新聞記事コーパスからの統計量表現の自動抽出と共起関係ネットワーク構築

齋藤悠† 河合英紀‡ 土田正明† 水口弘紀† 久寿居大†
†NEC インターネットシステム研究所 ‡NEC 中央研究所

1. はじめに

世の中の動向を把握するためには、世の中で起こっている現象を分析することが有効であるが、複雑な仕組みを持つ現代社会では、一つの現象のみに注目した分析だけでは十分でない場合がある。例えば昨今の環境問題に代表されるように、問題の一側面を取り上げて部分的に最適化するだけでは解決できない問題が数多く存在する。このように多くの要因が複雑に絡んだ問題へは、局所的な解法ではなく大域的な解法が求められる。そのためには複数の現象に関する情報を統合し、問題を総合的に捉え分析する必要がある。

一方、情報の電子化に伴い、大量の情報が利用可能となってきている。世の中の様々な現象に関する情報は動向情報として新聞記事に記述されていることが多い。新聞記事における動向情報は、実世界の現象をある観点から計測した値である統計量を基に作成されている。すなわち、新聞記事での動向情報とは、「出生率が 20%上昇」「不況により内閣支持率が低迷」等、統計量に対応する統計量表現とその動向表現、統計量に関連するイベント表現を用いて記述されていることが多い。統計量表現とは、統計量の測定内容を示す表現であり、「出生率」や「内閣支持率」等に相当する。動向表現とは統計量表現の動向を示す表現で「20%上昇」や「低迷」等に相当する。イベント表現とは統計量の推移と関連のある現象を示す表現で「不況」「暴動」「冷夏」等に相当する。

大量の新聞記事データから、世の中の様々な現象に関する動向情報を抽出し可視化できれば、世の中で起こっている現象を捉え動向を把握することができる。大量の新聞記事データから動向情報を抽出し可視化するためのアプローチは二つある。一つは動向情報の基である統計量の数量そのものの可視化であり、例えば統計量の時間推移や地理的な分布をグラフ化する方法である。これは実世界でのある現象に注目した動向情報の抽出、可視化であるといえる。もう一つは、統計量同士の関係性の可視化であり、例えば異なる統計量表現の因果関係を抽出しネットワーク表示する方法である。これは実世界での現象の因果関係や発生メカニズムに注目した動向情報の抽出、可視化であるといえる。

本稿では、世の中の現象を総合的に捉えることを目的に、後者の統計量表現の因果関係ネットワーク構築のアプローチによる動向情報の可視化を目指す。統計量表現の因果関係ネットワークを構築するためには、二つの課題がある。一つは統計量表現を自動で抽出することであり、もう一つは統計量表現の因果関係を抽出しネットワークを構築することである。統計量表現の自動抽出には、本稿では統計量表現の特徴的なパターンである suffix に注目した抽出方式を採用する。suffix に注目した理由は、統計量表現は測定値である統計量の内容を表すため、数値測定に関連する特有の表現を含んでいることが多いからである。統計量表現の因果関係抽出には、まず共起関係を抽出し、そこから各種関係を分類することで因果関係を抽出する。強い共起関係にある統計量表現は因果関係を持つ可能性が高いと予想するからである。因果関係抽出には、因果関係を表す特徴的な接続表現等に注目する方法も考えられるが、[6]によれば、文書中の因果関係の約 70%以上には手がかりとなる接続表現等が明記されていない。したがって本稿では、最初のステップとして因果関係を包含する共起関係でネットワークを構築する。そして共起関係に包含されている関係を分類することで、共起関係の中に因果関係がどのくらい含まれているか、抽出可能かどうか、について分析する。

2. 関連研究

統計量表現の抽出に関しては、統計量表現を抽出する方式の研究と抽出対象である統計量表現の分析、分類の研究がある。斉藤ら[1]は、数値表現とその周囲の品詞と助詞の出現パターンを手がかりに統計量表現を抽出する手法を提案している。藤畑ら[2]は、数値表現と係り受け関係にある名詞を統計量表現の候補とし、係り受けの種類で順位付けして上位候補を統計量表現として抽出している。上記[1][2]はいずれも統計量表現と数値表現の対応を抽出する目的であるので、数値表現を手がかりに対応する統計量表現を抽出している。一方、本研究では統計量表現の間の因果関係を抽出するため網羅的に統計量表現を抽出しなければならないという目的により、統計量表現の部分文字列である suffix を手がかりに統計量表現を抽出している。統計量表現は「失業率が年々増加傾向にある」のように必ずしも数値表現と対応する形で出現するわけではないため、抽出もれの生じる可能性があるからである。また村田ら[3]は、統計量表現自動抽出のための学習データ作成を目的として統計量表現を修飾語の内容で分類しているが、本研究では階層を考慮した統計量表現の因果関係を見つけることを目的として統計量表現を分類している。

