 和紙, 発電 + 風力

5.3.2 評価方法

5.1.2 節の評価方法と同様に評価する。ただし、テーマキーワードのみで作成したネットワークに含まれる単語は、役立つノードとして数えないこととする。

5.3.3 実験結果

アンドテーマ手法で作成したネットワークの評価結果を、表 5.6 に示す。表 5.6 には、各手法ごとにテーマキーワード 20 個を用いて作成したネットワーク及びスニペットが、一定個数以上の新規の情報取得に役立つノード (単語) を持つ比率を個数別に示している。また、テーマキーワードを「京都」とし、基本手法で作成したネットワークを図 5.5 に示し、テーマキーワードを「京都」とし、アンドテーマキーワードを「寺」としてアンドテーマ手法で作成したネットワークを図 5.6 に示す。同様に、テーマキーワードを「世界遺産」とし、基本手法で作成したネットワークを図 5.7 に示し、テーマキーワードを「世界遺産」とし、アンドテーマキーワードを「遺跡」としてアンドテーマ手法で作

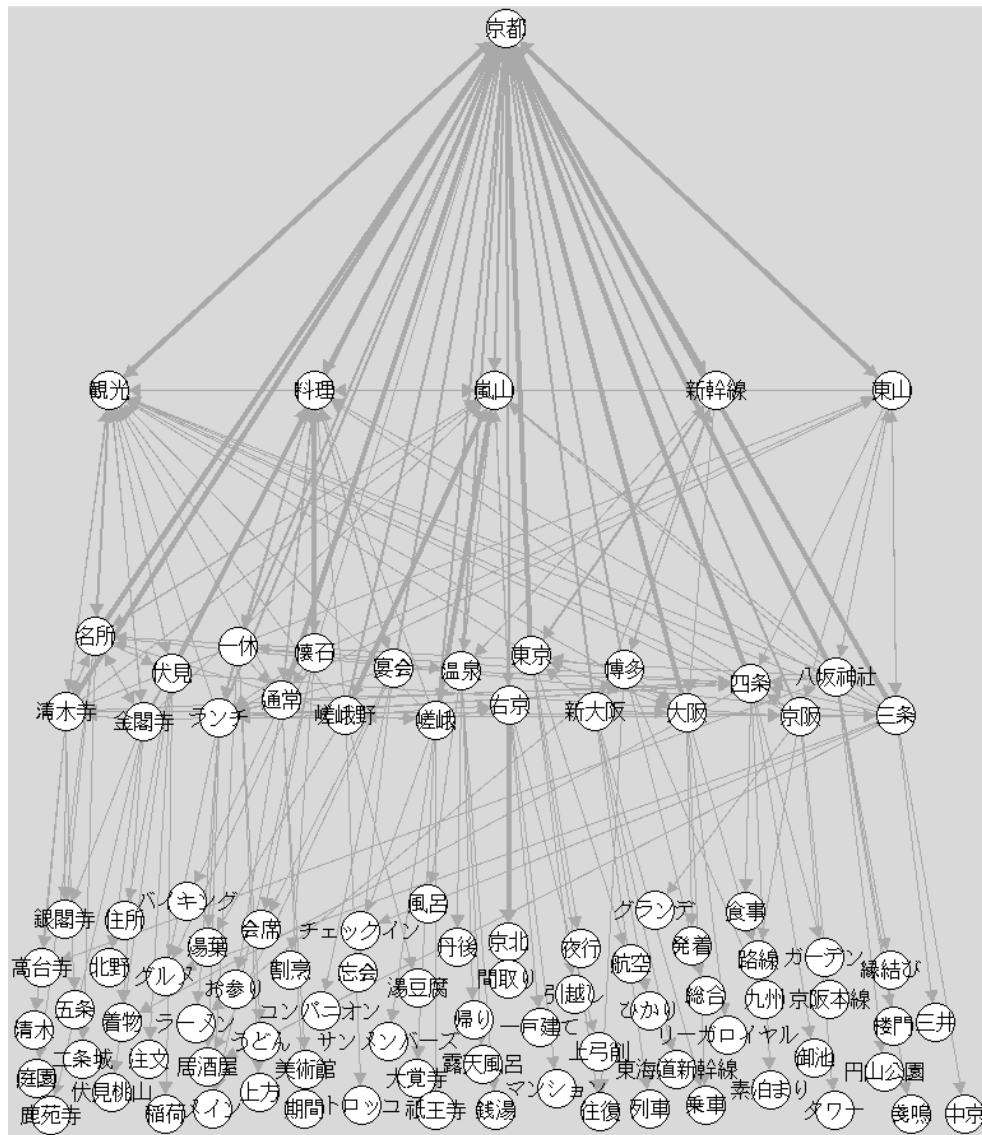
成したネットワークを図 5.8 に示す。さらに、ノードの取得テーマキーワードを「宇宙」とし、基本手法で作成したネットワークを図 5.9 に示し、テーマキーワードを「宇宙」とし、アンドテーマキーワードを「銀河」としてアンドテーマ手法で作成したネットワークを図 5.10 に示す。テーマキーワードを「京都」、アンドテーマキーワードを「寺」としたネットワークと、テーマキーワードを「世界遺産」、アンドテーマキーワードを「遺跡」としたネットワークは、アンドテーマ手法において比較的多くの新規の情報取得に役立つノードを取得できた例である。テーマキーワードを「宇宙」、アンドテーマキーワードを「銀河」としたネットワークは、アンドテーマ手法においてあまり新規の情報取得に役立つノードを取得できなかった例である。

図 5.5, 図 5.7, 図 5.9 の基本手法で作成したネットワークと図 5.6, 図 5.8 のアンド手法で作成したネットワークは、それぞれ新規の情報取得に役立つノードの個数が 5 個以上であるため、表 5.6 の役立つノード(単語)数が 5 個以上の部分に分類される。図 5.10 のアンド手法で作成したネットワークは、新規の情報取得に役立つノードの個数が 1 個であるため、表 5.6 の役立つノード(単語)数が 1 個以上の部分に分類される。

表 5.7, 表 5.11 のスニペットは、新規の情報取得に役立つ単語の個数が 1 個であるため、表 5.6 の役立つノード(単語)数が 1 個以上の部分に分類され、表 5.8, 表 5.12 のスニペットは、新規の情報取得に役立つ単語の個数が 5 個であるため、表 5.6 の役立つノード(単語)数が 5 個以上の部分に分類される。表 5.9 のスニペットは、新規の情報取得に役立つ単語の個数が 3 個であるため、表 5.6 の役立つノード(単語)数が 3 個以上の部分に分類され、表 5.10 のスニペットは、新規の情報取得に役立つ単語の個数が 0 個である。

