

三事象間の因果関係分析のための因果関係ネットワーク構築の一手法

津川 敦朗[†] 新妻 弘崇^{††} 太田 学^{†††}

[†] 岡山大学工学部情報工学科 〒700-8530 岡山県岡山市北区津島中3丁目1番1号
^{††}, ^{†††} 岡山大学大学院自然科学研究科 〒700-8530 岡山県岡山市北区津島中3丁目1番1号
 E-mail: [†], ^{†††}{tsugawa, ohta}@de.cs.okayama-u.ac.jp, ^{††}niitsuma@suri.cs.okayama-u.ac.jp

あらまし 今日, 我々は Web 等のメディアを通じて様々な出来事について情報を入手することができる. しかし, 複数の事象が複雑に絡み合う場合, それを深く理解することは容易ではない. そこで本研究では関連する事象間の因果関係に着目する. 事象間の因果関係を容易に把握できるように, Web 検索を利用して因果関係を抽出し, それらを因果関係ネットワークとして可視化する手法がいくつか提案されている. しかし, これらでは指定した一または二事象についての因果関係しか調べることができなかった. そこで本研究ではこの因果関係ネットワークを拡張し, 三つの事象間の因果関係を分析するための因果関係ネットワーク生成の方法を提案する.

キーワード 因果関係, 可視化, Web マイニング

A Construction Method of a Causal Network for Analyzing Causal Relations among Three Events

Atsuro TSUGAWA[†], Hirotaka NIITSUMA^{††}, and Manabu OHTA^{†††}

[†] Department of Information Technology, Faculty of Engineering, Okayama University
 3-1-1, Tsushima-naka, Kita-ku, Okayama, 700-8530 Japan
^{††}, ^{†††} Graduate School of Natural Science and Technology, Okayama University
 3-1-1, Tsushima-naka, Kita-ku, Okayama, 700-8530 Japan
 E-mail: [†], ^{†††}{tsugawa, ohta}@de.cs.okayama-u.ac.jp, ^{††}niitsuma@suri.cs.okayama-u.ac.jp

Key words causal relation, visualization, Web mining

1 はじめに

新聞やテレビといったメディアを通じて報じられるニュースの中には, 多くの事象が複雑に絡み合って引き起こされる出来事があり, それらを即座に深く理解することは容易ではない. このとき, 関連する事象間の因果関係を把握することは, 内容の理解を深め, リスク回避や意思決定などにも役に立つと考えられる. そこで青野ら [1] は, Web 検索結果から因果関係を自動抽出し, 抽出した要因を始点ノード, 結果を終点ノードとし, それらを結ぶエッジにより因果関係を可視化する因果関係ネットワークを提案した. この因果関係ネットワークによる要因検索システムではユーザが任意の事象を入力すると, 「に伴う」のような因果関係を表す手がかり表現と組み合わせて Web 検索し, その検索結果から要因を抽出する. しかし, このような要因抽出は手がかり表現に依存している上, 青野らは「X に伴う Y」のように, 要因 X が結果 Y の前に出現するような手がかり表現しか使用していなかった. 武田ら [2], [3] は有用な手がかり表

現を新たに追加し, 結果が要因よりも前に出現する因果関係を抽出することを可能とした. また, 青野らや武田らのシステムは, 入力を一事象を表すキーワードとしているため, 藤原ら [4] はシステムを拡張して, 二事象の入力に対してそれらの間に成り立つ因果関係を可視化する方法を提案した. 本稿では青野ら, 武田ら, 藤原らのシステムを拡張して, 三事象間の因果関係を因果関係ネットワークにより可視化する手法を提案する.

本稿の構成は次の通りである. 2 章で関連研究について述べ, 3 章で因果関係の抽出について説明する. 4 章では因果関係ネットワークの構築について, 因果関係を可視化する手法と, 三事象間の因果関係を検索する手法について説明する. 5 章では評価実験として要因結果検索の精度を示すとともに, 実験で生成した因果関係ネットワークについて考察する. 6 章でまとめと今後の課題について述べる.

2 関連研究

本研究で行う因果関係抽出と生成する因果関係ネットワーク,

単語間の類似度尺度についての関連研究について述べる。

2.1 因果関係抽出

文書から因果関係を自動抽出する研究としては、接続関係を用いる手法 [5], [6], [7], [8] と、手がかり表現を用いる手法 [7], [8], [9], [10] が提案されている。接続関係を用いる手法では、複文・重文を解析対象としており、それらを単文に分割したときの各単文の接続関係から因果関係を抽出する。佐藤ら [5] や佐藤・堀田 [8] の研究では、取り出した因果関係の表現形式を格フレームを用いて整理している。一方手がかり表現とは、「に伴う」、「を理由に」、「を原因として」などの要因と結果を結びつける表現のことであり、因果関係を含む文章を抽出する際の判断材料となる。なおこれを、乾ら [7] は「接続標識」、佐藤・堀田 [8] は「手がかり標識」と呼称している。坂地ら [9] や石井ら [10] の研究では、因果関係の持つ構文パターンを用いることによって、複文・重文のみならず、手がかり表現を含んでいる全ての文章から因果関係を抽出している。青野ら [1], 武田ら [2], [3], 藤原ら [4] は坂地ら、石井らの手法を用いてる。本研究でも坂地ら、石井らの手法を用いる。

山田ら [11] はクローズドキャプション中の名詞ペアと名詞ペア間の構文構造に注目し、二つの名詞間に因果関係があるかどうかを EM アルゴリズムを用いて判定している。また Khoo ら [6] は医療関係の文書から因果関係を抽出するために、医療の分野に特化した構文パターンなどを人手で作成している。高見ら [12] はブログにおいて同じ話題に関するバースト特性 (話題の盛り上がり度) がコミュニティ毎に異なる点に着目し、ある話題と因果関係にある事象を発見している。大森ら [13] は不具合事例例文に対して、SVM を利用した因果関係抽出を適用し、製品や部品に生じた不具合の原因に関する記述を取り出す手法を提案している。