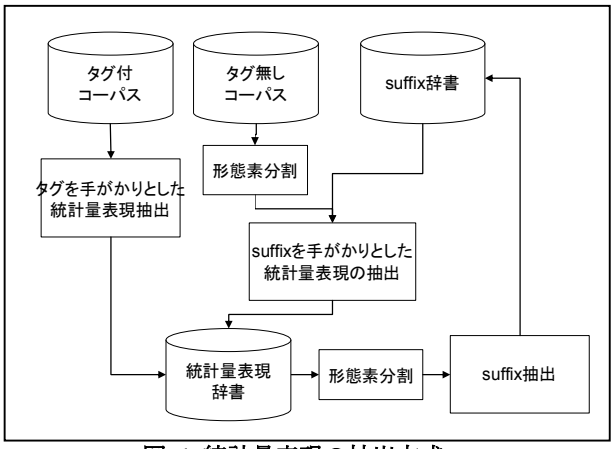

図 1 統計量表現の抽出方式

因果関係の抽出に関しては、「ため」「伴って」等の接続表現や各フレーム等を用いて表層文字列上の因果関係を抽出する研究がある[4][5]。一方で、文書中に記述されている因果関係のうち、約 70%以上のものは表層文字列上、明確な接続表現がないという研究報告もなされている[6]。そこで本研究では、因果関係を包含している可能性の高い共起関係を抽出し、そこに包含されている関係を分類することで因果関係の抽出を試みる。

3. 統計量表現の抽出と分類

本節では、統計量表現に特徴的な suffix を手がかりに統計量表現をコーパスから抽出する方式を提案する。また、因果関係ネットワーク構築での統計量表現の抽象度を揃えることを目的とした統計量表現の分類の定義を示す。

3.1. suffix を利用した統計量表現抽出方式

本稿では、少数の suffix を手がかりに、多数の統計量表現をコーパスから抽出する。すなわち、種となるいくつかの統計量表現を用意し、その文字列特徴である suffix を末尾に含む名詞句を統計量表現としてコーパスから抽出する。図 1 では、本稿での統計量表現の抽出のための処理手順を示している。まずタグ付きコーパスから統計量表現を表すタグに囲まれた単語を種の統計量表現として抽出し、統計量表現辞書に格納する。次に、種の統計量表現を形態素分割し、末尾の 1～3 形態素を統計量表現の suffix として抽出する。最後に、形態素分割したタグ無しコーパスから suffix を末尾とした名詞句を統計量表現として抽出し辞書へ追加することで統計量表現の増殖を行う。

種となる統計量表現は、MuST タグ付きコーパスの<unit stat>タグおよび<name>タグで囲まれた文字列とする。MuST タグの仕様[7]によれば、<unit stat>および<name>は、可視化対象の統計量に言及している部分および統計量の名前を表すタグであり、我々が抽出対象として定義した統計量表現と一致している。したがって本稿では、上記二種類のタグに囲まれた文字列を統計量表現として扱う。

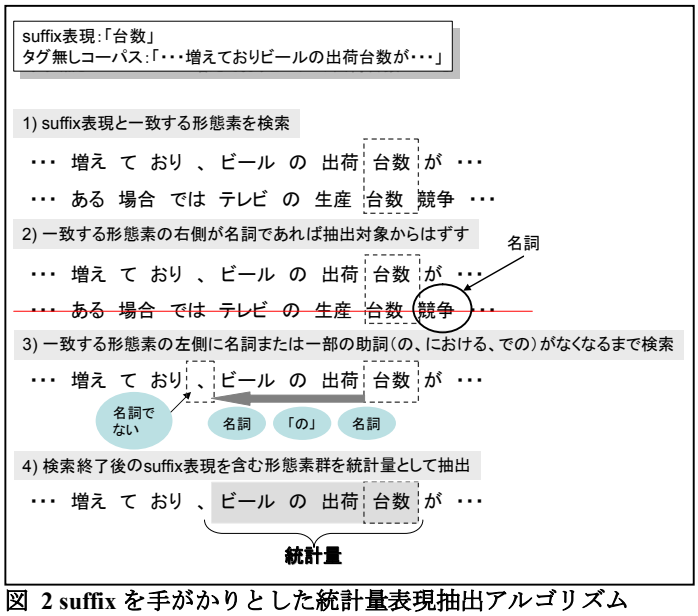

図 2 suffix を手がかりとした統計量表現抽出アルゴリズム

suffix を手がかりにした統計量表現の抽出では、suffix を末尾とした修飾表現までを含む名詞句を統計量表現として抽出する。具体的には、suffix と一致する名詞部分をコーパスから特定し、その前方の名詞あるいは助詞の連続部分と suffix とを統計量表現として抽出する。図 2 の例を用いてより詳細にアルゴリズムを説明する。図 2 では suffix を「台数」とし、コーパスを 2 文「・・・増えており、ビールの出荷台数が・・・」「・・・ある場合ではテレビの出荷台数競争・・・」とした場合の例である。