表 5.6: アンドテーマ手法における一定個数以上の新規の情報取得に役立つノード(単語)を持つネットワーク及びスニペットの比率

| 手法 | 役立つノード(単語)数 | | |
|----------|-------------|------------|------------|
| | 1 個以上 | 3 個以上 | 5 個以上 |
| 基本手法 | 0.90(9/10) | 0.80(8/10) | 0.80(8/10) |
| アンドテーマ手法 | 1.00(10/10) | 0.80(8/10) | 0.60(6/10) |
| スニペット法 | 0.80(8/10) | 0.40(4/10) | 0.40(4/10) |

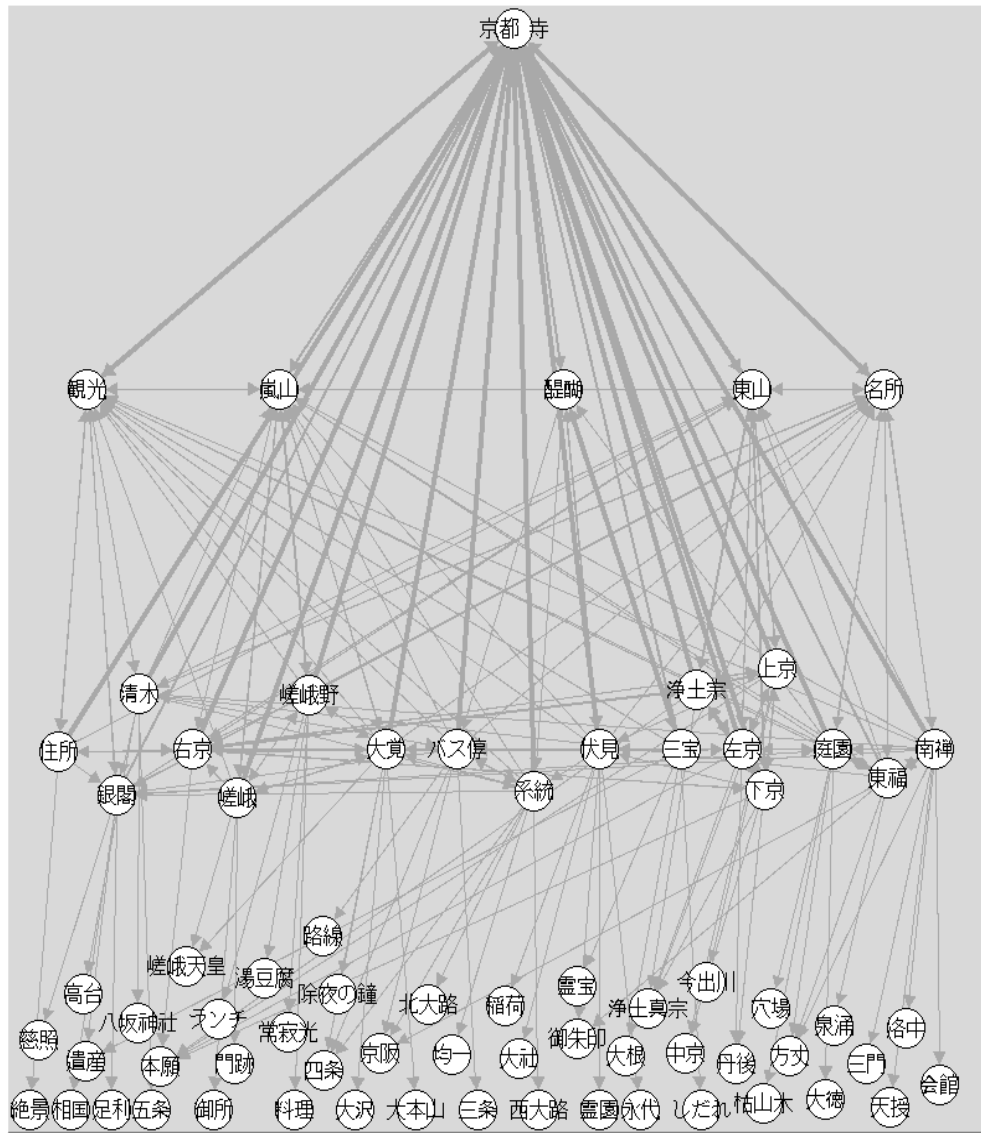


- 役立つとしたノード：東山，嵯峨，八坂神社，大覚寺，縁結び

図 5.5: 基本手法によるテーマキーワード「京都」のネットワーク

表 5.7: テーマキーワード「京都」のスニペット

| |
|---|
| JR 東海の京都観光情報【そうだ 京都、行こう。公式サイト】。京都の風景、お寺や神社、イベント、お祭り、お食事やお土産などの情報をご案内。京都が楽しくなる特集記事やインタビューも。 |
| 2016/05/19 関西に住んでいる作者がおススメの京都の観光名所をまとめました定番から穴場のスポットを随時まとめていきます！是非、京都観光の参考にして下さい。 |
| 1月6日「京都観光関連産業正規雇用化促進事業（多能力人材等育成訓練実施業務）」に係る公募型プロポーザルの実施について1月6日 宇治茶世界文化遺産登録推進シンポジウム～アジアから見る「宇治茶の文化的景観」～を開催！ |
| 奥深い京都の魅力をお伝えするため、数多くの京都観光情報をお伝えしています。イベント情報や花だより、施設情報など、京都観光に役立つ情報満載です。 |
| 京都（きょうと、みやこ、きょうのみやこ、英: Kyoto [1] [2] [3] ）は、日本の都市の1つである。都、京とも言い、歴史的には794年に日本の首都に定められた都城・平安京で、当時は日本の政治・文化の中心地であった。 |
| 役立つとした単語：宇治茶 |



- 役立つとしたノード：醍醐，三宝，東福，御朱印，泉桶

図 5.6: アンドテーマ手法によるテーマキーワード「京都」，アンドテーマキーワード「寺」のネットワーク

表 5.8: テーマキーワード「京都」, アンドテーマキーワード「寺」のスニペット
旅行者のクチコミをもとにした人気ランキングから京都の寺・神社情報を探す
ことができます。日本最大級の旅行クチコミサイト フォートラベルで京都寺・神社
情報をチェック！

