人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第12回)
SIG-AM-12-08

社会情勢の変化を表す表現の自動収集と可視化の検討

Automatic Collection and Visualization of the Expressions that Represent Changes of Situations in the Society

尾亦　智弘[1*]　渋木　英潔[2]　森　辰則[2]
Tomohiro Omata[1]　Hideyuki Shibuki[2]　Tatsunori Mori[2]

[1] 横浜国立大学大学院環境情報学府
[1] Graduate School of Environment and Information Sciences, Yokohama National University
[2] 横浜国立大学大学院環境情報研究院
[2] Graduate School of Environment and Information Sciences, Yokohama National University

Abstract: In this paper, we discuss automatic collection and visualization of the expressions that represent changes of situations in our society. In creating new products or making a new plan, it is necessary to grasp the change in social situation. There are various ways in which we may find changes of situation in our society. For some of them, such as statistical manners, we have to begin with discussion of how to interpret them. We focus on expression on Web documents that seem to be written as authors' own interpretation of chages of situations in our society. We propose a method to collect them automatically from the Web documents. We also discuss a visualization method to help users to understand the collected expressions easily.

1　はじめに

ある技術を活用し新しいものを作る時や，企業などが新企画を立てる際には，それがどれだけ社会のニーズに答えられるかが重要となる．そのため，社会がどのような状況にあるのか，どのような変化が起きているのかといった社会情勢の変化を把握することが必要である．しかし，社会情勢の変化が現れる対象は多岐にわたる上に，統計情報のように解釈の仕方から議論しなければならないような情報もある．そのため，社会情勢の調査を支援するような技術が求められている．

しかし，そのような状況にありながら，現在，社会情勢の変化を収集するような手法は確立されていない．そのため，ニュースや新聞を見るといった手作業での調査が行われている．そこで本稿では，社会情勢の調査時の負担軽減を目的とし，Web 上の文書群から，社会情勢の変化を表すものであると書き手が解釈しているであろう表現を自動的に収集する手法を検討する．また，収集した社会情勢の変化を表す表現を利用者に見やすく提示するために可視化についての検討も行う．ここで，社会情勢の変化を表す表現とは,「少子化」や「女性の社会進出」など，我々の社会の状態の変化を指し示す名詞や名詞句である．

社会情勢の変化を表す表現を含む文にはどのようなものがあるか調査を行ったところ,「少子化といった時代の変化 … 」や「少子化などの世の中の流れ … 」といった文が見られた．この文では,「といった」や「などの」といった，例示の表現によって,「時代の変化」や「世の中の流れ」といった，社会情勢の変化の具体事例が属するであろう上位クラスに対して，その実例として「少子化」という表現により指し示されるコトが存在すること示されている．つまり，この文は「＜社会情勢の変化の実例を指し示す表現＞＜例示の表現＞＜社会情勢の変化の具体事例が属する上位クラスを指し示す表現＞」という構成をしている．このことから，社会情勢の変化の具体事例が属する上位クラスを指し示す表現から，社会情勢の変化の実例を指し示す表現を抽出するという，上位下位関係の抽出によって，社会情勢の変化を表す表現が収集できるのではないかと考えた．また，上位クラスを指し示す表現の種類を増やすことで，より多くの社会情勢の変化を表す表現が獲得できると考えられる．そのため，ブートストラッピング法により，上位クラスを指し示す表現を拡充させることを考える．本稿では，これら考えに基づき，社会情勢の変化を表す表現の収集手法を検討する．

*連絡先： 横浜国立大学大学院環境情報学府
〒 240-8501 神奈川県横浜市保土ケ谷区常盤台 79-1
E-mail: omata_t@forest.eis.ynu.ac.jp

2 関連研究

本節では提案手法に用いた要素技術について述べる．

2.1 定型表現を用いた上位下位関係抽出手法

大量のテキストデータから用語間の関係を判別する手法はこれまでに数多く提案されている．X を上位語，Y を下位語としたときに,「Y などの X」といった定型表現を用いる手法が有力な手法として存在している [1, 2]．上記の定型表現を用いることで，X と Y は上位下位関係にあるということが判別できる．このほかにも，安藤ら [3] は,「Y といった X」「Y のような X」などのパターンも上位下位関係を判定するために有力な定型表現であることを分析している．

