3K3-10

新聞記事における統計量表現の共起ネットワーク

Co-occurrence Network of Statistical Terms in Newspaper

河合 英紀[*1] 齋藤 悠[*2] 土田 正明[*2] 水口 弘紀[*2] 國枝 和雄[*1] 山田 敬嗣[*1]
Hideki Kawai Haruka Saito Masaaki Tsuchida Hironori Mizuguchi Kazuo Kunieda Keiji Yamada

[*1]NEC C&C イノベーション研究所
NEC C&C Innovation Research Laboratories

[*2]NEC 共通基盤ソフトウェア研究所
NEC Common Platform Software Research Laboratories

In this paper, we will describe a simplifying method for global dynamics visualization. Global dynamics of various events and statistics are important to analyze complex international issues such as environmental, economic and political problems. We have been developing a system which can extract a co-occurrence network of statistical terms like birth rates, oil prices or energy consumption by matching a suffix patterns of statistical terms. However, the network structure consisting of thousands of statistical terms is too complicate to understand their causal relations briefly. So we propose a method for simplifying the network structure based on network complexity and language expressions. Our experimental result shows that a clique of the statistical terms corresponds to a certain topic or issue and causal relations can be described as a chain of the cliques on the network structure.

1. はじめに

現代社会では、様々な事象がお互いに影響を及ぼしあっているため、単一の事象にのみ注目して分析や最適化を行うと、思わぬところに副作用が出てしまうおそれがある。例えば、石油を取り巻く情勢や政党支持率の動き、地球規模での気候の変化など、入り組んだ構造を持つ経済・政治・環境問題等では、様々な事象のグローバルなダイナミクスを把握しなければ、全体最適な解決策を見出すことは困難である。そのため、現実世界の多種雑多な情報の中から動向情報を編集・可視化する情報編纂技術の重要性が高まっている [加藤 07]。

筆者らは、グローバルダイナミクスの可視化手法として、統計量表現の共起ネットワークを構築する方法を研究している [齋藤 07, Saito 07]。「出生率」や「失業率」、あるいは「エネルギー消費量」といった統計量表現は、世の中の様々な事象をある側面から定量化した指標の名前であるため、統計量表現同士の依存関係を観察することによって、様々な事象間の因果関係の理解の助けとなるからである。これまで、「率」や「量」など、統計量表現の suffix に着目したパタンマッチングによる統計量表現の抽出や、統計量表現の共起ネットワークにおける関数従属性などに関する調査を行ってきた。

しかし、統計量表現の共起ネットワークは非常に構造が複雑で、その全体像を可視化しても理解が難しいという問題があった。その主な原因としては、(1) ネットワーク構造の複雑性と (2) 統計量表現の構造の複雑性の 2 種類がある。

(1) ネットワーク構造の複雑性とは、ノードやエッジの数が多いことに加え、これらが「人口」や「株価」のように多くの統計量表現と共起する統計量表現がハブとなって、非常に密に接続したクラスタを形成してしまう問題である。また、(2) 統計量表現の構造の複雑性とは、統計量表現に含まれる測定期間や測定条件を表す修飾語を区別すべきか否かの判断が困難であるという問題である。例えば、「アメリカの失業率」「男性の失業率」「6 月の失業率」のように、測定地域・対象・期間の異なる統計量表現を厳密に区別すべきか、それとも全ての場合において「失業率」として同一視しても問題ないか、といった判断は自明ではない。

そこで本稿では、ネットワーク構造と統計量表現の構造の両方に着目したグローバルダイナミクスの要約・可視化手法を提案する。実験では、提案手法を新聞記事データに適用し、得られたネットワーク構造について議論する。

連絡先: 河合英紀, NEC C&C イノベーション研究所
〒 630-0101 奈良県生駒市高山町 8916-47, 0743-72-3684
h-kawai@ab.jp.nec.com

2. 関連研究

情報編纂は言語や数値などの多種雑多な情報を知的に編纂し、要約・可視化する技術である。新聞記事における動向表現を行う関連研究として、統計量表現とその値を表す数値表現の組を抽出する技術が挙げられる。数値表現を手がかりにその周辺の統計量表現を見つける手法として、品詞と助詞の出現パターンを利用する方法 [斉藤 98] や係り受け関係を利用する方法 [藤畑 01] などが挙げられる。また、統計量表現のタグセットを定義して、アノテーション付コーパスを構築し、機械学習を使って自動抽出する方法もある [森 07]。本研究では、統計量表現に共通してよく出現する suffix に着目したパタンマッチングを利用した統計量表現抽出を行っている点が特徴である。

また、因果関係を抽出する関連研究としては、「ため」「伴って」等、因果関係を表す特徴的な接続表現を用いる手法 [佐藤 06] や、格フレームを用いて表層文字列上の因果関係を抽出する手法 [佐藤 99] がある。一方で、文書中に記述されている因果関係のうち約 70%以上は表層文字列上、明確な接続関係が出現していないという調査報告 [乾 05] もなされている。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に着目し、共起ネットワークを構築・観察する実験を行う。また、因果関係の対象として統計量表現にフォーカスしている点や、人間にとって理解しやすいシンプルな共起ネットワークを得るために、ネットワーク構造と統計量表現の両方に着目した簡略化方法を行う点も本研究の特徴である。