1) 形態素分割したタグ無しコーパスから suffix と一致する形態素を検索

図 2 では、suffix「台数」が 2 文ともに一致している。

2) suffix と一致する形態素の右側形態素をチェックし名詞であれば、抽出しない

図 2 では、後者の文は、「台数」の右側に名詞「競争」があり、統計量表現ではない別の複合名詞の一部である可能性が高い。したがって、この文からの抽出は行わない。

3) suffix と一致する形態素の左側の形態素に、名詞または一部の助詞のいずれも出現しなくなるまで検索

図 2 では、「台数」の左側形態素が「出荷（名詞）」「の（助詞「の」）」「ビール（名詞）」と続き、「、(名詞/助詞ではない)」になったところで検索が終了する。

4) 基点の形態素から左側終了地点の形態素までを統計量表現として抽出

図 2 では、「、」で検索が終了となったので、その直後の形態素から suffix までの「ビールの出荷台数」を統計量表現として抽出する。

ただし、4) で抽出された統計量表現のうち、「の」で始まるものは名詞句として適切でないため除外する。

さらに、一致する suffix の形態素数と出現頻度を元にスコアを定義し、抽出した各統計量表現のスコア付けを行う。これは上記アルゴリズムで抽出した統計量表現の中からより統計量表現としてふさわしいものだけを抽出するためである。 スコア付けは 1) 形態素数の多い suffix に一致し、2) 出現頻度の高い suffix に一致するほどより統計量表現らしい、という立場で行う。すなわち、形態素数が N で出現頻度 R の suffix に一致して抽出された統計量表現のスコア S

を以下のように定義する。

$$S = 100^{(N-1)} \times R$$

3.2. 統計量表現の分類

前節で自動抽出した統計量表現の抽象度を分析するために、その修飾語の種類に注目して統計量表現を分類する。統計量表現の分類は文献[3]においても研究されているが、本研究では、統計量表現の因果関係ネットワーク構築を目的とした分類を試みる。我々は統計量表現の中から因果関係のあるものを繋いでネットワークを構築するとき、ネットワーク内の統計量表現の抽象度を揃える必要があると考えている。例えば、「率」を suffix に持つ統計量表現には「率」「失業率」「アメリカの失業率」「国内の失業率」「9 月の失業率」等があるが、ここで、「失業率」と「アメリカの失業率」とを考えると、それぞれ因果関係ネットワーク構築に必要な統計量表現が異なるはずである。なぜなら、「失業率」の因果関係ネットワーク構築には、抽象度が高く失業率一般と因果関係にある統計量表現が必要であるのに対し、「アメリカの失業率」の因果関係ネットワーク構築は、アメリカ独自の事象を示す統計量表現が必要だからである。

上記の目的により、1) 多くの統計量表現に共通する種類の修飾語である、2) 同じ分類ラベルに属する修飾語は同時に複数付与されない、という観点で 4 種類の分類ラベルを定義し、4 種類のどれにも属さない修飾語をその他とする。分類ラベルの内容を図 3 の分類ラベル階層図の例を用いて説明する。

- 基底統計量表現
  統計量表現の中で、形態素数が最少でかつ統計量の内容が判別可能な名詞句。図 3 では「失業率」が相当する。

a) 対象物
統計量表現の中で統計量の測定対象となっている人、団体、ものを示す修飾語。図 3 では「サラリーマンの」が相当する。

b) 主体
統計量表現の中で統計量に対して制御もしくは影響力を及ぼすことのできる人、団体、ものを示す修飾語。図 3 では「IT 企業の」「大手自動車メーカーの」が相当する。