2.2 因果関係ネットワーク

抽出した因果関係を可視化する研究としては青野ら [1], 武田ら [2], [3], 藤原ら [4], 佐藤・堀田 [8], 石井ら [10] の研究がある。佐藤・堀田は、因果関係を含む文節から得られる重要単語を、事象ノードを表すキーワードとしている。エッジはノード間、すなわち事象間の因果関係と共起関係を表現しており、それぞれ片方向のエッジ、双方向のエッジで表現される。共起関係の強さは事象ノードのキーワードの概念的な類似度とし、これをノード間の距離の近さに置き換えて表現している。石井らの研究では、日本語文法の SVO 構造に注目し、因果関係を含む文節から抽出した主語、動詞、目的語の組によって事象を表現している。また、事象間でキーワードを、WordNet を用いて概念レベルで比較することで、類似した事象対を発見して結合している。さらに、オンラインニュース記事を一日単位で取得して、新たに抽出された因果関係からネットワークを更新できるようにしている。これにより、抽出した因果関係は日々蓄積していくため、他のノードとほとんどマージされない因果関係を削除することで因果関係ネットワークの可読性を高めている。青野らは、入力キーワードを結果事象に持つ因果関係を抽出し、入力キーワードの要因となっている事象をさらに要因検索することで、階層的に因果関係を獲得し、それらのつながりをネットワークによ

表 1 要因検索で使用する手がかり表現

優先度	手がかり表現
1	に伴う, に伴い, を理由に, が理由で, の影響で
2	により
3	の原因として, の原因は

表 2 結果検索で使用する手がかり表現

優先度	手がかり表現
1	の原因として, の原因は
2	に伴う, に伴い, を理由に, が理由で, の影響で
3	により

り可視化し、因果関係を表現している。武田らはそれを拡張し入力キーワードを要因に持つ因果関係を抽出することを可能とし、藤原らは二事象の入力に対してそれらの間に成り立つ因果関係を可視化している。

本研究では青野らが提案した因果関係ネットワークを基に、三事象の入力キーワードを用いて抽出した因果関係を可視化する手法を提案する。

2.3 語と文書の共起に基づく類似尺度

文書間の類似度を測る指標としてコサイン類似度が一般的に用いられている。また単語の共起の度合いを測る指標として Jaccard 係数, Simpson 係数などを用いる方法がある。Web 文書の場合、単語が同一の Web ページに同時に現れることを共起とし、Web 検索エンジンの検索結果数を用いて共起頻度とする方法が広く用いられている。例えば、二単語の AND 検索の検索結果数によってそれらの単語の共起頻度を見積もることが出来る。多くの場合、検索結果数が極端に少ないものの影響を除くために閾値を定めている。また、加藤ら [14] はブートストラップ法を用いて少数の語のペアを基に、Web 文書から同様の類似関係にある語のペアを大量に取得し、Web を用いて単語間の意味的類似度を測る指標である WebPMI により、取得したペアを評価している。

3 手がかり表現による要因結果検索

本研究では手がかり表現を用いて因果関係が記述されている文書を取得する。本節では武田ら [2], [3] の定めた八つの手がかり表現、「に伴う」、「に伴い」、「を理由に」、「が理由で」、「の影響で」、「により」、「の原因として」、「の原因は」を用いて要因事象と結果事象を取得する方法を述べる。なお、この手がかり表現を用いて要因事象を取得することを以下では「要因検索」と呼び、この手がかり表現を用いて結果事象を取得することを以下では「結果検索」と呼ぶ。

3.1 手がかり表現

本研究では前述した八つの手がかり表現を用いて検索する。武田らの提案している方法を参考に、まずこれらの手がかり表現を因果関係を表す度合いに基づいて三つのグループに分ける。因果関係の検索では、グループ毎に優先度を変えてこれらの手がかり表現を使用する。検索において使用する手がかり表現のグループの優先度、グループの分け方を表 1, 2 に示す。検索に使用する際は同グループ間にも優先度が存在し、リストの初

めに出てくるものほど優先度が高いものとしている。これらは 3.2.1 節及び 3.2.2 節で説明する検索式の生成の優先度に関係する。グループ分け自体は 3.3.3 節で説明する因果関係の重みに関係する。

3.2 検索式と使用順序

因果関係を取得するために、表 1 の手がかり表現に因果関係を調べたい事象を組み合わせた検索式を生成し、Bing Search API [17], [18] を利用して検索結果を取得する。以下ではこの調べたい事象を、検索によって取得した他の事象と区別するために事象 X と呼ぶ。

3.2.1 要因検索

要因検索では入力事象の要因を検索するので、事象 X' を因果関係における結果事象と見なし、以下の八つの検索式を生成する。

$FQ_1 = \text{“ に伴う } X' \text{”}$
 $FQ_2 = \text{“ に伴い } X' \text{”}$
 $FQ_3 = \text{“ を理由に } X' \text{”}$
 $FQ_4 = \text{“ が理由で } X' \text{”}$
 $FQ_5 = \text{“ の影響で } X' \text{”}$
 $FQ_6 = \text{“ により } X' \text{”}$
 $FQ_7 = \text{“ } X' \text{ の原因として ”}$
 $FQ_8 = \text{“ } X' \text{ の原因は ”}$

ただし、この $FQ_1 \dots FQ_8$ の全てについて検索を実行するとは限らない。表 1 の優先度を考慮し、 FQ_1 の検索結果数が十分でない場合、 FQ_2 で検索し、それでも不十分なら FQ_3 というように続け、取得した累計検索結果数が十分な数に達した時点で検索を終了する。

3.2.2 結果検索

結果検索では入力事象の結果を検索するので、事象 X' を因果関係における要因事象と見なし、要因検索の場合と同様に以下の八つの検索式を生成する。

$RQ_1 = \text{“ の原因として } X' \text{”}$
 $RQ_2 = \text{“ の原因は } X' \text{”}$
 $RQ_3 = \text{“ } X' \text{ に伴う ”}$
 $RQ_4 = \text{“ } X' \text{ に伴い ”}$
 $RQ_5 = \text{“ } X' \text{ を理由に ”}$
 $RQ_6 = \text{“ } X' \text{ が理由で ”}$
 $RQ_7 = \text{“ } X' \text{ の影響で ”}$
 $RQ_8 = \text{“ } X' \text{ により ”}$