本研究では,「Y などの X」「Y といった X」「Y のような X」の 3 つ定型表現を抽出に用いる．

2.2 ブートストラッピング法

ブートストラッピング法 [4, 5] とは，獲得対象となるクラスのインスタンスをシードとして与え，コーパスからインスタンスと共起するパターンを抽出し，抽出した共起パターンを用いて新たなインスタンスを抽出する．といった手順を反復的に繰り返すことで，少数のインスタンスから大規模なインスタンス集合を再帰的に獲得する手法である．ブートストラッピング法は，語義曖昧性解消，固有表現抽出および関係抽出など自然言語処理の様々なタスクに利用されている．本研究では，ブートストラッピング法を関係抽出に利用する．

ブートストラッピング法では，反復処理を繰り返していくうちにシードインスタンスとは関係のないインスタンスを抽出してしまう「意味ドリフト」という問題がある [6]．これは，ブートストラッピング法の反復過程において再現率は高いが適合率が低いパターン (ジェネリックパターン) を抽出してしまうことに起因する現象である．

意味ドリフトを回避する方法としては，ジェネリックパターンを抽出する前に反復を打ち切ることがあるが，反復を停止する最適な回数はタスクにより異なり，事前に決定することは困難である．Esspresso アルゴリズム [7] は精巧なスコアリング関数を用いて相互再帰的にインスタンスとパターンのスコアを定義し，意味ドリフトによる問題を軽減している．

2.3 Espresso アルゴリズム

Espresso アルゴリズムのスコアリング関数は，信頼度の高いパターンと頻繁に共起するインスタンスは信頼度が高く，信頼度の高いインスタンスと頻繁に共起するパターンは信頼度が高いという考えに基づいている．パターン p とインスタンス i のスコアはそれぞれ $r_\pi(p)$ と $r_l(i)$ であり，以下の式を用いて信頼度を計算する．

$$r_\pi(p) = \frac{1}{|I|}\sum_{i\in I}\frac{pmi(i,p)}{maxpmi}r_l(i) \quad (1)$$

$$r_l(i) = \frac{1}{|P|}\sum_{p\in P}\frac{pmi(i,p)}{maxpmi}r_\pi(p) \quad (2)$$

$$pmi(i,p) = \log\frac{|i,p|}{|i,*||*,p|} \quad (3)$$

P はパターン集合，I はインスタンス集合であり，$|P|$ と $|I|$ はパターンとインスタンスの数を表す．$|i,*|$ はインスタンス i の頻度，$|*,p|$ はパターン p の頻度を表す．$|i,p|$ はインスタンス i とパターン p が共起する回数である．pmi はインスタンスとパターン間の自己相互情報量を表しており，$maxpmi$ は全てのインスタンスとパターンの組み合わせの間における pmi の最大値である．なお，$r_\pi(p)$ と $r_l(i)$ の初期値はそれぞれ 1 である．

Espresso アルゴリズムでは，反復過程において (1) 式と (2) 式を適用することで，精度を高く保ちながら再現率を大幅に向上させている．

2.4 自己組織化マップ（SOM）

Kohonen によって提案されたニューラルネットワークモデルのひとつであり，トポロジカルマッピングによる教師なし学習を行うことで，多次元の属性値から 2 次元マップを生成し，視覚化することで，効果的なデータ分類手法として注目されている．自己組織化マップ上では類似度の高いデータどうしは近くに，類似度の低いデータどうしは遠くに配置される．[8]

3 社会情勢の変化を表す表現

本節では，本研究で抽出対象とする，社会情勢の変化を表す表現について述べ，それが現れる文章にはどのようなものがあるか調査する．そして，どのようにしたらそれを収集できるのかについて考察する．

3.1 社会情勢の変化を表す表現

社会情勢の変化は，注目しているある時点での社会の状態と，それ以前の社会の状態の間に違いが見られるものであり，その上で，その状態が変わっていく方向が見られるものが社会情勢の変化を表す表現だと言える．例としては,「少子化」や「女性の社会進出」などのものである.「少子化」では，注目している時点において，子供の数が少なく，それ以前の時点ではそれよりも数が多かったことが分かり，それから,「化」という文字によって，多い状態から少ない状態への推移が分かる．なお，どのような変化を表す表現を抽出すべきかは利用者のニーズによる．そのニーズとして，今の社会情勢の変化が知りたいということが多いため，過去の内容を指し示す表現は対象にしない．