c) 期間
統計量表現の中で統計量の測定対象期間を示す修飾語。図 3 では「9 月の」「8 月の」「98 年度上期」が相当する。

d) 地域
統計量表現の中で統計量の測定範囲の場所を示す修飾語。図 3 では「アメリカの」「国内の」が相当する。

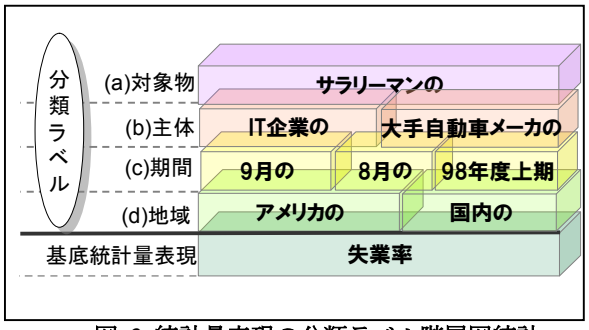

図 3 統計量表現の分類ラベル階層図統計

a)の対象物と b)の主体との違いは、統計量の数値の測定対象そのものと、数値を変化させることができるものとの違いである。すなわち、「サラリーマン」は失業率を測定する対象ではあるが失業率に影響を及ぼすことはできない。一方「IT 企業」は従業員の失業についての決定権があり、失業率に影響を及ぼすことができる。

このような分類ラベルを定義することで、統計量表現の抽象度が分かる。すなわち、多くの分類ラベルが付与された統計量表現ほど抽象度が低く複雑であると判断できる。例えば「9 月の国内の IT 企業サラリーマンの失業率」と「サラリーマンの失業率」では前者の方が多くの分類ラベルが付与されるため抽象度が低く複雑な統計量表現であることが分かる。

4. 統計量表現の共起ネットワーク

本節では、抽出された統計量表現を使った共起ネットワークの構築方式を説明する。

本稿では、システムが出力した統計量表現の共起関係から必要な統計量表現の関係だけを人手で選択し、インタラクティブにネットワークを作成していく。このような作成方法を採用する理由は、共起関係を持つ統計量表現から我々が目的とする因果関係を持つ統計量表現のみを抽出しネットワークを構築したいからである。

本稿では、二つの統計量表現が一つの記事に同時に出現するとき、それら二つの統計量表現は共起関係にある、または共起するという。また、二つの統計量表現が共起している記事の数をそれら二つの統計量表現の共起頻度という。

以下、本稿でのインタラクティブなネットワーク構築方式について説明する。まず、共起関係にある統計量表現を抽出する。具体的には、各統計量表現に対して出現記事 ID を付与しておき、共通の記事 ID を持つものを共起する統計量表現ペアとして抽出する。

次に、フィルタリングを行い、多くの統計量表現と共起する統計量表現を除外する。統計量表現をノードとする共起ネットワークを構築するとき、これらのノードがハブとなり、ほとんどのノードを繋げてしまうからである。本稿では 50 以上の統計量表現と共起するものをフィルタリングの対象とする。例えば「割合」「経常利益」といった統計量表現は 50 以上の統計量表現と共起するため除外する。

最後に、注目すべき統計量表現に共起する統計量表現を 1 ホップ表示し、興味ある統計量表現を目検で選択しそれに共起する統計量表現をまた 1 ホップ表示させて選択する、ということを繰り返してネットワークをインタラクティブに展開していく。最も外側のノードに共起する新たなノードが見つからなくなったところで展開を終了する。

5. 実験結果と考察

本節では、suffix に注目した統計量表現抽出方式で、多数の統計量表現が抽出できたことを示す。また、抽出された統計量表現を修飾語の内容で分類したところ、分類は期間、対象物の順で多く、複数の分類を持つ統計量表現は少なかったことを示す。統計量表現の共起関係ネットワーク構築では、一見、関連が推測困難な統計量表現についてその関係を発見できたことを、例を用いて示す。さらに共起する統計量表現のペアを分類したところ、直接、間接関係にあるペアが多かったことを示す。

5.1. 統計量表現抽出結果

5.1.1. 実験方法

3.1 節の suffix に注目した統計量表現抽出方式で、新聞記

事コーパスから統計量表現の抽出実験を行った。対象としたコーパスは1998年と1999年の毎日新聞約23000記事で、タグ付きの約1000記事とタグ無しの約22000記事で構成されている。タグ付コーパスから抽出した統計量表現は86単語であり、これを種の統計量表現とする。

表 1に、タグ付きコーパスから抽出した統計量表現の一例を示す。また、種の統計量表現から抽出したsuffixは146であり、表 2に抽出したsuffixの一例として、頻度ごとにサンプリングしたものを示す。このsuffixを用いて3.1節の方式で統計量表現を抽出しスコアを付与した。

抽出した統計量表現のうち、スコア値5以上16000語からランダムサンプリングした200語、およびスコア値最上位から1000語、の二種類を対象に、目視確認での精度評価を行った。評価方法は、各統計量表現について3名の評価者が各々に評価し、統計量表現であるかそうでないかを判定する。評価が異なる場合は「統計量表現でない」とし、評価対象の全統計量表現のうち3名の評価者全員が「統計量表現である」と判定した語の割合を精度とした。