ただし、この $RQ_1 \dots RQ_8$ の全てについて検索を実行するとは限らない。表 2 の優先度を考慮し、 RQ_1 の検索結果数が十分でない場合は、 RQ_2 で検索し、それでも不十分なら RQ_3 というように続け、取得した累計検索結果数が十分な数に達した時点で検索を終了する。

3.2.3 二つの事象を用いた要因検索

二事象を組み合わせた検索式を生成し Web を検索する。これらの検索式は 3.2.1 節で述べた検索式より要因検索に用いられる優先度が高い。生成される検索式は、因果関係を調べる二事象をそれぞれ X' , X'' とすると以下になる。

$FRQ_1 = \text{“ に伴う } A \text{” AND “ } B \text{”}$

$FRQ_2 = \text{“ に伴い } A \text{” AND “ } B \text{”}$

$FRQ_3 = \text{“ を理由に } A \text{” AND “ } B \text{”}$

$FRQ_4 = \text{“ が理由で } A \text{” AND “ } B \text{”}$

$FRQ_5 = \text{“ の影響で } A \text{” AND “ } B \text{”}$

$FRQ_6 = \text{“ により } A \text{” AND “ } B \text{”}$

$FRQ_7 = \text{“ } A \text{ の原因として ” AND “ } B \text{”}$

$FRQ_8 = \text{“ } A \text{ の原因は ” AND “ } B \text{”}$

$(A, B) = (X', X''), (X'', X')$

ただし、この $FRQ_1 \dots FRQ_8$ の全てについて検索を実行するとは限らない。表 1 の優先度を考慮し、 FRQ_1 の検索結果数が十分でない場合は、 FRQ_2 で検索し、それでも不十分なら FRQ_3 というように続け、取得した累計検索結果数が十分な数に達した時点で検索を終了する。 FRQ_8 まで検索をしても十分な要因検索数が取得できていない場合は 3.2.1 節で述べた検索に移行する。

3.2.4 二つの事象を用いた結果検索

二事象を組み合わせた検索式を生成し Web を検索する。これらの検索式は 3.2.2 節で述べた検索式より結果検索に用いられる優先度が高い。生成される検索式は、因果関係を調べる三事象をそれぞれ X' , X'' とすると以下になる。

$FRQ'_1 = \text{“ の原因として } A \text{” AND “ } B \text{”}$

$FRQ'_2 = \text{“ の原因は } A \text{” AND “ } B \text{”}$

$FRQ'_3 = \text{“ } A \text{ に伴う ” AND “ } B \text{”}$

$FRQ'_4 = \text{“ } A \text{ に伴い ” AND “ } B \text{”}$

$FRQ'_5 = \text{“ } A \text{ を理由に ” AND “ } B \text{”}$

$FRQ'_6 = \text{“ } A \text{ が理由で ” AND “ } B \text{”}$

$FRQ'_7 = \text{“ } A \text{ の影響で ” AND “ } B \text{”}$

$FRQ'_8 = \text{“ } A \text{ により ” AND “ } B \text{”}$

$(A, B) = (X', X''), (X'', X')$

ただし、この $FRQ'_1 \dots FRQ'_8$ の全てについて検索を実行するとは限らない。表 2 の優先度を考慮し、 FRQ'_1 の検索結果数が十分でない場合は、 FRQ'_2 で検索し、それでも不十分なら FRQ'_3 というように続け、取得した累計検索結果数が十分な数に達した時点で検索を終了する。 FRQ'_8 まで検索をしても十分な要因検索数が取得できていない場合は 3.2.2 節で述べた検索に移行する。

3.2.5 三つの事象を用いた要因検索と結果検索

三事象を組み合わせた検索式を生成し Web 検索を行う。これらの検索式は 3.2.1 節、3.2.2 節、3.2.3 節、3.2.4 節の検索式よりも要因検索、結果検索共に用いられる優先度が高い。生成される検索式は、因果関係を調べたい三事象をそれぞれ X' , X'' , X''' とすると以下になる。

$Q_1 = \text{“ } A \text{ に伴う } B \text{” AND “ } C \text{”}$

$Q_2 = \text{“ } A \text{ に伴い } B \text{” AND “ } C \text{”}$

$Q_3 = \text{“ } A \text{ を理由に } B \text{” AND “ } C \text{”}$

$Q_4 = \text{“ } A \text{ が理由で } B \text{” AND “ } C \text{”}$

$Q_5 = \text{“ } A \text{ の影響で } B \text{” AND “ } C \text{”}$

$Q_6 = \text{“ } A \text{ により } B \text{” AND “ } C \text{”}$

$Q_7 = \text{“ } A \text{ の原因として } B \text{” AND “ } C \text{”}$

$Q_8 = “A \text{ の原因は } B” \text{ AND } “C”$

$(A, B, C) = (X', X'', X'''), (X', X''', X''), (X'', X', X'''),$
 $(X'', X''', X'), (X''', X', X''), (X''', X'', X')$

この検索式は 3.1 節で述べた優先度を無視し、全て等しく検索され重みづけされる。ただし、この $Q_1 \dots Q_8$ の全てについて検索を実行するとは限らない。 Q_1 の検索結果数が十分でない場合は、 Q_2 で検索し、それでも不十分なら Q_3 というように続け、取得した累計検索結果数が十分な数に達した時点で検索を終了する。 Q_8 まで検索をしても十分な要因検索数が取得できていない場合は 3.2.3 節で述べた検索に移行する。同様に、 Q_8 まで検索をしても十分な結果検索数が取得できていない場合は 3.2.4 節で述べた検索に移行する。