3.2 社会情勢の変化を表す表現が現れる文章

次に，社会情勢の変化を表す表現であることが明示されている文章について調査した．その結果を「少子化」を例に，以下に記す．

① 上位クラスの例示として示されているもの

> 若者の都市部集中、車離れ、晩婚化、**少子化**といった時代の流れとは別の経済活動や行動様式を持っている。

② 直後の表現によって変化であることが示されるもの

> 急速な**少子化**の進行とそれに伴う人口減少は、社会経済全般にわたり、さまざまな影響を及ぼすことが想定されます。

①では,「時代の流れ」という，社会情勢の変化の具体事例が属する上位クラスを指し示すもの表現があり，その中の例示として「少子化」が示されている．②では,「進行」というある物事が進んでいることを表す表現によって,「少子化」が変化であることが示されている．これらのうち①が抽出に有用な表現だと考えられる．その理由としては,「といった」という例示であることを表す表現によって「＜社会情勢の変化の実例を指し示す表現＞＜といった＞＜社会情勢の変化の具体事例が属する上位クラスを指し示す表現＞」という構造をしていることが挙げられる．②は，直後の表現の膨大さや，文の意味を解釈する必要があるといったことが考えられる．そのため，本研究では①のように，上位クラスを指し示す表現の例示として示されるものを対象に収集を行う．

3.3 社会情勢の変化を表す表現の収集に向けたアプローチ

前小節にて，上位クラスの例示として示されるものを対象に収集を行うことを述べた．本小節では，これについてさらに調査し，どのようにして社会情勢の変化を表す表現の収集に取り組むのかというアプローチについて述べる．前小節の①の例に出てきた，例示を表す「といった」と，社会情勢の変化の具体事例が属する上位クラスを指し示す表現の「時代の流れ」の 2 つを組み見合わせた「といった時代の流れ」という表現で検索を行ったところ，以下のような①とは違う文章が得られた．

> グローバル化、市場の変化、販売方法の多様化**といった時代の流れ**、スピードは今後益々想像を絶する勢いで推移することでしょう。

この文章からは,「グローバル化」「市場の変化」「販売方法の多様化」が社会情勢の変化を表す表現として得られる．このため,「＜例示の表現＞＜社会情勢の変化の具体事例が属する上位クラスを指し示す表現＞」という定型表現を検索質問として検索することで，定型表現の前に現れる，社会情勢の変化を表す表現が得られるのではないかと考えられる．

例示の表現は「といった」のほかにも「などの」「のような」といったものが考えられこれらについても,「時代の流れ」と組み合わせた表現を用いて検索を行い，以下の文章を得た．

> 少子・高齢化や高度情報化、グローバル化**などの時代の流れ**により、個人の価値観の多様化や行動の変化が進むほか、雇用システムの変化、社会保障制度の見直し、環境問題への取り組み、男女共同参画の推進など、社会経済システムの改革が迫られている。

> 文明は、生活環境の大幅な変化、或は産業革命の**ような時代の流れ**があった時に使われているように思います。

これらの例からも，社会情勢の変化を表す表現を得ることができる．

また，例示の表現の場合と同様に，社会情勢の変化の具体事例が属する上位クラスを指し示す表現も「時代の変化」「世の中の流れ」のように,「時代の流れ」以外の表現が考えられる．これについて,「といった」と組み合わせた表現を用いて検索を行い，以下の文章を得た．

社会が急激に変化する中、我が国の教育も、知識基盤社会、グローバル化、人口減少社会といった**時代の変化**に即した対応が求められており、教育を支える教員についても新たな時代にふさわしい資質能力を備える必要がある。