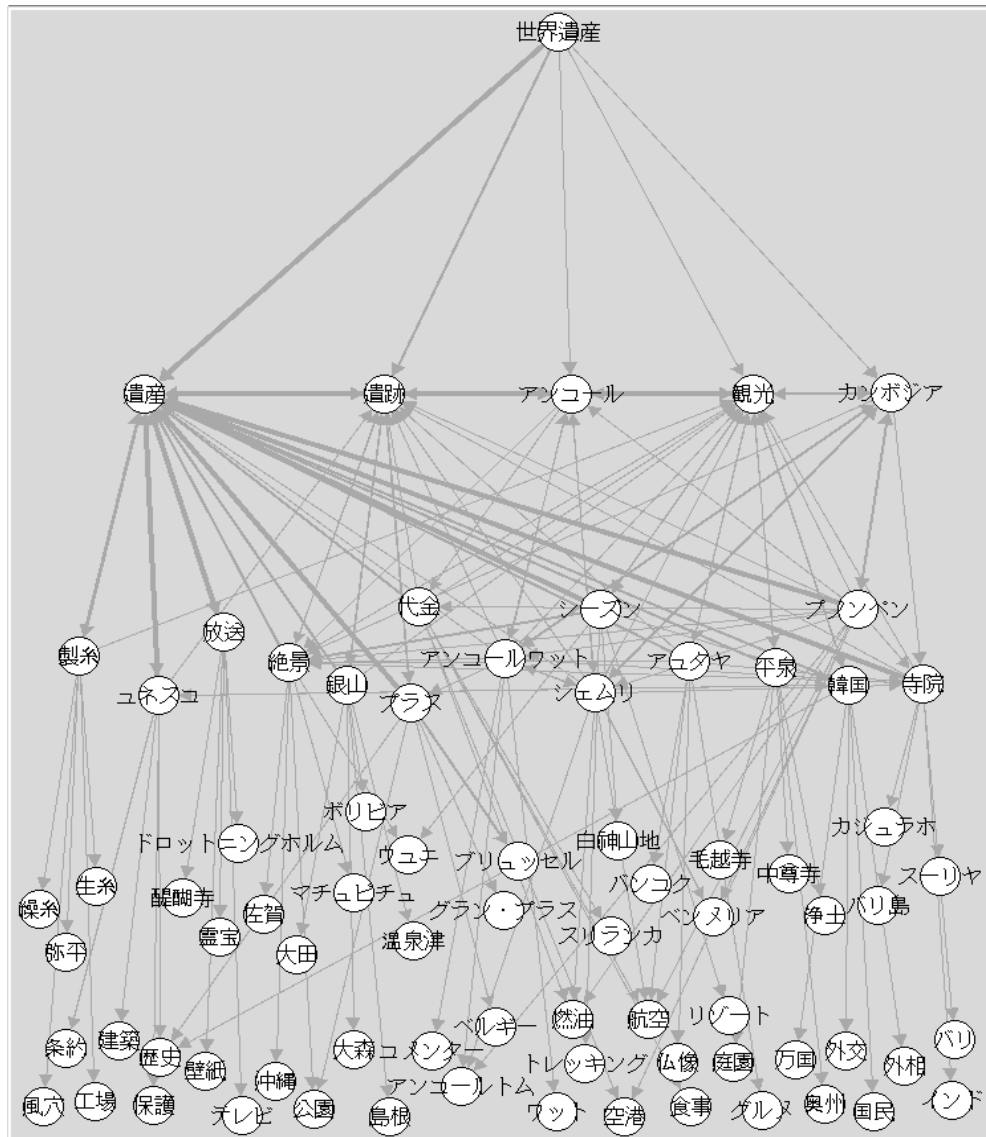
清水寺 京都の名所、東山を旅行。豊臣家ゆかりの神社仏閣が集まり、三十三間
堂、清水寺、知恩院、高台寺と坂道を散策。日本の食と文化の源・祇園。白川沿
いは料亭から漏れる光と街灯の淡い光が川面を照らし、京都情緒 ...

京都の寺・神社情報をご紹介します。... 南禅寺(ナンゼンジ) 京都五山の
上位に列せられた禅刹。歌舞伎にもある石川五右衛門の伝説で有名な三門など、
堂塔伽藍が並ぶ。

2017/01/16 京都観光といえば、やはり寺社。京都には色々な寺社があつてどこ
いけばいいか迷う人も多いのではないのでしょうか？今回はそんな時にオススメな
京都の寺社をレキジョならぬ歴男な筆者がランキング形式丁寧にとまとめました。