3.3 因果関係抽出

3.3.1 因果関係の出現パターン

本研究で因果関係抽出を試みる文は、要因と結果の出現順によって以下の 2 パターンに分けられる。

パターン 1. 要因が手がかり表現の直前に出現する。

パターン 2. 要因が手がかり表現の直後に出現する。

パターン 1 に基づく因果関係の抽出は、係り受け解析結果を基に手がかり表現の前の文節を要因とし、手がかり表現を含む文節の係り先を結果として抽出する。ただし、そのままでは要因と結果共に文節（基礎文節と呼称する）となり事象としてはふさわしくない。そこでこの文節から 3.3.2 節で述べるルールに基づいて名詞の塊を生成し基礎文節の代わりに要因と結果とする。パターン 2 からの因果関係の抽出は、パターン 1 と逆に、係り受け解析情報を基に手がかり表現の前の文節を結果とし、手がかり表現を含む文節の係り先を要因として抽出する。ただし、そのままでは要因と結果共に基礎文節となり事象としてはふさわしくない。そこでこの文節から 3.3.2 節で述べるルールに基づいて名詞の塊を生成し要因と結果とする。

3.3.2 要因と結果の抽出

因果関係を調べる事象を「放射線」、「作業員」、「汚染」とし、これらを検索キーワードとして「現在、放射線汚染土壌などの除染作業が行われているが、汚染拡大の原因として作業員の熟練度不足が挙げられている」という文を取得したとする。この文は「の原因として」が手がかり表現としてあるのでパターン 2 に該当する。この文を Cabocha [15]、Mecab [16] を用いて係り受け解析及び形態素解析を行うと表 3 のようになる。表 3 の 5 番目の文節が手がかり表現を含む文節であるので、一つ前の 4 番目の文節が要因の基礎文節である。また、5 番目の文節の係り先である 8 番目の文節が結果の基礎文節である。この文を例に要因と結果を抽出する方法を説明する。まず要因及び結果の構成要素を以下のように定義する。

- 名詞（代名詞と非自立の名詞は除く）
- 英数字のみで構成される要素
- 記号「・」

これらの形態素を連結する事により、要因及び結果を表す語を生成する。連結のルールは以下の通りである。

(1) 基礎文節内の名詞を結合

基礎文節内の形態素を基に、頭から連続する名詞を結合する。表

表 3 係り受け解析及び形態素解析の出力結果の例

ID	文節	形態素	品詞	係り先
0	現在、	現在、 、	名詞、副詞可能 記号、読点	3
1	放射線汚染土壌などの	放射線 汚染 土壌 など の	名詞、一般 名詞、一般 名詞、サ変接続、 助詞、副助詞 助詞、連体化	2
2	除染作業が	除 染 作業 が	名詞、サ変接続 名詞、一般 名詞、サ変接続 助詞、格助詞、一般	3
3	行われているが、	行わ れ て いる が 、	動詞、自立 動詞、接尾 助詞、接続助詞 動詞、非自立 助詞、接続助詞 記号、読点	8
4	汚染拡大の	汚染 拡大 の	名詞、サ変接続 名詞、サ変接続 助詞、連体化	5
5	原因として	原因 として	名詞、一般 助詞、格助詞、連語	8
6	作業員の	作業 員 の	名詞、サ変接続 名詞、接尾、一般 助詞、連体化	7
7	熟練度不足が	熟練 度 不足 が	名詞、サ変接続 名詞、接尾、一般 名詞、サ変接続 助詞、格助詞、一般	8
8	挙げられている	挙げ られ て いる	動詞、自立 動詞、接尾 助詞、接続助詞 動詞、非自立	—

3 の 4 番目の文節は、手がかり表現である「原因として」の一つ前の文節であるので要因の基礎文節である。この基礎文節に注目すると「汚染」と「拡大」がそれぞれ名詞なので結合され、「の」は助詞なので結合されず、「汚染拡大」が一つん塊として抽出される。

(2) 基礎文節の最初の品詞が動詞または形容詞である場合
基礎文節の一つ前の文節を確認し、係り先がその基礎文節かを確認する。係っていればその一つ前の文節を新たに基礎文節とし、その最初の品詞が名詞であるならば続く限り結合する。表 3 の 8 番目の文節は手がかり表現を含む文節の係り先であるので結果の基礎文節である。この文節に注目すると、始まりが動詞「挙げる」であり、一つ前の 7 番目の文節の係り先が結果の基礎文節なのでルール (2) に該当する。あらたに 7 番目の文節を結果の基礎文節とし、「熟練」、「度」、「不足」の三つの名詞を結合させ「熟練度不足」が一つの塊として抽出される。

(3) 基礎文節の最初の品詞が名詞である場合
基礎文節の一つ前の文節を確認し、係り先がその基礎文節であ

ることを確認する。係っていればその文節の最後の単語を確認し、「の」または「が」である場合、かかっている文節の名詞を結合したものと、係り先の文節の名詞を結合したものを「の」でさらに結合する。表3の例の8番目の文節に注目すると、ルール(2)より「熟練度不足」が一つの塊として抽出され、その後さらに一つ前の文節に戻り、「作業」、「員」の二つの名詞を連結した「作業員」を「の」で結合した「作業員の熟練度不足」が一つの塊として抽出される。

上記ルールを例文に当てはめると表3の例文からは、「汚染拡大」という要因と、「作業員の熟練度不足」という結果が抽出される。[1], [2], [3], [4] では手がかり表現のすぐ後の文節から要因・結果の塊を生成しようとするため、「汚染拡大の原因として、現場で働いている作業員の熟練度不足が挙げられている」という文から、「現場」という重要度の低い単語を抽出してしまう。しかし本研究ではこのように、より適切な事象を抽出できる。

3.3.3 因果関係の重要度

本研究の要因検索可視化システムでは、抽出した因果関係に重みを付与し、その値が大きいものを一つの事象に対し三つ可視化する。因果関係 c_x の重み $weight(c_x)$ は青野らの方法に基づいて、以下のように定義した。

$$weight(c_x) = \sum_d \frac{cf_d(c_x)}{cf_d(C)} * group$$

ここで、 d は因果関係 c_x が抽出された文書、 $cf_d(c_x)$ は d から抽出された因果関係 c_x の数、 C は d から抽出された因果関係の集合、 $cf_d(C)$ は d から抽出された因果関係の総数である。また $group$ は3.1節で説明した、手がかり表現のグループを表す重みで、優先度1のグループは10000、優先度2のグループは1、優先度3のグループは0.0001とし優先度が高いグループほど重みが大きくなるようにした。