我々は皆様のよき相談相手となれるよう、日々研鑽するとともに、コンプライアンスや内部統制といった、**世の中の流れ**に合った支援をさせて頂きます。

これらの例から，これまでとは異なる社会情勢の変化を表す表現が獲得できていることが分かる．

以上の調査から,「＜例示の表現＞＜社会情勢の変化の具体事例が属する上位クラスを指し示す表現＞」という定型表現を用いた社会情勢の変化を表す表現の抽出は有効であると考えられる．また，例示の表現と社会情勢の変化の具体事例が属する上位クラスを指し示す表現の組み合わせを変えて定型表現の種類を増やすことで，様々な社会情勢の変化を表す表現が抽出できると考えられる．これらの考えに基づき，再帰的処理によって少数のシードから，多くのインスタンスを獲得するブートストラッピング法を用いることを考える．ブートストラッピング法では，まず，抽出したい表現 X が属すクラス C を考えたとき,「X といった C」のような表現に注目し，C を固定しつつ，X に対応する表現を集める．その後，C の種類を増やすために，新たに得られた表現 X' を用いて,「X' といった C'」のような表現を集め，C' の候補を増やす，ということを反復的に繰り返すことによって，少数のシードから徐々に，X ならびに C に対応する表現を拡充させている．

4 定型表現を用いたブートストラッピング法による社会情勢の変化を表す表現の収集

本節では，前節にて述べたアプローチに基づく提案手法について述べる．

4.1 概要

本研究では，以下のステップにより社会情勢の変化を表す表現の収集を行う．

① Espresso アルゴリズムによる社会情勢の変化を表す表現の獲得

② Espresso アルゴリズムによる社会情勢の変化を表す上位クラス表現の候補の獲得

③ 人手による②で得られた表現のフィルタリング

④ ①②③を繰り返す

図 1 に上記ステップの流れを概略図として示す．上記ステップの詳細について次節で述べていく．

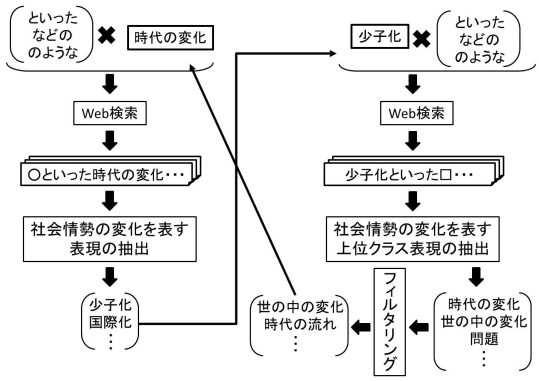

図 1: プロセス概要

4.2 Espresso アルゴリズムによる社会情勢の変化を表す表現の獲得

このステップでは，入力として社会情勢の変化の具体事例が属する上位クラスを指し示す表現を与え，例示の表現を組み合わせて定型表現を作成する．作成した定型表現をクエリに Web 検索を行い，定型表現を含む文章を得る．本研究では，例示の表現として「といった」「などの」「のような」の 3 つの表現を用いた．つまり，クエリは「といった○○」「などの○○」「のような○○」(○○は入力として与えた表現) となる．また，Web 検索には Bing[1] を用いた．

得られた文章から検索に用いた定型表現を含む 1 文を抜き出し，KNP により係り受け解析を行う．KNP による解析の結果，定型表現にかかっていると解析されたものを社会情勢の変化を表す表現として抽出する．その際,「少子・高齢化や高度情報化、グローバル化などの時代の流れにより...」の例のように，社会情勢の変化を表す表現が並列していることがある．そのため，句点「、」，助詞「や」，助詞「と」で区切って抽出をする．上記の例では,「少子・高齢化」「高度情報化」「グローバル化」に分けられて抽出される．

その後，Espresso アルゴリズムのスコアリング関数を用いて，得られた表現の信頼度を計算し，その上位 3

[1] http://www.bing.com/?cc=jp

件を次のステップである，社会情勢の変化の具体事例が属する上位クラスを指し示す表現を抽出する際の入力とする．実験的に抽出を行ったところ,「現在」「私」「今」などのような社会情勢の変化を表す表現ではなく，高頻度で現れるため Espresso アルゴリズムの信頼度が高くなってしまう語があった．そのため，これらの語を排除するため，2 文字以下の語は信頼度を計算せず，入力にならないようにした．

4.3 Espresso アルゴリズムによる社会情勢の変化の具体事例が属する上位クラス表現の候補の獲得

このステップでは，入力として社会情勢の変化を表す表現を与え，例示の表現と組み合わせて定型表現を作成する．前のステップと同様に，例示の表現としては「といった」「などの」「のような」の 3 つの表現を用いる．そのため，クエリは「△△といった」「△△などの」「△△のような」(△△は入力として与えた表現) となる．その後，作成した定型表現をクエリに Web 検索を行い，定型表現を含む文章を得る．