表 1 タグ付きコーパスからの統計量表現

タグ付きコーパスからの統計量表現
男女別完全失業者数
PHSの加入台数
発泡酒のメーカー別出荷量シェア
国内出荷台数
公示地価
主要メーカーの総生産台数
実質消費支出
実況判断DI
花粉の飛散量
レギュラーガソリンの全国平均店頭価格

表 2 統計量表現からのsuffixとその頻度例

統計量表現のsuffix	頻度
数	12
率	6
加入台数	5
相場	4
量	4
指数	4
失業者数	3
の下落幅	3
支出	2
数量	2
全国売上高	2
出生率	1
原油価格	1
国内生産台数	1
割合	1

5.1.2. 実験結果

33100語の統計量表現が抽出でき、16000語からランダムサンプリングした200語については精度約0.67、スコア上位1000件については精度約0.84であった。また、3名の評価者の評価結果の一致度は、サンプリング200語については87%、上位スコア1000語については88%であった。これより提案方式は統計量表現の増殖には効果的であることが確認できる。

表 3にスコア上位10件の統計量表現を示す。また、図 4にランダム200件の統計量表現のスコア値精度評価の結果を、図 5に統計量表現のスコア順位と精度評価の結果を、それぞれ累積グラフで示す。図 4ではx軸がスコア値、y軸がスコア値100000からxまでの累積精度を表している。図 5ではx軸がスコア順位、y軸が1位からx位までの累積精度を表している。図 4ではスコア値が下がると累積精度がなだらかに落ちて行くことが分かる。スコアを上位1000件に絞った図 5では、スコア順位が低下すると精度も低下することがさらに明確に読み取れ、スコアと精度に相関があることが分かる。これより本稿で定義したスコアの正当性が確認できる。スコア上位16000件までの統計量表現の精度は約0.67以上であり、特に上位1000件については約0.84以上であった。

誤り例は、「最悪の完全失業率」「1回の割合」「年間最高の127万の患者数」「学校数や在学者数」等であった。

5.1.3. 考察

・ 網羅性について

本実験では、統計量表現に特徴的なsuffixを用意し統計量表現を抽出した。抽出された統計量表現の精度は高かったが、網羅性についての課題が残る。例えば、統計量表現の中には、「GDP」等のsuffixに当てはまらないものも存在する。今後は網羅性へも対応した、suffix以外の統計量表現抽出方式を検討する必要がある。

・ イベント表現抽出について

本実験では、事象の抽出対象を統計量表現に限定していた。統計量表現以外に動向情報に関連するものとして、「暴動」「冷夏」等のイベント表現がある。イベント表現と統計量表現とには因果関係がある場合が多いため抽出する必要がある。suffixは、統計量表現以外の動向表現やイベント表現等の事象表現の抽出への適用が困難である。なぜなら動向表現やイベント表現はsuffixに代わる特徴的な文字列表現を見つけにくいからである。今後は、統計量表現以外の事象の抽出へも適用できる方式を検討する必要がある。

表 3 suffixを手がかりに抽出した統計量表現（スコア上位10件）

suffixを手がかりに抽出した統計量表現	スコア
完全失業率	3620000
日経平均株価	3310000
完全失業者数	1560000
内閣支持率	1000000
有効求人倍率	920000
失業者数	810000
3月の完全失業率	560000
政党支持率	500000
4月の完全失業率	400000
男性の完全失業率	400000