3. グローバルダイナミクスの可視化

本節では、グローバルダイナミクスの可視化手法として、統計量表現の共起ネットワーク構築方法の概要を説明し、ネットワーク構造と統計量表現の両方に着目した簡略化方法として、

ノードの次数制限と、共通 suffix による統計量表現のクラスタリング手法を提案する。

3.1 統計量表現の抽出

本研究では、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 [齋藤 07]。まず、種となる数十語程度の統計量表現の集合を用意し、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 suffix として抽出する。最後に、形態素分割したコーパスから suffix を末尾とした名詞句を統計量表現として抽出し辞書に追加することで統計量表現の増殖を行う。統計量表現抽出処理は、図 1 に示すように、4 つのステップからなる。

1. 形態素分割したコーパスから suffix と一致する形態素を検索し基点とする。
2. 基点の形態素の右側の形態素をチェックし名詞であれば抽出しない。
3. 基点の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する。

上記の方法では、統計量が測定された地域・期間・条件等を表す修飾語も付属した状態で統計量表現が抽出される。本研究では、統計量表現は基底表現に、(a) 対象物、(b) 主体、(c) 期間、(d) 地域のいずれかまたは複数の修飾語が付与された文字列から構成されていると定義している [齋藤 07]。

基底表現とは、統計量表現の中で、統計量の内容が判別可能な最小単位の形態素列である。例えば、「9 月のアメリカの失業率」という統計量表現では、「失業率」が基底表現にあたる。(a) 対象物とは、統計量の測定対象となっている人、団体、ものを示す修飾語である。例えば、「エアコンの出荷台数」という統計量表現では「エアコン」が対象物にあたる。(b) 主体とは、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、ものを示す修飾語である。例えば、「NEC のパソコンの出荷台数」という統計量表現の中では「NEC」が主体に相当する。(c) 期間は、統計量表現の測定対象期間を示す修飾語である。「9 月の」「98 年度上期」などの表現がこれに当たる。(d) 地域は、統計量の測定対象地域を示す修飾語である、「アメリカの」「首都圏の」などの表現がこれに相当する。

統計量表現同士の因果関係を観察する際は、上記の修飾語がどこまで有効かを考慮する必要がある。例えば、「アメリカの失業率」について考えてみると、国や地域を問わず「失業率」一般に当てはまる因果関係もあれば、アメリカだけに特徴的に見られる因果関係もあるかもしれない。これを決定するためには、統計量表現の表層文字列だけでなく、ネットワーク上で隣接する統計量表現同士の関係性考も慮する必要がある。

3.2 共起ネットワークの簡略化手法

統計量表現が抽出できたら、統計量表現同士の共起関係を求めて共起ネットワークを構築することができる。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共起頻度と呼ぶ。統計量表現をノードとし、共起頻度が閾値 θ 以上の統計量表現間をエッジで結ぶことによって共起ネットワークが得られる。

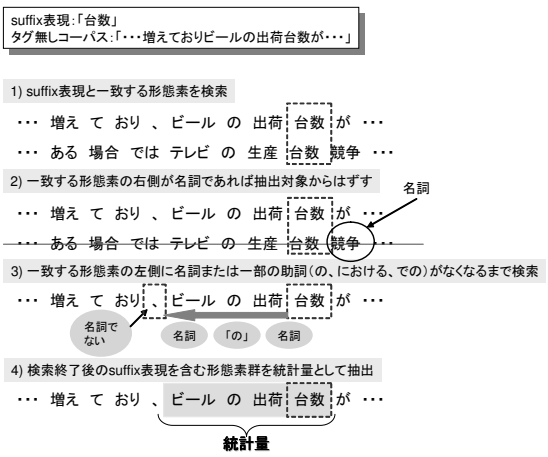

図 1: 抽出アルゴリズムの概要

しかし、このようにして構築された共起ネットワークは、非常に多くのノードが緊密に接続しあっているので、そのまま可視化しても複雑すぎて関係性を把握することはできない。図 2 に、共起ネットワーク構造の例を示す。図 2 では、統計量表現「リサイクル率」を中心とする 2 ホップ先までの統計量表現を可視化しているが、「人口」のように多くの統計量表現と共起し、ネットワークの中でハブとなっている統計量表現が密度の高いクラスタを形成しているため、エッジをたどって統計量表現同士の関係性を評価することが非常に困難である。そこで本稿では、ノードの次数制限と、統計量表現のクラスタリングによって複雑な構造の共起ネットワークを簡略化する。

ノードの次数制限とは、各統計量表現について、共起頻度の上位 ω 語に対してのみエッジを張ることによって、ネットワークの緊密性を緩和する方法である。これにより、ハブノードによる密度の高いクラスタの形成を抑えながら、ネットワークの主要な構造を浮き彫りにすることが期待できる。