京都の寺の一覧です。このサイトの記事に記載した京都の寺を一覧表にしてみま
した。建物や庭や花など何らかの特徴で有名な寺の一覧です。

役立つとした単語：三十三間堂，南禅寺，禅刹，知恩院，堂塔伽藍



- 役立つとしたノード：カンボジア，寺院，マチュピチュ，白神山地，カジュラホ

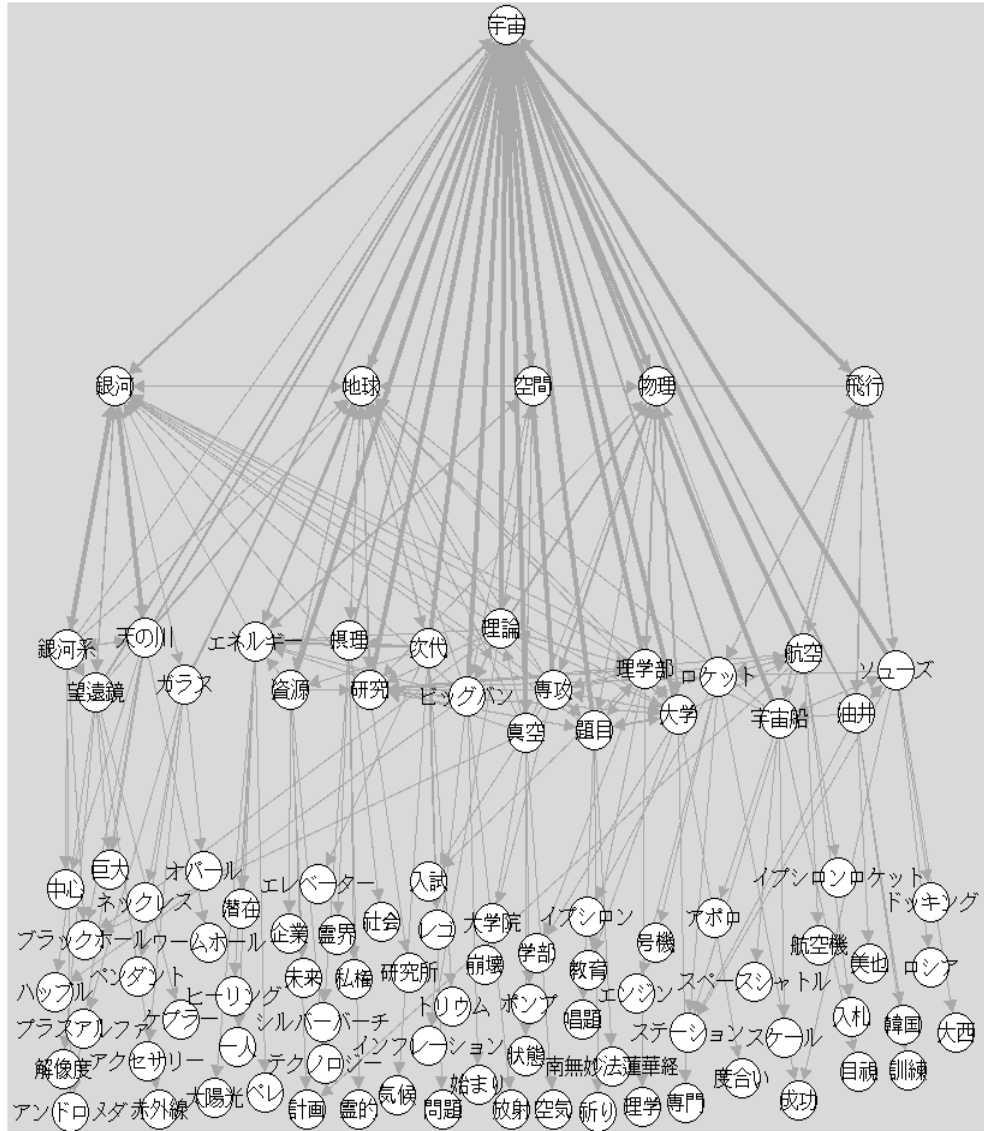
図 5.7: 基本手法によるテーマキーワード「世界遺産」のネットワーク

表 5.9: テーマキーワード「世界遺産」のスニペット

| |
|---|
| <p>世界遺産への貢献のため、あなたのご支援が必要です よくお寄せいただく質問とその答をまとめました。 世界遺産活動 世界遺産一覧 地図上の地域名をクリックすると地域別、国別の世界遺産リストをご確認いただけます。 2015年12...</p> |
| <p>世界遺産(せかいいさん)は、1972年のユネスコ総会で採択された「世界の文化遺産及び自然遺産の保護に関する条約」(世界遺産条約)に基づいて世界遺産リストに登録された、遺跡、景観、自然など、人類が共有すべき「顕著な...</p> |
| <p>毎週日曜日午後6時から放送中のTBSテレビ「THE世界遺産」公式サイトです。過去の放送内容、世界遺産の一覧などはこちらから! ... 次回予告 2017年1月8日 鉄道で登る! アルプス名峰と氷河 ~ スイス・アルプス ユングフラウ-アレッチ...</p> |
| <p>日本の世界遺産(にっぽんのせかいいさん、にほんのせかいいさん)は、ユネスコの世界遺産に登録されている日本国内の文化・自然遺産の一覧。Category:日本の世界遺産に索引がある。</p> |
| <p>世界遺産 ランキング一覧 世界遺産写真画像 世界遺産登録リスト名 国名 種別 ランキング 1位 アンコール遺跡群 カンボジア 世界遺産 一覧/地図 世界文化遺産 順位をつける 世界三大仏跡の一つで、神のオーラを感じることができる。</p> |
| <p>役立つとした単語: 世界三大仏跡, アンコール遺跡群, カンボジア</p> |

表 5.10: テーマキーワード「世界遺産」、アンドテーマキーワード「遺跡」のスニペット

| |
|--|
| 「遺跡へ行くぞ!」は、日本最古の旅行者向けの遺跡ホームページです。世界遺産に登録されている歴史的建造物や、登録されていないけれど貴重な遺跡など、ありとあらゆる歴史的スポットを紹介しています。このサイトの最大の ... |
| 2012/07/22 Google マップで見ることができる遺跡や史跡や世界遺産をまとめてみました。Google アース用の KML ファイルもダウンロードできます。更新日: 2012 年 07 月 22 日 |
| 世界遺産 ランキング一覧 世界遺産写真画像 世界遺産登録リスト名 国名 種別 ランキング 1 位 アンコール遺跡群 カンボジア 世界遺産 一覧/地図 世界文化遺産 順位をつける 世界三大仏跡の一つで、神のオーラを感じることができる。 |
| 世界各地の世界文化遺産・世界自然遺産へ行こう! 各地の世界遺産や遺跡、名所など人気おすすめスポットをピックアップ! 世界遺産旅行・ツアーを楽しみたいならエイビーロードで! |
| 2014/04/22 世界には多くの遺跡がありますが、その中に奇妙な共通点があります。石で造られた古代遺跡の奇妙な共通点についてまとめます。更新日: 2014 年 04 月 22 日 |
| 役立つとした単語: なし |