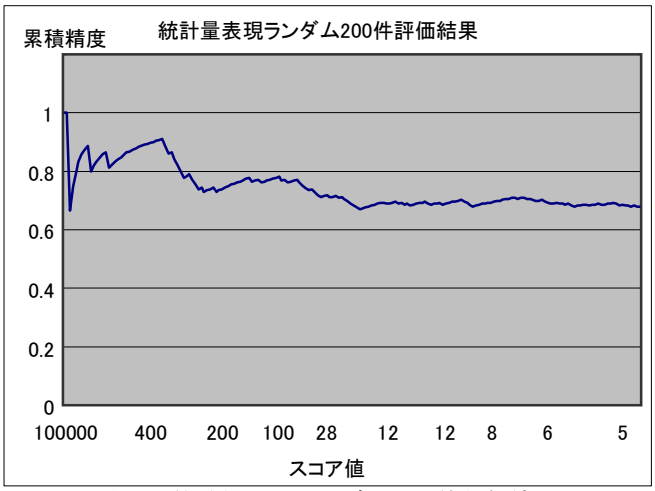

図 4 統計量表現ランダム 200 件評価結果

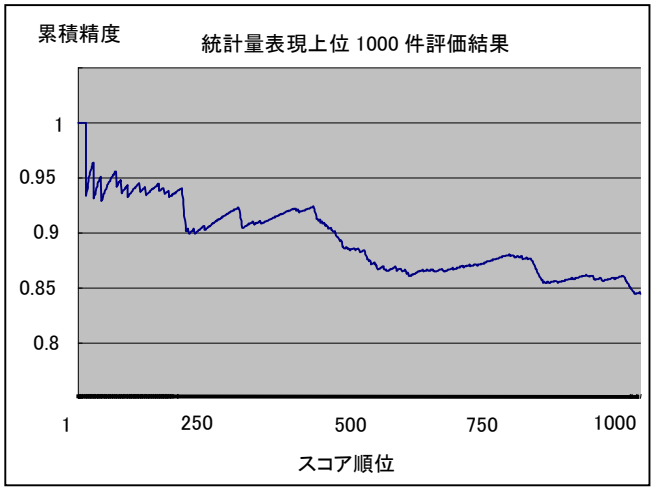

図 5 統計量表現スコア上位 1000 件評価結果

5.2. 統計量表現の分類結果

5.2.1. 実験方法

抽出した統計量表現のうち正しく抽出された統計量表現スコア上位 200 件について、3.2 節の分類基準を適用し人手による分類を行った。

表 4 統計量表現の分類（上位 200 件）

分類ラベル (件数)[割合]	分類例
(a)対象物 (48) [24%]	PHS の加入台数、ダイヤの売上高、橋本内閣支持率
(b)主体 (4) [2%]	百貨店売上高、府立高校の中退者数
(c)期間 (97) [49%]	2 月の国内卸売り物価指数、昨年の総入場者数
(d)地域 (38) [19%]	2 月の国内卸売り物価指数、関空出国者数
(e)基底統計量表現 (18) [9%]	観客数、出入国者数、退職者数、支持率、失業率、失業者数
その他 (18) [9%]	総合卸売り物価指数、地域別の完全失業率

5.2.2. 実験結果

括弧内の数値は該当する統計量名の数を示している。表 4 の結果から、「期間」の表現を含む統計量表現が最も多く、「地域」の表現を含む統計量表現が最も少なかった。また、修飾語である「対象物」「主体」「期間」「地域」が付与された統計量表現 180 語のうち、1 種類のラベルの付いた統計量表現は 156 語(87%)、2 種類のラベルの付いた統計量表現は 23 語(13%)、3 種類のラベルのついた統計量表現は 1 語(0.5%)であった。これより、スコア上位 200 語の統計量表現には、修飾語ラベルが複数付与される抽象度の低い統計量表現が少ないことが分かった。

5.2.3. 考察

複数の分類ラベルを付与された統計量表現が少なかったという結果より、本稿での suffix に注目した統計量表現抽出方式は、抽象度の低い複雑な統計量表現を抽出するには適さないことが分かる。抽象度が低く複雑な統計量表現は修飾される語が多いが、その修飾語すべてを一つの名詞句に含んでいる場合は少ない。多くの場合、タイトルや前の文等で既に修飾され、それが複数回積み重なることで抽象度が低く複雑な統計量表現となると考えられる。したがって、共起する統計量表現のうち基底統計量表現が共通しているものは修飾語を残して纏め上げる処理等で、抽象度が低く複雑な統計量表現の抽出にも対応していく必要がある。

5.3. 共起関係ネットワーク構築結果

5.3.1. 実験方法

統計量表現とタグ無しコーパスから、4 節のインタラクティブなネットワーク構築方法で共起関係ネットワークを構築した。統計量表現は 5.1 節で抽出した 33100 語、タグ無しコーパスは約 22000 記事で統計量表現抽出に用いたものと同様である。ただし、統計量表現 33100 語には統計量表現でなく誤って抽出された語も含まれている。使用したネットワーク表示ツールは、本研究所にて独自に研究開発されたものである[8]。