得られた文章から検索に用いた定型表現を含む 1 文を抜き出し，JUMAN により形態素解析を行う．解析の結果，定型表現の後に続く名詞句を社会情勢の変化の具体事例が属する上位クラス表現として抽出する．本研究では，名詞，助詞「の」，接尾辞，連体詞，形容詞のいずれかが続く限り名詞句と見なしている．

その後，Espresso アルゴリズムのスコアリング関数を用いて，得られた表現の信頼度を計算し，表現の信頼度の順位付けを行う．実験的に抽出を行ったところ,「問題」「課題」「私たち」のような意味ドリフトを起こしやすい表現が Espresso アルゴリズムでの信頼度が高くなっていた．そのため，3 文字以下の語は入力を計算せず，入力にならないようにした．

4.4 人手による表現のフィルタリング

Espresso アルゴリズムによる社会情勢の変化を表す表現獲得の予備実験を行ったところ，反復を繰り返すにつれ社会情勢の変化とは関係のない表現が多く獲得された．これは，1 度目の反復の時点で,「取り組み」や「キーワード」のような意味ドリフトを起こすインスタンスが選択されるため，後の反復で意味ドリフトが起き，関係のないインスタンスばかりが獲得されるためである．そこで，各反復において社会情勢の変化を表す上位クラス表現に対して，人手でフィルタリングをかけることを考える．獲得されるインスタンスは 300 程度あり，その中から社会情勢の変化を表す表現かどうかを確認するのは大きなコストがかかる．これに対し，人手でのフィルタリングで確認する件数としては，毎回数件を選べばよいため，各回で 10 件もない．そのため，社会情勢の変化を表す表現かどうかを確認するコストよりも小さいといえる．そのため，人手でのフィルタリング処理を加えても，十分に本研究の目的を果たせると考えられる．人手でのフィルタリングにより，関係のないインスタンスを取り除き，意味ドリフトの問題を抑えることができるようになることが期待される．

人手によるフィルタリングは，Espresso アルゴリズムによって順位付けされた表現を上位から見ていき，1 度用いた表現は選択しないという条件で行い，3 件の表現を獲得するまで行うことによって行う．

5 評価実験

4 章で述べた手法により，社会情勢の変化を表す表現がどの程度獲得できるのかを確認するために実験を行った．

5.1 実験方法

加藤らの研究 [9] の実験方法を参考にした．社会情勢の変化の具体事例が属する上位クラス表現のシードとして,「時代の変化」「時代の流れ」「世の中の変化」「世の中の流れ」の 4 つの表現を与え，ブートストラップによる抽出の反復を 3 回行う．そして，社会情勢の変化を表す表現として得られた文字列の集合に対して人手で評価を行った．なお，本実験の Web 検索には Bing を用い，1 回の検索で取得する Web 文書数は 50 件とした．

5.2 実験結果

本研究では Web 文書を対象にしているため，再現率を本質的に求めることはできない．そのため，ブートストラップを 3 回繰り返して得られた候補に対して人手で確認を行い，そのなかで社会情勢の変化を表す表現であると判断されたものを全正解集合と仮定する．そのときの適合率と再現率のグラフを図 2 に，反復回数と，得られた表現の異なり数，正解の異なり数の増加の様子を図 3 に示す．また，抽出された社会情勢の変化を表す表現の正解例とその上位クラス表現の例を，それぞれ表 1 と表 2 に示す．なお，図 2 における，0 回目とは，入力のシードで抽出されたものを指す．

[2]企業の社会的責任 (corporate social responsibility) 企業が利益を追求するだけでなく、組織活動が社会へ与える影響に責任をもち、あらゆるステークホルダー（利害関係者 : 消費者、投資家等、及び社会全体）からの要求に対して適切な意思決定をすることを指す。