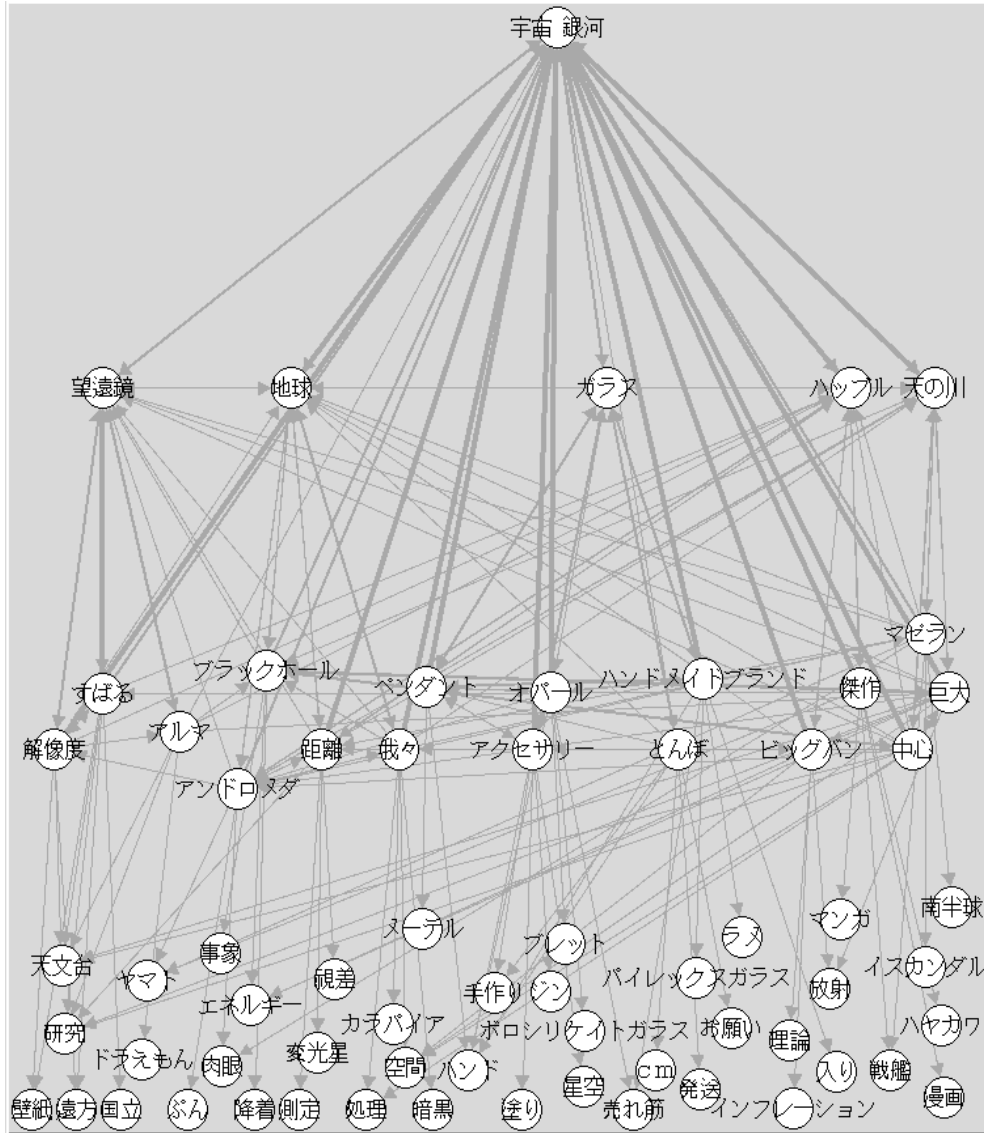


- 役立つとしたノード：望遠鏡，ビッグバン，ハッブル，ケプラー，イプシロンロケット

図 5.9: 基本手法によるテーマキーワード「宇宙」のネットワーク

表 5.11: テーマキーワード「宇宙」のスニペット

| |
|---|
| <p>宇宙について説明するにあたり、まず人類がどのように宇宙の理解を深めてきたか、おおまかな流れを解説する。宇宙がいかに始まったか。この議論は宗教や哲学上の問題として語られて続けている [5]。宇宙に関する説・研究などは ...</p> |
| <p>2013/12/23 宇宙の事を考えると、壮大すぎて人生がちっぽけに思えてきますね。 更新日: 2013年12月23日 ... 暗黒物質（あんこくぶっしつ、dark matter）とは、宇宙にある星間物質のうち電磁相互作用をせずかつ色電荷を持たない、光学的 ...</p> |
| <p>2017/01/20 「宇宙の果てに近い、たいへん遠い天体が見つかった」というようなニュースを目にすることがあります。「宇宙の果て」という言葉から考えて、「そこまでいくと、そこから先には宇宙がなくなってしまうような境目」と ...</p> |
| <p>2015/06/01 太陽の300兆倍、宇宙一明るい銀河を発見 初期宇宙に巨大ブラックホールの謎 ... 新たに発見された銀河 WISE J224607.57-052635.0 のイラスト。太陽の300兆倍以上という、宇宙一の明るさを誇る。(Illustration by NASA)</p> |
| <p>2016/06/24 謎に包まれている宇宙。宇宙についてゾッとする話をまとめました 更新日: 2016年06月24日 ... 出典【宇宙】おまえらが宇宙の凄い所を語って俺を眠らせなくするスレ 宇宙&物理 2ch まとめ</p> |
| <p>役立つとした単語：暗黒物質</p> |



- 役立つとしたノード：マゼラン

図 5.10: アンドテーマ手法によるテーマキーワード「宇宙」, アンドテーマキーワード「銀河」のネットワーク

表 5.12: テーマキーワード「宇宙」, アンドテーマキーワード「銀河」のスニペット

| |
|---|
| <p>銀河は、主に楕円型・渦巻型（渦巻・棒渦巻）・レンズ状を含む不定形がある [19]。ハッブル分類はこれをより包括的に記述した分類である [19]。しかしあくまで外観上の特徴を捉えた考察であるため、スターバースト銀河のように星形 ...</p> |
| <p>太陽系を飛び出して近傍銀河までの距離を見してみる 太陽は恒星。水星、金星、地球、火星、木星、土星、天王、海星は惑星。太陽系は、太陽の重力の影響が及ぼす範囲までなので（異説多数あり）、太陽以外の恒星は、当然、太陽系 ...</p> |
| <p>銀河系（ぎんがけい、英: Milky Way）は、宇宙に数ある銀河の中でも、人類の住む地球・太陽系を含む銀河の名称である [6] [7]。局部銀河群に属する。以前は渦巻銀河の一種と考えられていたが、20 世紀末以降は棒渦巻銀河であるとする説 ...</p> |
| <p>2015/06/01 1000 億個、あるいはそれ以上の銀河がひしめく宇宙で、最大の輝きを放つ銀河が見つかった。新たに発見された銀河 WISE J224607.57-052635.0 の明るさは、太陽の 300 兆倍を超える。米 NASA ジェット推進研究所（JPL）の天 ...</p> |
| <p>埋め込まれている動画【宇宙ヤバイ】我々がいる銀河系は、直径 5 億光年の「超銀河団」の端に位置する可能性が高い ... その超銀河団がさらに無数にあんだろ？地球人の他にも知的生命体がない（いなかった）わけないよな</p> |
| <p>役立つとした単語：ハッブル, スターバースト銀河, 渦巻銀河, 超銀河団, 局部銀河群</p> |

5.4 DFの置き換えによる変化の評価

5.4.1 実験条件

DFの計算に使用するデータを以下に示す。

- Wikipedia (2016年10月まで, 2101083記事)
- 毎日新聞 (2012年1年分, 110587記事)
- IDF=1 (DFの計算にデータを用いず, IDFの値を常に1として実験を行う)