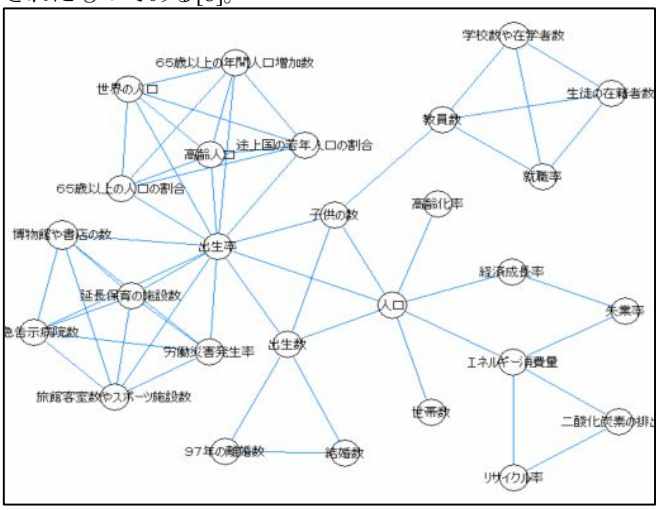

図 6 “出生率”中心の共起関係ネットワーク

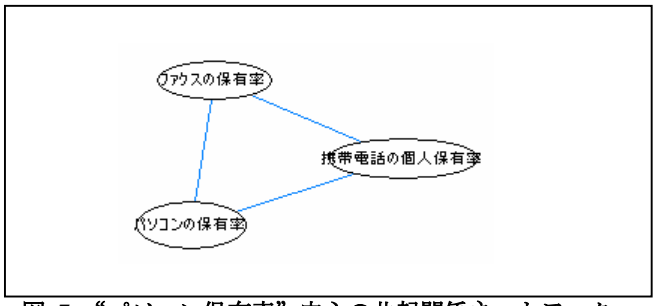

図 7 “パソコン保有率”中心の共起関係ネットワーク

5.3.2. 実験結果

図 6 では「出生率」を中心に、図 7 では「パソコン保有率」を中心に、3 ホップまで展開した統計量表現の共起関係ネットワークを示している。各ノードは統計量表現であり、アークで繋がれているノード同士は共起関係にある。

5.3.3. 考察

図 6 では「出生率」と関連する統計量表現として「リサイクル率」や「二酸化炭素の排出量」といった興味深いものが表示されている。「出生率」と各ネットワークの島の関係の種類を出現記事から調査したところ、左側の「博物館や書店の数」「延長保育の施設数」「緊急告示病院」等は、関連性はあるが因果関係はなかった。これらの統計量表現は都道府県の豊かさを表す指標である。一方、右側の「人口」「エネルギー消費量」等とは因果関係があった。したがって本稿での共起関係ネットワークを因果関係ネットワークに絞り込めば、統計量表現同士の興味深い関係の抽出が期待できる。今後は共起関係の中から因果関係を抽出する技術が必要である。

一方、図 7 ではネットワークが広がらず、「パソコン保有率」に関連する興味深い統計量表現は発見できなかった。「パソコン保有率」を基点に、2 ホップ目に新たに追加できる統計量表現がなかったためである。このようにノードを展開できなくなってしまう原因として、ノードにイベント表現が含まれてなかったことが挙げられる。本稿では統計量表現をノードとして共起関係ネットワークを構築したが、事象の興味深い関連性はイベント表現を介して説明されることが多い。そのため、今後はイベント表現も考慮したネットワーク構築技術が必要である。

5.4. 共起関係の分類結果

5.4.1. 実験方法

5.3 節で抽出した共起関係の中から、共起強度の高い順に上位 200 ペアについて、評価者 3 名の人手による分類を行った。共起強度は、二つの共起する統計量表現で出現が少ない方の出現回数に対する共起回数の割合とした。分類は「同義」「直接的な関係」「間接的な関係」「無関係」の 4 種類とした。直接的な関係とは二つの統計量表現が従属関係にあることで、間接的な関係とは二つの統計量表現は独立であるが、共通の要素と従属関係にあることとする。評価者は、評価対象の統計量表現 200 ペアについて、各ペアがこれら 4 種類のどの関係に属するかを評価者の持つ知識によって判定した。また、各ペアが出現する 617 文を読み、4 種類の関係が文書上に明示されているか、明示されているものについては接続表現など手がかりとなる表現があるか、を目視確認した。評価者により判定が分かれた場合は多数決により判定内容を決定した。

5.4.2. 実験結果

共起強度の高い統計量表現 200 ペアには、直接的、間接的な因果関係がこの順で多く含まれており、それぞれ約半分は文書上でその関係が明示されていた。しかし手がかりとなる表現は少なかった。3 人の評価者による評価の一致度は関係が文書上に記述があるかどうかの判定については 48%、手がかり表現があるかどうかの判定については 49%といずれも低く、ゆれが大きかった。表 5 では、分類結果の詳細を示している。共起関係に含まれる各関係について、その件数と、文書上に関係が明記されているかどうかの件数、手がかりとなる文字列上の表現があるかどうかの件数を示している。関係が「不明」とは、評価者の多数決では関係の種類が決定できなかったペアを表す。括弧内の数値は割合を示しており、件数の列では 200 ペアに対するそれぞれの関係の割合、文書上の記述の有無もしくは手がかり表現の有無の列では、各関係の件数に対する割合を示している。例えば、直接的因果関係のペアは 119 件あったので、200 件中の割合としては 60%である。その中で文書上での明記があったものは 68 件で 119 件に対する割合は 57%である。