4 因果関係ネットワークの構築

本研究ではWebから抽出した因果関係を因果関係ネットワークとして可視化する。3.3節で説明したように、CaboCha [15], MeCab [16] を用いて検索結果のスニペットの係り受け解析及び形態素解析を行い、因果関係を抽出したのち、抽出された要因と結果のうち関連度の大きいものについては、その要因と結果を用いて繰り返し検索する場合がある。最後にJUNG [19] を用いて、抽出された因果関係を可視化する。本節では因果関係の可視化の方法と繰り返し検索の条件を説明をする。

4.1 因果関係の可視化

因果関係はその重みによってランク付けするので、その上位の因果関係のみを用いて因果関係ネットワークを構築する。青野ら [1] の手法では要因を始点ノード、結果を終点ノードとして配置することで因果関係ネットワークを可視化している。すなわち、因果関係ネットワークは有向グラフで表され、その中の二つのノードとそれを結ぶエッジが一つの因果関係を表している。本研究では三つの入力キーワードのノードがつながっている状態、すなわち因果関係で説明できる状態を最終結果として可視

化する。一度(一段階)の検索で三つの入力キーワードのノードすべてがつながれば検索を終了し可視化する。つながらなければ4.2節で述べる繰り返し検索をすることで因果関係検索を続ける。二段階の繰り返し検索でパスが生成されなければ三段階、四段階と繰り返し検索をする。

4.2 繰り返し検索

繰り返し検索は最初に行う検索(一段階目の検索)と同様に三事象を入力キーワードとする。選ぶ三事象の組み合わせはランダムに選ぶと検索回数が増えコストがかかるので、関連が高いと判断する事象にて繰り返し検索を行う。繰り返し検索を重ねるごとに二段階、三段階と呼び方を変える。このとき事象間の関連度は共起度にて判断する。共起度の指標としてSimpson係数、Jaccard係数を用いる。二つの事象間の共起度を全ての組み合わせで算出し共起度の最も高かった組み合わせに一事象を加えた三事象で検索をする。共起度について4.2.1節で、三事象を選ぶルールについて4.2.2節で説明する。

4.2.1 事象間の共起度

可視化されたノードの事象名を検索エンジンで検索し、検索結果数を取得する。さらに共起度を測るノードの事象名とのAND検索を行い検索結果数を取得する。そして、これらの検索結果数を用いてSimpson係数、Jaccard係数で共起度を測る。事象名 w_1 を含むWebページの検索結果数を $|W_1|$ 、事象名 w_2 を含むWebページの検索結果数を $|W_2|$ 、事象名 w_1 と w_2 を両方含むWebページの検索結果数を $|W_1 \cap W_2|$ とすると、事象名 w_1 と w_2 のSimpson係数、Jaccard係数は以下の式で計算される。

$$Simpson(w_1, w_2) = \frac{|W_1 \cap W_2|}{\min(|W_1|, |W_2|)}$$

$$Jaccard(w_1, w_2) = \frac{|W_1 \cap W_2|}{|W_1| + |W_2| - |W_1 \cap W_2|}$$

4.2.2 繰り返し検索の対象事象の選び方

一段階の因果関係検索の例を図1、図2に示す。どちらも入力A、B、Cとかかれた黄色のノードが因果関係を調べる第一、第二、第三の入力キーワードを表し、要因・結果とかかれた赤色のノードが因果関係検索によって得られた要因・結果を表している。図2に現れている「事象Z」というノードは入力Aの要因事象であり入力Bの結果事象であるため、このノードを介して入力Aと入力Bの因果関係が分かる。しかし、図1、2共に三つの入力キーワードすべてがパスでつながっているわけではないので繰り返し検索をすることとなる。繰り返し検索の対象の選び方は四種類あり、チャンクの数と何段階目の検索かによって異なる。チャンクとは三つの入力キーワードに注目した時のパスがつながっている部分グラフの数とする。図1では入力Aに対して要因一つ結果二つの合計四つのノードからなるチャンク、入力Bに対して要因一つ結果二つの合計四つのノードからなるチャンク、入力Cに対して要因二つ結果一つの合計四つのノードからなるチャンク、合わせて三つのチャンクがあることが分かる。図2では入力Aと入力Bが「事象Z」ノードによってつながっているため、入力A、Bを含む七つのノードからなる

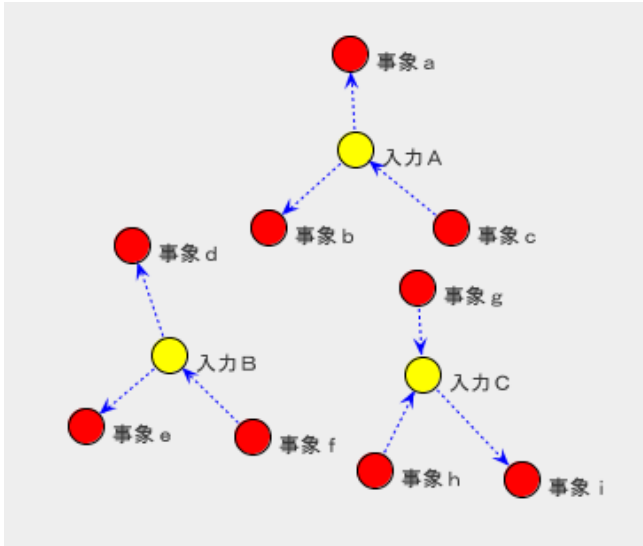


図 1 一段階の因果関係検索の可視化例 (3 チャンク)