また, 使用するテーマキーワードを以下に示す。

- イタリア, ギリシャ, 宇宙, 円安, 健康, 政策, 地震, 将棋, 通貨, 名産,

5.4.2 評価方法

5.1.2節の評価方法と同様に評価する。

5.4.3 実験結果

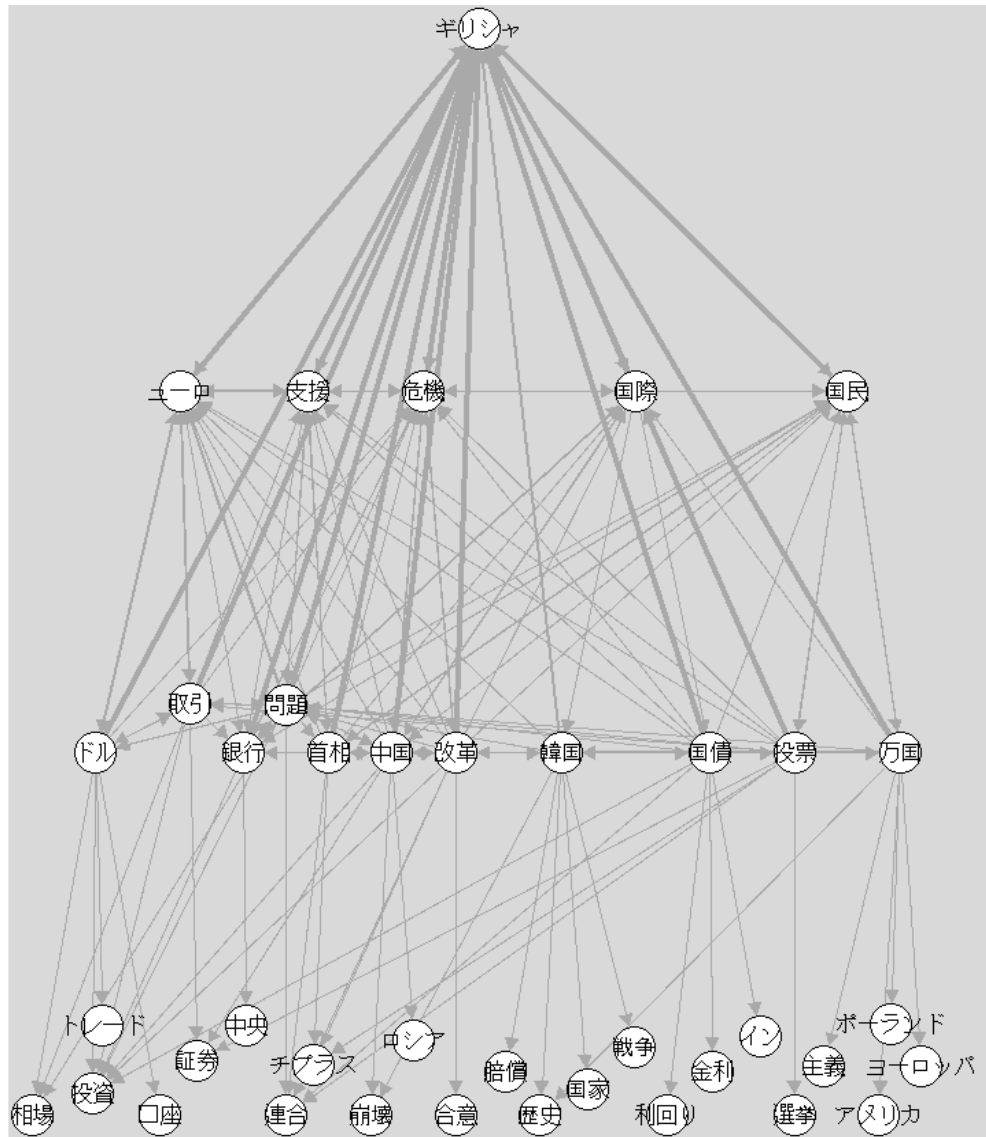
DFの計算を置き換えて作成したネットワークの評価結果を, 表5.13に示す。表5.13には, 各手法ごとにテーマキーワード20個を用いて作成したネットワーク及びスニペットが, 一定個数以上の新規の情報取得に役立つノード(単語)を持つ比率を個数別に示している。また, テーマキーワードを「ギリシャ」とした場合に, IDF=1で作成したネットワークを図5.11に, 基本手法で作成したネットワークを図5.12に示す。

図5.12の基本手法で作成したネットワークは, 新規の情報取得に役立つノードの個数が5個以上であるため, 表5.13の役立つノード(単語)数が5個以上の部分に分類される。なお, 図5.11の先行手法で作成したネットワークは, 新規の情報取得に役立つノードの個数が0個となっている。

表5.14のスニペットは, 新規の情報取得に役立つ単語の個数が5個であるため, 表5.13の役立つノード(単語)数が5個以上の部分に分類される。

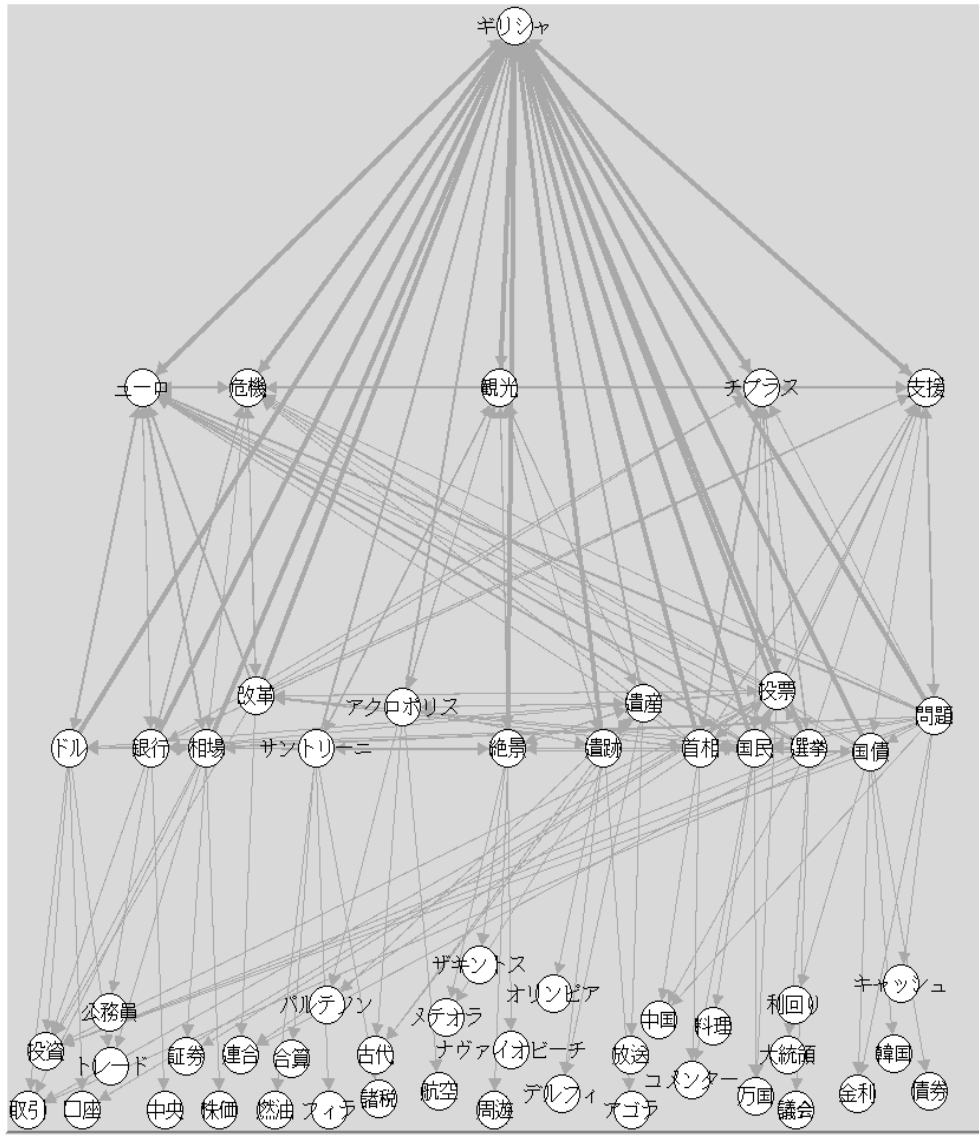
表 5.13: DF の置き換えにおける一定個数以上の新規の情報取得に役立つノード（単語）を持つネットワーク及びスニペットの比率