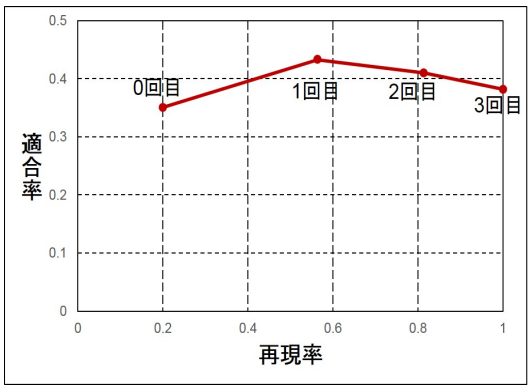

図 2: 収集した社会情勢の変化を表す表現の適合率-再現率グラフ

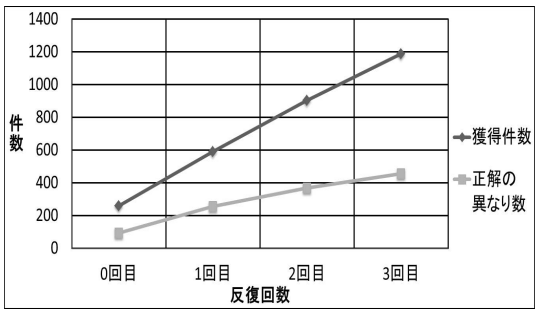

図 3: 獲得件数と正解の異なり数

表 1: 抽出された社会情勢の変化を表す表現の正解例

社会情勢の変化を表す表現	
少子化	高齢化
グローバル化	情報化
女性の社会進出	国際化
人口減少	景気の低迷
経済成長	核家族化
少子・高齢化	技術の進展
CSR[2] の広がり	コモディティ化
産業構造の変化	共働き世帯の増加

表 2: 抽出された社会情勢の変化を表す上位クラス表現の例

社会情勢の変化の具体事例が属する上位クラス表現
環境変化
社会情勢の変化
社会構造の変化
環境の変化
社会の変化
社会環境の変化

6 考察

図 2，3 から，1 回目の反復で再現率を大幅に上昇させ，その後も多くの正解を得ていることが分かる．また，適合率が反復を重ねても大きく低下するようなことは起こっていない．これは，人手でのフィルタリングが効果的に働いたためと考えられる．

しかし，適合率が 40 %程度とやや低調な値であった．誤抽出された表現に対して分析をしたところ，原因は以下の 4 つに大別された．

① 明らかに現在の社会情勢を表していないもの

- 1917 年のロシア革命によるソヴィエト政権の成立
- 大日本帝国の滅亡
- 武士の台頭

② 比喩の表現として用いられていたもの

- 今の濁流
  もとの文章：**今の濁流**のような世の中の流れに取り残されようとしてる日本と私（極めて私的な私）がいます。
- この怒涛
  もとの文章：**この怒涛**のような時代の変化の渦にのみ込まれて沈まないように、本年は、時代の流れを見据えた運営計画を立案し着実に実行していきたいと思います。

③ 係り受けの誤りによるもの

- 最短でその日のうちに自宅まで届けるサービスで
  もとの文章：ネットスーパーは、インターネット上で注文した生鮮品や食料品・日用品を、**最短でその日のうちに自宅まで届けるサービスで**、働く女性や高齢化世帯の増加、インターネットの普及といった社会環境の変化などで、利用者が急速に拡大している。

④ 並列して述べられていたものが区切ったことにより社会情勢の変化であることが分からなくなってしまったもの

- 個人の価値観
  もとの文章：当社では、賃貸住宅事業を住宅事業のもうひとつの柱とすべく、「三井の賃貸」ブランドのもと、**個人の価値観**やライフスタイル、家族構成の多様化といった社会情勢の変化を踏まえた、都心部における良質な賃貸住宅供給を目指した“パーク

アクシス”シリーズを中心に事業展開しております。

①について，抽出された表現のみで現在の社会情勢の変化を表しているかを判断することは難しいと考えられる．そのため，もとの文章の周りの文脈を見て，話題を判断し，それが過去のものだと判断できる場合は，抽出対象にしないようにするというような処理の改善が必要だと考えられる．

②について，これらの表現は「のような」を用いた定型表現で抽出した際のものであった．例示の表現として使用されていることを期待していたが，比喩の表現として使われていたものがほとんどであった．このことから,「のような」は本研究においてはあまり効果的な表現でなかった可能性があり，今後の十分な検討が必要になる．