この表から、共起強度の強い関係からは直接的な因果関係があるものが多かったことが分かる。それらの約半数は関係が文書上で記述されているが、接続表現などの手がかりとなる表現はなかったことが読み取れる。対照的に同義関係については、その関係のほとんどは文書上で記述されており、かつ、手がかり表現があったことが分かる。

見つかった手がかり表現の例としては、直接的な関係では「～ので」「の結果」等、間接的な関係では「～の背景にある」「～や…が急上昇」等、同義関係では「() 括弧表現」「～を示す」等であった。

また、文書上に関係が明記されている場合とされていない場合を詳細に分析したところ、例えば、失業率と失業者数など明らかに関係が分かる統計量表現ペアについては、文書上にその関係は明記されておらず、大口電力消費量と百貨店売上高など一見無関係な統計量表現ペアについては文書上その関係が明記されていた。これは記事執筆者が読者の知識レベルを想定して関係を明記するかどうか判断した結果であると考えられる。

5.4.3. 考察

本実験での分類結果より、共起関係の中には直接的、間接的を合わせると関係があるものは 173 件（約 87%）であったことが分かる。しかしその中で手がかり表現があったものは 39 件（約 22%）であり、文書中に現れる因果関係のほとんどには接続表現などの手がかりとなる表現がないことが推測される。一方、同義関係については、文書上に明記されているものについては括弧表現等の手がかり表現が明記されていた。したがって、無関係は含まれるが、共起強度の高い共起関係から手がかり表現を用いて同義表現を除くことで、本実験では精度約 92%での因果関係の抽出が可能となる。

表 5 共起関係の分類結果

関係の種類	件数 [%]	文書上の記述		手がかり表現	
		あり [%]	なし [%]	あり [%]	なし [%]
直接的	119[60]	68[57]	51[43]	32[27]	87[73]
間接的	54[27]	23[43]	21[57]	7[13]	47[87]
同義	12[6]	10[83]	2[17]	10[83]	2[17]
無関係	12[6]	0[0]	12[100]	0[0]	12[100]
不明	3[1.5]	--	--	--	--

6. おわりに

本稿では、多角的な視点から世の中の動向を把握するための一手法として、統計量表現の因果関係をネットワーク表示することを提案した。そのためにまず、新聞記事コーパスから統計量表現の suffix を手がかりに、統計量表現 16000 語を約 67%の精度で、1000 語を約 84%の精度で抽出した。また抽出した統計量表現について抽象度を整理するという目的で分類し、抽象度の低い統計量表現がほとんどないことを示した。次に、抽出した統計量表現の共起関係をネットワーク表示し、統計量表現について興味深い関連性が見つかった例と見つからなかった例についてその原因を考察した。また、抽出した共起関係が包含する関係を分類し、直接的あるいは間接的な因果関係が多く含まれているが、手がかり表現は少ないことを示した。今後は、接続表現や動向表現等に注目し、統計量表現やイベント表現間の因果関係の抽出に取り組む予定である。

参考文献

[1] 斉藤公一，迫田昭人，中江富人，岩井禎広，田村直良，中川裕志.“数値表現をキーとした新聞記事からの情報抽出.”自然言語処理研究会報告 1998-NL-125，情報処理学会，1998.

[2] 藤畑勝之，志賀正裕，森辰則.“係り受けの制約と優先規則に基づく数量表現抽出.”自然言語処理研究会報告 2001-NL-145，情報処理学会，2001.

[3] 村田一郎，森辰則.“新聞記事中の統計量名の学習による自動抽出.”NTCIR-5 Pilot Workshop MuST 成果進捗報告会論文集，2006.

[4] 佐藤浩史，笠原要，松澤和光.“テキスト上の表層的因果知識の獲得とその応用.”信学技報 TL98-23，pp. 27—34，電子情報通信学会 1998.

[5] 佐藤岳文，堀田昌英.“Web マイニングを用いた因果関係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.

[6] 乾孝司，奥村学.“文書内に現れる因果関係の出現特性調査.”計量国語学 Vol.25, No.3, 2005.

[7] 松下光範，加藤恒昭.“動向情報に基づく情報可視化の基礎検討.”2005 年度人工知能学会全国大会，1E3-03 , 2005.

[8] Hironori Mizuguchi, Dai Kusui, Taku Oshima, Shigehiko Kanaya, Hirotada Mori,“KAREIDMAP: A System for Predicting and Mining Gene Regulatory Networks.” Genome Informatics 14: 382-383 , 2003.