| 手法 | 役立つノード（単語）数 | | |
|---------|-------------|------------|------------|
| | 1 個以上 | 3 個以上 | 5 個以上 |
| DF=Wiki | 1.00(10/10) | 0.70(7/10) | 0.70(7/10) |
| DF=新聞 | 0.90(9/10) | 0.70(7/10) | 0.70(7/10) |
| IDF=1 | 0.70(7/10) | 0.50(5/10) | 0.40(4/10) |
| スニペット法 | 0.60(6/10) | 0.20(2/10) | 0.10(1/10) |



- 役立つとしたノード：なし

図 5.11: IDF=1 としたテーマキーワード「ギリシャ」のネットワーク



- 役立つとしたノード：サントリーニ，アクロポリス，遺産，パルテノン，メテオラ

図 5.12: DF=Wiki としたテーマキーワード「ギリシャ」のネットワーク

表 5.14: テーマキーワード「ギリシャ」のスニペット

| |
|--|
| <p>ギリシャ共和国 (ギリシャきょうわこく、ギリシャ語: μ) 通称ギリシャは、東南ヨーロッパに位置する国である。2011 年国勢調査によると、ギリシャの人口は約 1,080 万人である。アテネは首都及び最大都市 ...</p> |
| <p>ギリシャへ観光しよう やっぱり人気、ギリシャ旅行 ‘ギリシャ’ と聞かされてどんなイメージが沸くだろうか。筆者としてはやはりギリシャ神話だと考えている、そう思っている人は多いはず。ならばそれ以外で他にないかと ...</p> |
| <p>地球の歩き方がおすすめの観光スポットや旅行情報をご紹介します。ギリシャの天気や通貨、治安などの現地情報に加え、格安航空券、最安値ツアー、ホテル予約まで、ギリシャの旅行・観光情報も満載。ギリシャ旅行に関する ...</p> |
| <p>ギリシャ人 5 言語 現代ギリシャ語 6 宗教 ギリシャ正教 7 国祭日 3月25日 (独立記念日) 8 略史 年月 略史 1821年 オスマン帝国からの独立戦争 1832年 コンスタンティノーブル条約によりギリシャの独立が承認され、ギリシャ王国が成立。 ...</p> |
| <p>ギリシャ旅行・ツアー・観光 ヨーロッパ文明発祥の地、そしてギリシャ神話の舞台として有名な国、ギリシャ。アテネ、サントリーニ島、ミコノス島、クレタ島他、お得なギリシャ旅行情報満載！H.I.S. のおすすめツアーから、こだわり ...</p> |
| <p>役立つとした単語：コンスタンティノーブル条約，オスマン帝国，サントリーニ島，ミコノス島，クレタ島</p> |

第6章 考察

6.1 主観によるネットワークの評価の考察

5.1.2 節の評価方法に基づき評価した結果，新聞データを用いる先行手法より検索エンジンを用いる基本手法のほうが，新規で情報取得に役立つノードが多かった．そのうち，先行手法ではテーマキーワード 20 個のうち約 4 割は新規で情報取得に役立つノードが 5 個以上見られたが，基本手法ではテーマキーワード 20 個のうち約 8 割で新規で情報取得に役立つノードが 5 個以上見られた．また，スニペット法ではテーマキーワード 20 個のうち 1 個しか，情報取得に十分であるネットワークを作成できなかった．しかし，スニペット法に用いるスニペット上位 5 件分は，合計文字数がネットワークのノードの合計文字数の約 2 倍ほどであるため，必ずしも文字数が多いほうが評価方法に有利になるわけではないことがわかる．

基本手法のみが情報取得に十分とされるネットワークを作成できる場合も見られた．例として，図 5.1 より，テーマキーワードを「鳥取」として先行手法で作成したネットワークには，テーマキーワードである「鳥取」に関する情報が少なかった．しかし，図 5.2 より，同じくテーマキーワードを「鳥取」として基本手法で作成したネットワークには，テーマキーワードである「鳥取」に関する「鳥取砂丘」「大山」といったノードが見られた．これは，先行手法で用いている新聞データに，テーマキーワードに関する情報が少ないためであると考えられる．

しかし，基本手法で作成したネットワークが，先行手法で作成したネットワークよりも新規で情報取得に役立つノードが少ない場合もあった．図 5.4 のテーマキーワードを「石油」として基本手法で作成したネットワークには，新規で情報取得に役立つノードが「シェールオイル」のみであったが，図 5.3 のテーマキーワードを「石油」として先行手法で作成したネットワークは「シェールオイル」のノードを含み，新規で情報取得に役立つノードが 5 個以上であった．これは，収集した検索結果 50 件のデータが新規で情報取得をするのに不足していたためであると考えられる．

6.2 複数被験者によるネットワークの評価の考察

主観で評価した場合と比較して，被験者実験では情報取得に役立つノードに選ばれた個数が少なかった．これは，テーマキーワードに関係のない情報であるノードがネットワークに多く，リンクから読み取れる情報が不足していたことが影響していると考えられる．また，選ぶノードの個数は被験者の知識の量によって，ある程度変化する．