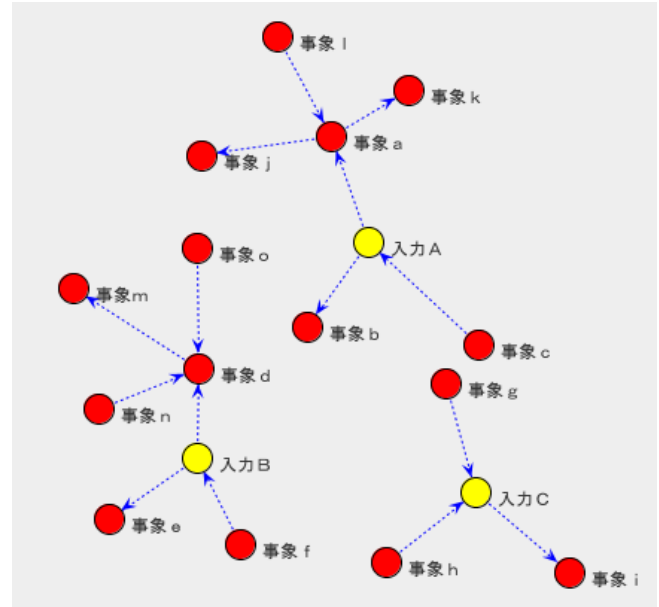


図 3 二段階の因果関係検索の可視化例 (3 チャンク)

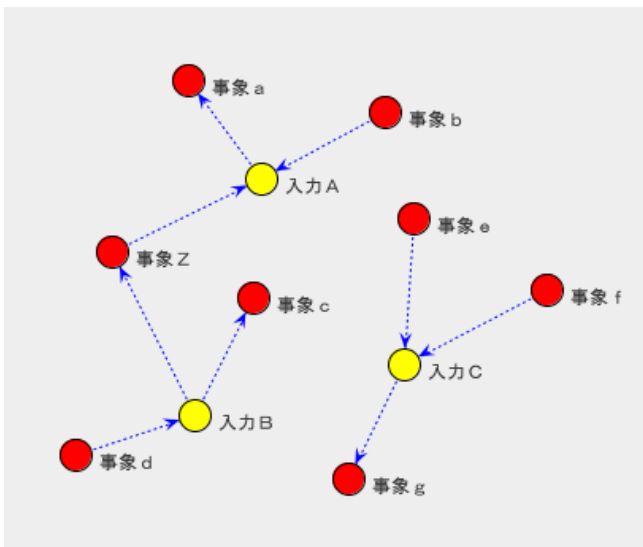


図 2 一段階の因果関係検索の可視化例 (2 チャンク)

チャンクと、入力Cを含む四つのノードからなるチャンク、合わせて二つのチャンクがあることが分かる。以下に繰り返し検索のルールを図を用いて説明する。[t]

(1) 3 チャンクからなる場合 (二段階目)

図 1 を例に説明する。入力 A に注目しパスのつながっている三つのノード事象 a, b, c から一つ、パスのつながっていないノード事象 d, ..., i, 入力 B, 入力 C の八つから一つ選び、合計 24 通りの組み合わせで共起度を測る。その中で最も共起度が高かった組み合わせと入力 A を加えた三事象で新たに因果関係を検索し可視化する。この際に共起度が最大だった組み合わせの二事象からそれぞれ三つずつ取得した因果関係を可視化するが、入力 A からは可読性の低下を考慮し因果関係を増やさない。これらの処理をしたのちにチャンク数を数え、三つのままならば三段階目以降の 3 チャンクでの繰り返し検索をする。入力 A と入力 B もしくは入力 A と入力 C の間のパスが生成されることでチャンク数が二つになった場合は、三段階目以降は、(2) に示す

2 チャンクでの繰り返し検索を行う。

(2) 2 チャンクからなる場合 (二段階目)

図 2 を例に説明する。他の入力キーワードへのパスが生成されていない入力キーワード (独立した入力キーワード) に注目する。図 2 の場合は入力 C が入力 A と入力 B とともパスでつながっていないので、入力 C に注目する。入力 C にパスのつながっている三つのノード事象 e, f, g から一つ、パスのつながっていないノード事象 a, b, c, d, Z, 入力 A, 入力 B の七つから一つ選び、合計 21 通りの組み合わせで共起度を測る。この中で最も共起度が高かった組み合わせと入力 C を加えた三事象で新たに因果関係を検索し可視化させる。この際に共起度が最大だった二事象からそれぞれ三つずつ取得した因果関係を可視化するが、入力 C からは可読性の低下を考慮し因果関係を増やさない。これらの処理をしたのちにチャンク数を数え、二つのままならば三段階目も 2 チャンクで繰り返し検索をすることになる。入力 A, 入力 B, 入力 C の間でパスが生成されることでチャンク数が一つになった場合検索を終了し、最終の可視化結果を表示する。

(3) 3 チャンクからなる場合 (三段階目以降)

図 3 を例に説明する。図 3 は図 1 にルール (1) をあてはめ繰り返し検索をし、事象 a と事象 d の共起度が最も高かった場合の例を示している。ここで共起度が最も高かった組み合わせのうち入力 A とパスでつながっているノードに注目する。図 3 の場合は事象 a である。このノードから新たにパスが生成された事象 j, k, l から一つ、パスのつながっていないノード事象 d, e, f, g, h, i, m, n, o, 入力 B, 入力 C の 11 個から一つ選び、合計 33 通りの組み合わせで共起度を測る。その中で最も共起度が高かった組み合わせに事象 a を加えた三事象で新たに因果関係検索しを可視化する。この際に共起度が最大だった二事象からそれぞれ三つずつ取得した因果関係を可視化するが、事象 a からは可読性の低下を考慮し因果関係を増やさない。これらの処理をしたのちにチャンク数を数え、三つのままならば新たに、共起

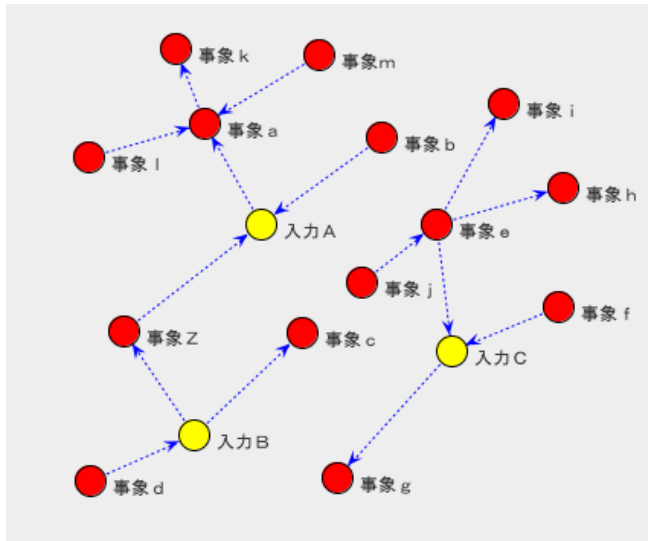


図 4 二段階の因果関係検索の可視化例 (2 チャンク)