③について，まず上記の例の係り受けの結果を以下の図 4 に示す．この例では,「最短でその日のうちに自

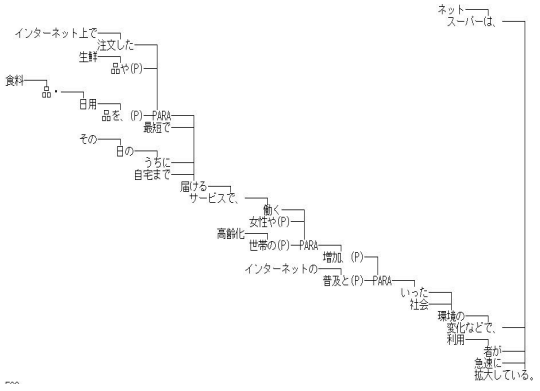

図 4: KNP での解析結果

宅まで届けるサービスで」という表現が「といった社会環境の変化」という表現にかかっていると誤解析されているため，誤って抽出された．これについては，構文解析の以外の所での解決が必要だと思われる．

④について，上記の例は,「個人の価値観の多様化」「ライフスタイルの多様化」「家族構成の多様化」の 3 つが並列して表記されたものである．これについては，KNP での並列構造の解析結果を上手に利用することでの解決が可能と考えられる．

7 SOM による社会情勢の変化を表す表現の可視化

前節まででは，社会情勢の変化を表す表現の自動収集手法について述べたが，図 3 を見ると非常に多くの表現が収集されていることが分かる．また，この表現の中には,「少子高齢化」,「少子・高齢化」,「少子化」,「高齢化」のように同じような意味の表現も含まれている．この状態で収集した表現の全てを提示するのでは，利用者に負担がかかってしまうと考えられる．そのため，近い意味の表現がまとまるように収集した表現を SOM により可視化し，利用者に見やすく提示する．

SOM で収集した表現を可視化するためには，その表現を対応する高次元データで表す必要がある．本研究では，分布仮説 [10] に基づき，文脈ベクトルを表現に対応する高次元データとする．

7.1 文脈ベクトルの作成

文脈ベクトルの作成方法について述べる．対象の表現が収集されたときの文脈だけでは，その表現の文脈としては偏っている可能性がある．そのため，対象の表現を Web で検索し得られ対象の表現を含む一文の集合加えることで，当初とは異なる文脈を集め，その対象の表現が現れうるの文脈を拡張する．また Web 検索時のヒット件数が少なかった表現については文脈ベクトルを作成しない．これはその表現が多様な文脈で使われているかを調べ，長い表現や意味をなさないような表現を対象にしないためである．今回はヒット件数が 10 件に満たないものについては文脈ベクトルを作成しなかった．このようにして拡張した文脈に対して，形態素解析を行い内容語を抽出する．各内容語をベクトルの次元とし，その語の頻度を値としてベクトルを作成する．作成したベクトルはそのベクトルの長さで割ることで正規化する．

7.2 SOM による可視化

収集プロセスの 1 回目で得られた 248 表現について，前小節で述べた方法により文脈ベクトルを作成し，SOM により可視化をする．SOM 作成時の学習回数は 1000 × 1000 回，一辺のノード数は 40 個とした．作成した SOM を図 5 に示す．

マップの右下に,「少子・高齢化」,「少子高齢化」,「高齢化」がまとまっていることや，マップの右上に「国際化」,「グローバル化」,「「グローバル化」」,「社会の国際化」,「経済のグローバル化」がまとまっており，近い意味の表現がまとまっていることが確認できる．

7.3 考察

マップの右下に注目すると,「少子・高齢化」,「少子高齢化」,「高齢化」に加えて「複合化」も近くに配置されており，近い意味でない表現までまとまっているため，あまり望ましくない結果となった．これらにつ