被験者実験の結果，検索エンジンを用いる基本手法のほうが新聞データを用いる先行手法より，幅広いネットワークで情報取得に十分であるネットワークを作成できることが確認できた．また，Web サイトの要約文であるスニペットと比較しても，基本手法で作成したネットワークが新規の情報取得に適していた．

6.3 アンドテーマ手法の評価の考察

アンドテーマ手法でネットワークを作成した結果，テーマキーワードのみを用いる基本手法で作成したネットワークでは見られないノードが多く見られる場合があった．例として，テーマキーワードを「世界遺産」，アンドテーマキーワードを「遺跡」としてアンドテーマ手法で作成した図 5.8 には，テーマキーワードのみを用いて基本手法で作成した図 5.7 には見られない「ボロブドゥール」「プレアヴィヒア」といった「遺跡」に関するノードが見られた．なお「ボロブドゥール」は，インドネシアのジャワ島中部のケドゥ盆地に所在する大規模な仏教遺跡で世界的な石造遺跡であるボロブドゥール遺跡のことを示し「プレアヴィヒア」は，カンボジアとタイ王国の国境にあるダンレク山地内のプレアヴィヒア州に位置する遺跡のプレアヴィヒア寺院のことを示す．また，アンドテーマ手法ではアンドテーマキーワードに関するノードが基本手法で出現した位置よりも上位に見られたため，情報を取得できる範囲が拡大された．

アンドテーマ手法でネットワークを作成したが，アンドテーマキーワードに関する新規で情報取得に役立つノードを得られなかった場合もあった．テーマキーワードを「宇宙」，アンドテーマキーワードを「銀河」としてアンドテーマ手法で作成した図 5.10 のネットワークは，テーマキーワードを「宇宙」として基本手法で作成した図 5.9 のネットワークでは見られない「マゼラン」といったノードが見られたが，アンドテーマキーワードと関連性が低いとされるノードも多く見られた．これは，設定したアンドテーマキーワードに関する追加の情報が少なかったため，新規で情報取得が可能なノードを引き出すことができなかったためであると考えられる．

ネットワークの内容が絞られ、興味がある内容を得やすくなる場合も見られた。テーマキーワードを「京都」、アンドテーマキーワードを「寺」としてアンドテーマ手法で作成した図 5.6 は、テーマキーワードのみを用いて基本手法で作成した図 5.7 よりもノードが少なくなり、「醍醐」「銀閣」といった「寺」に関するノードが上位に見られ、アンドテーマキーワードである「寺」に沿った内容となった。

6.4 DF の置き換えによる変化の考察

テーマキーワードを「ギリシャ」として、 $IDF=1$ とした図 5.11 の場合、どの置き換えにおいても共通して見られるような一般的なノードが多くなり、新規で情報取得に役立つノードが少なくなった。テーマキーワードを「ギリシャ」として、DF の計算に Wiki データを用いた図 5.12 では、テーマキーワードである「ギリシャ」の地名、名所や遺跡等もノードに出現し、新規で情報取得に役立つノードやリンクを見つけやすいネットワークとなった。また、使用した 10 個のうち 3 個が、DF の計算に Wiki データや新聞データを用いた場合においても情報取得に役立つノードが 3 個以上見られなかったが、これはテーマキーワードについて周知の情報が多く、新規の情報を引き出しづらかったためであると考えられる。

6.5 今後の課題

作成したネットワークのノードが多すぎることで、重要な情報が把握しづらい場合も見られた。重要な情報が把握しづらい場合、短時間で内容を把握することが困難になり、情報の見落としにもつながる。そのため、今後は、得られた情報を使用者に理解しやすい状態で提示することで、さらに役立てることができると考えている。

第7章 おわりに

本研究では、検索エンジンを用いて単語ネットワークの作成を行った。ネットワークの主となる概念をテーマキーワードとして設定し、テーマキーワードに関する情報を検索エンジンで収集した。そして、収集したテーマキーワードに関する情報をもとにネットワークを作成した。これにより、幅広いテーマキーワードで情報量が十分なネットワークを作成できた。

実際に20個のテーマキーワードでネットワークを作成し、新規で内容把握や発想支援に役立つとされるノードに着目して、評価を行った。評価の結果、作成したネットワークのうち、約8割のネットワークが情報取得に効果的であることが確認できた。また、新聞を用いた場合に十分なネットワークが作成できなかった8個のテーマキーワードのうち、6個のテーマキーワードは検索エンジンを用いた提案手法で情報取得に十分とされるネットワークが作成できた。さらに、検索エンジンを用いた提案手法では、新聞を用いた場合には見られなかったノードが多く見られた。

複数名の被験者に基づく実験も実施し、検索エンジンを用いる基本手法のほうが新聞データを用いる先行手法より、幅広いネットワークで情報取得に十分であるネットワークを作成できることが確認できた。また、Webサイトの要約文であるスニペットと比較しても、基本手法で作成したネットワークが新規の情報取得に適していた。

アンドテーマ手法では、テーマキーワードのみで作成されたネットワークを確認し、情報をより引き出すことができるアンドテーマキーワードを設定した。これにより、アンドテーマキーワードに沿った追加の情報を取得することができた。

しかし、作成したネットワークのノードが多すぎることで、重要な情報が把握しづらい場合も見られた。今後は、得られた情報を使用者に理解しやすい状態で提示することで、さらに役立つことができると考えている。

謝辞

最後に，1年間研究を進めるに当たって，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へ深く感謝するとともに心から御礼申し上げます．また，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.
- [2] 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. 言語処理学会第16回年次大会発表論文集, pp. 119–133, 2016.
- [3] 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.
- [4] 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.
- [5] 松尾豊, 友部博教, 橋田浩一, 石塚満. Webから人間関係ネットワークの抽出と情報支援. 人口知能学会第17回全国大会講演論文, pp. 1–4, 2003.
- [6] 友部博教, 松尾豊, 武田英明, 安田雪, 橋田浩一, 石塚満. Semantic webのための人の社会ネットワーク抽出と利用. 情報処理学会論文誌, Vol. 46, No. 6, 2005.
- [7] M. Harada, S. Sato, and K. Kazama. Finding authoritative people from the web . in proceeding of the joint conference on digital libraries. pp. 306–313, 2004.
- [8] 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.
- [9] 大澤幸生, ネルスE. ベンソン, 谷内田正彦. Keygraph : 語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会論文誌, Vol. J82-D-I, No. 2, pp. 391–400, 1999.
- [10] Bing search api. <https://datamarket.azure.com/dataset/bing/search>.