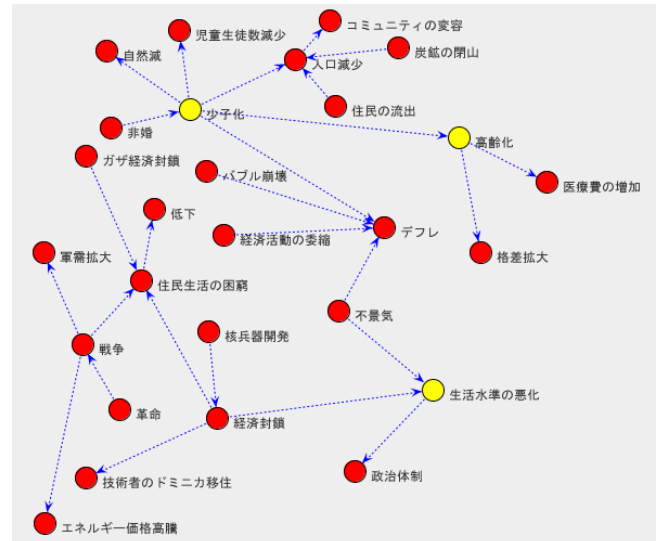


図 5 因果関係ネットワークの可視化例

表 4 3つの入力クエリ

	キーワード 1	キーワード 2	キーワード 3
クエリ 1	水質汚染	排水	酸素
クエリ 2	少子化	高齢化	生活水準の悪化
クエリ 3	物価の上昇	東日本大震災	戦争

表 5 Simpson 係数に基づく因果関係ネットワーク

	クエリ 1	クエリ 2	クエリ 3
検索段階数	3	2	3
全ノード数	21	15	22
最長パスの長さ	5	4	5
最短パスの長さ	2	1	2
適合率 (適切な事象名/ノード数)	0.72 (13/18)	0.92 (11/12)	0.50 (6/12)
適合率 (妥当な因果関係/エッジ数)	0.76 (16/21)	0.93 (13/14)	0.67 (14/21)

度が最も高かった組み合わせのうち入力 A とパスでつながっているノードに注目して、三段階目以降の 3 チャンクでの繰り返し検索をする。

(4) 2 チャンクからなる場合 (三段階目以降)

図 4 を例に説明する。図 4 は図 2 にルール (2) をあてはめ繰り返し検索をし、事象 a と事象 e の共起度が最大だった結果を示している。ここで共起度が最大の組み合わせのうち孤立した入力キーワードとパスがつながっているノードに注目する。図 4 の場合は事象 e である。このノードから新たにパスが生成された事象 h, i, j から一つ、パスのつながっていないノード事象 a, b, c, d, k, l, m, Z, 入力 A, 入力 B の 10 個の中から一つ選び、合計 30 通りの組み合わせで共起度を測る。その中で最も共起度が高かった組み合わせに事象 e を加えた三事象で新たに因果関係を検索し可視化させる。この際に共起度が最大だった二事象からそれぞれ三つずつ取得した因果関係を可視化するが、事象 e からは可読性の低下を考慮し因果関係を可視化させない。これらの処理をした後にチャンク数を数え、二つのままならば新たに、共起度が最も高かった組み合わせのうち入力 C をパスのつながっているノードに注目して、三段階目以降の 2 チャンクでの繰り返し検索をする。

5 評価実験

5.1 ネットワーク生成実験

生成実験では、3 つの入力キーワードを用いて因果関係ネットワークを生成した。表 4 に実験で使用した三つのクエリを示す。キーワード「少子化」、「高齢化」、「生活水準の悪化」を入力クエリとして生成したネットワークを図 5 に示す。

5.2 ネットワーク比較実験

比較実験では、因果関係ネットワークの生成において何段階目の要因・結果検索で三事象が全てパスでつながるか、可視化される全ノード数、入力キーワード間の最短パスと最長パスの長さを算出する。さらに、可視化したノードの適合率 (事象名として適切かどうか)、因果関係を表すエッジの適合率を評価す

る。また、4.2.1 節で説明した二つの尺度に基づいて繰り返し検索を行った場合生成される因果関係ネットワークにどのような違いがあるか比較する。そして、共起度を考慮せずに無作為に繰り返し検索の対象を選んだ場合の因果関係ネットワークを比較して、共起度を用いることが有用かどうか考察する。可視化した事象名とエッジの表す因果関係が妥当かどうかの判断は著者が行った。表 4 に実験で使用した三つのクエリを示す。また、表 5、表 6 にはそれぞれ Simpson 係数、Jaccard 係数に基づいて繰り返し検索を行って生成した因果関係ネットワークについての各指標をまとめた。表 7 には共起度を考慮せずに無作為に繰り返し検索の対象を選んだ場合の因果関係ネットワークの評価をまとめた。なお、表 7 の因果関係ネットワークを生成する際の繰り返し検索をするノードの決定方法については、4.2 節の繰り返し検索対象の選び方を採用し、共起度にて候補の中から 1 組を選ぶのではなく候補の中から無作為に 1 組を選んだ。

三つ全ての入力キーワードを含む因果関係ネットワークを生成するために何段階の検索をしたかに関しては、Simpson 係数、

表 6 Jaccard 係数に基づく因果関係ネットワーク

	クエリ 1	クエリ 2	クエリ 3
検索段階数	2	3	3
全ノード数	15	20	16
最長パスの長さ	4	3	3
最短パスの長さ	2	1	2
適合率 (適切な事象名/ノード数)	0.67 (8/12)	0.88 (15/17)	0.62 (8/13)
適合率 (妥当な因果関係/エッジ数)	0.74 (11/15)	0.95 (20/21)	0.60 (9/15)