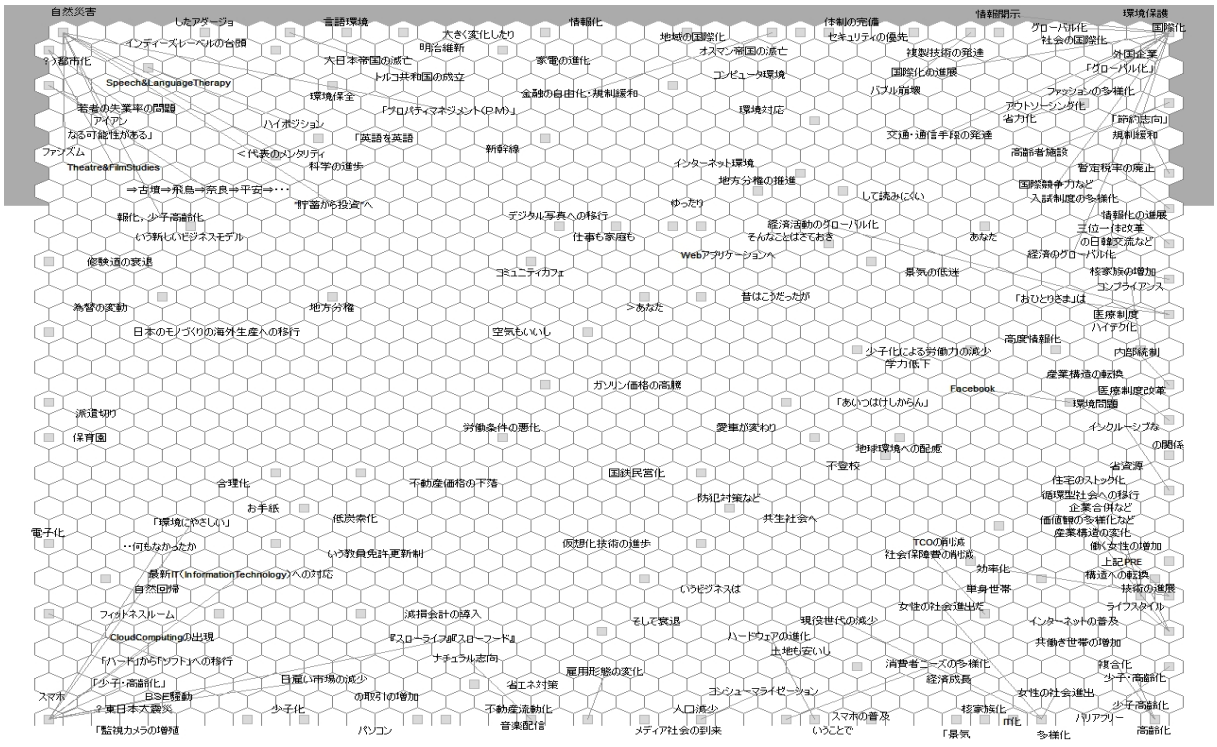

図 5: 作成した自己組織化マップ

いて文脈ベクトルを確認すると,「情報」,「出来る」といった一般的な語が多く共通していることが分かった．これらの一般的な語の影響あると考えられるため，idf などによる語の重み付けや，ストップワードの検討などが必要だと考えられる．

8 まとめ

本研究では，例示の表現に着目し，定型表現とブートストラッピング法を組み合わせて，社会情勢の変化を表す表現を自動収集する手法を提案した．適合率が 40 %程度とやや低調な値ではあったが，多くの社会情勢の変化を表す表現が獲得できた．また，収集した表現を SOM により可視化した．一般的な語の扱いなどの文脈ベクトルの構築方法については今後の検討が必要である．

参考文献

[1] Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the 14th International Conference on Computational Linguistic, pp.539-545,1992.

[2] 安藤まや， 関根聡， 石崎俊 : 定型表現を利用した新聞記事からの下位概念単語の自動抽出，情報処理学会研究報告,2003-NL-157，pp.77-82,2003.

[3] 安藤まや， 関根聡 : 上位語・下位語を含む連体修飾表現の言語的分析，言語処理学会第 10 回年次大会，2004.

[4] Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp.189-196 ,1995.

[5] Abney, S.: Understanding the Yarowsky Algorithm, Computational Linguistics, Vol.30, No.3, pp.365-395,2004.

[6] Curran, J. R., Murphy, T., and Scholz, B.: Minimising semantic drift with Mutual Exclusion Bootstrapping, in Proceedings of the 10th Conference of the Pacific Association for Computational inguistics, pp.172-180,2007.

[7] Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, in Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp.113-120,2006.

[8] T. Kohonen.: The self-organizing map, Proceedings Of The IEEE, Vol.78, No.9, pp1464-1480,1990.

[9] 加藤誠, 大島裕明, 小山聡, 田中克己 : 共起に基づく Web からの類似関係のブートストラップ抽出，日本データベース学会論文誌， Vol.8， No.1， pp.11-16,2009.

[10] 柴田知秀，黒橋禎夫: 超大規模ウェブコーパスを用いた分布類似度計算，言語処理学会第 15 回年次大会発表論文集, pp.705-708, 2009.