表 7 ランダムで繰り返し検索をした因果関係ネットワーク

	クエリ 1	クエリ 2	クエリ 3
検索段階数	4	4	6
全ノード数	27	27	38
最長パスの長さ	4	3	9
最短パスの長さ	2	1	2
適合率 (適切な事象名/ノード数)	0.63 (15/24)	0.83 (20/24)	0.49 (17/35)
適合率 (妥当な因果関係/エッジ数)	0.52 (14/27)	0.92 (24/26)	0.64 (25/39)

Jaccard 係数のどちらを用いた場合でも平均は同じだった。同様に、全パスのノード数、入力キーワード間の任意の 2 ノード間の最短パスの長さに関してもほとんど差は認められない。ノードの適合率の項目に関しては、Jaccard 係数を用いた場合の方が精度が高い。これは Simpson 係数では分母に単語 w_1, w_2 の検索結果数 $|W_1|, |W_2|$ の少ない方を用いるため、いずれかの検索結果数が少なすぎる場合 Simpson 係数に与える影響が大きくなり、関連性が低い事象の組み合わせであるにもかかわらず繰り返し検索されていることに起因している。同様の理由で繰り返し検索の回数が増えるため入力ノード間の最長パスが Jaccard 係数よりも Simpson 係数を用いたネットワークの方が長くなっている。表 7 では、共起度を基にしてノードを選んだ因果関係ネットワークに比べてパスが繋がった段階数と全ノード数が多い。これは共起度が低い組み合わせを選んだ際に、有用な因果関係を抽出できる確率が低くなることに起因する。エッジの適合率が低いのは関連性の低い事象の組み合わせが繰り返し検索の対象に選ばれた際に、ため妥当ではない事象や因果関係が抽出されやすくなるためである。妥当ではない因果関係を取得する例を示す。「東日本大震災」を入力キーワードにして要因検索及び結果検索をすると「東日本大震災に伴うお知らせ」という文章が 1 万件近く得られる。これを係り受け解析すると要因が「東日本大震災」、結果が「お知らせ」となるのだが、この因果関係は妥当であるとはいえない。検索式を生成して検索した際に、取得した要因または結果の事象名が適切かどうかを判断する処理を設けることで解決できる。また、それにより因果関係がわかりやすいネットワークになる。

6 おわりに

本稿では、青野ら、武田ら、藤原らのシステムを拡張して、調べたい三事象間の因果関係を Web より抽出し、因果関係ネットワークとして可視化する手法を提案した。また係り受け解析の条件を改良することで、より重要度の高い語句を抽出する手法も提案した。

今後の課題として、任意の四つ以上の事象入力に対する因果関係ネットワークの構築を検討したい。また、事象が増えることで可読性の低下が起ころうるので、多数の事象を表示させつつ可読性の低下を抑えた可視化方法を検討したい。

文 献

- [1] 青野荘志, 太田学: 要因検索による因果関係ネットワークの構築と因果知識の獲得, DEIM Forum2010, B9-1, 2010.
- [2] 武田真輝, 太田学: 因果関係ネットワーク構築のための要因検索法の改良, WebDB Forum2011, 2G-1, 2011.
- [3] 武田真輝, 太田学: 要因検索における要因抽出法の改良と結果検索の提案, WebDB Forum2012, B1-1, 2012.
- [4] 藤原彰彦, 太田学: 二つの事象間の因果関係分析のための因果関係ネットワークの提案, 電子情報通信学会 2013 年総合大会 ISS 特別企画「学生ポスターセッション」, 2013
- [5] 佐藤浩次, 笠原要, 松沢和光: テキスト上の表層的因果知識の獲得とその応用, 電子情報通信学会技術研究報告, Vol. 98, No. 640, pp. 27-32, 1999.
- [6] Khoo, C.S.G., Chan, S. and Niu, Y.: Extracting Causal Knowledge from a Medical Database Using Graphical Patterns, In: Proceedings of 38th Annual Meeting of the ACL, Hong Kong, pp. 336-343, 2000.
- [7] 乾孝司, 乾健太郎, 松本裕治: 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得, 情報処理学会論文誌, Vol. 45, No. 3, pp. 919-933, 2004.
- [8] 佐藤岳文, 堀田昌英: Web マイニングを用いた因果ネットワークの自動構築手法の開発, 社会技術研究論文集, Vol. 4, pp. 66-74, 2006.
- [9] 坂地泰紀, 竹内康介, 増山繁, 関根聡: 構文パターンを用いた因果関係の抽出, 言語処理学会第 14 回年次大会論文集, pp. 1144-1147, 2008.
- [10] 石井裕志, 馬強, 吉川正俊: 概念と構造を考慮した事象の類似判定に基づく因果関係ネットワークの増分構築, WebDB Forum2010, 3B-3, 2010.
- [11] 山田一郎, 小早川健, 三浦菊佳, 住吉英樹, 八木伸行, 崔紀鮮: クローズドキャプションを対象とした因果関係知識抽出の検討, 第 4 回情報科学技術フォーラム, No. 2, E-001, pp. 113-114, 2005.
- [12] 高見真也, 田中克己: ブログのコミュニティ分析による因果関係事象の抽出, 電子情報通信学会研究報告. データベース・システム研究会報告 2006(78), pp. 285-290, 2006.
- [13] 大森信行, 森辰則: 不具合事例文書からの製品・部品に関する因果関係抽出の検討, 言語処理学会第 18 回年次大会発表論文集, pp. 1192-1195, 2012.
- [14] 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ストラップ抽出, DBSJ Journal, Vol. 8, No. 1, pp. 11-16, 2009.
- [15] 日本語係り受け解析器 CaboCha
<http://chasen.org/~taku/software/cabocha/>
- [16] 日本語形態素解析器 MeCab
<http://chasen.org/~taku/software/mecab/>
- [17] Microsoft が提供する検索 Web API
<http://datamarket.azure.com/dataset/bing/search/>
- [18] Microsoft が提供する検索 Web API (Web Results Only 版)
<http://datamarket.azure.com/dataset/bing/searchweb/>
- [19] ネットワーク可視化・分析ツール JUNG
<http://jung.sourceforge.net/>