統計量表現のクラスタリングとは、ネットワーク上で隣接する統計量表現同士の関係性を考慮しながら、共通の基底表現を持つ統計量表現同士を一つのノードに集約する方法である。提案手法の概要を図 3 に示す。図 3 の (a) では、2 つの統計量表現「国内の失業率」と「アメリカの失業率」が、それぞれ「内閣支持率」と「経済成長率」と共起していることが分かる。一方、「銃所持率」は「アメリカの失業率」としか共起していない。この場合、図 3 の (b) のように、「国内の失業率」と「アメリカの失業率」を共通の基底表現「失業率」にクラスタ化することによって、「内閣支持率」と「経済成長率」に対するエッジをまとめることができる。一方、「銃所持率」と「アメリカの失業率」の関係はアメリカ独自の事情であるので、クラスタ化しないまま残しておく必要がある。

提案手法を一般化すると以下のように記述できる。共通の基底表現 BE と任意の修飾語 M から構成される 2 つの統計量表現 $S_1 = \{M_{11}, M_{12}, ..., M_{1m}, BE\}$、$S_2 = \{M_{21}, M_{22}, ..., M_{2n}, BE\}$ について、共起ネットワーク上で共通のノードと共起している場合は、S_1 と S_2 の共通部分からなる統計量表現 $S_3 = S_1 \cap S_2$ でクラスタ化し、共通ノードへのエッジを S_3 にまとめる。一方、S_1、S_2 にそれぞれ独自に接続しているノードがある場合は、そのエッジは残しておく。

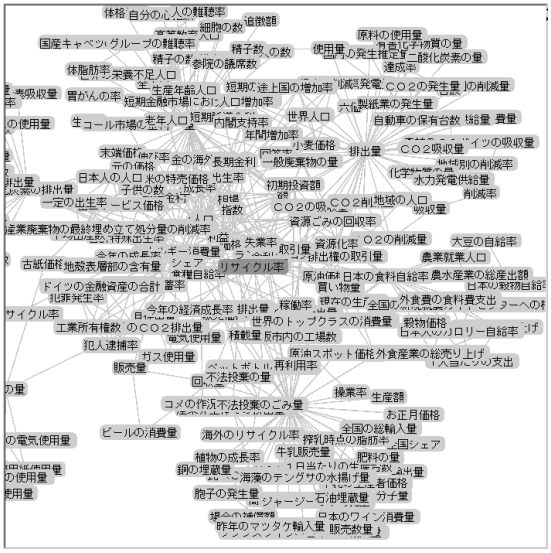

図 2: 共起ネットワーク構造の例

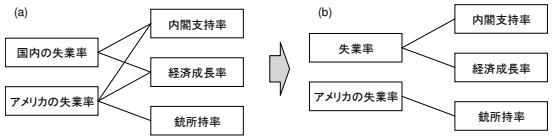

図 3: 共通基底表現とネットワーク構造によるクラスタリング

4. 実験方法

本研究では、テキストコーパスとして国立情報学研究所主催で実施されている評価型ワークショップ NTCIR-6 におけるパイロットタスク「動向情報の要約と可視化に関するワークショップ*1」で提供されている MuST コーパスを利用した。MuST コーパスは、毎日新聞の 1998 年 1 月〜1999 年 12 月の 2 年分の新聞記事（約 22 万記事）から構成されており、内閣支持率やパソコン出荷台数など 27 トピックについて統計量表現とその数値表現にタグ付けが行われている [加藤 04]。

初期の統計量表現の集合として、MuST のタグ付コーパスに登録されている 86 単語を用いた。また、suffix パタンを用いた抽出によって、統計量表現数を 100 倍の 8600 語に増殖した。これら 8600 語の統計量表現について、2 年分の新聞記事における共起頻度を計数し、閾値 $\theta = 1$ 以上の統計量表現間をエッジで結んで共起ネットワークを得た。ここで、閾値 θ を 1 以上としたのは、2 年分の新聞記事における統計量表現同士の共起頻度は大半が 1 回しかなかったからである。コーパスの量を増やし、高い共起頻度を持つ統計量表現同士の関係を使うことは今後の課題である。

さらに予備実験で、ノードの次数制限のパラメータ ω を変化させた場合に得られる共起ネットワークの最大クラスタサイズを調査した。その結果、$\omega < 10$ では最大クラスタサイズが急激に小さくなることが分かったため、本実験では $\omega = 10$ とした。最終的に簡略化された共起ネットワークの表示には、オープンソースの情報可視化ツールである prefuse*2 を用いた。

*1 http://must.c.u-tokyo.ac.jp/
*2 http://prefuse.org/

5. 結果および考察

実験の結果、簡略化されたネットワーク構造の上で、個別のトピックに対応したクリーク同士の関係を観察できるようになった。本節では、得られたネットワーク構造を紹介するとともに、提案手法の効果について考察を行う。

5.1 共起ネットワーク簡略化

図 4 に、簡略化された共起ネットワークの例を示す。図 4 は図 2 と同じく、統計量表現「リサイクル率」から 2 ホップ先までを表示している。一見して分かるとおり、ネットワークの密度が下がっており、ノード間の関係を観察しやすくなっている。図 4 では、統計量表現「リサイクル率」の周辺には「ごみの量」や「二酸化炭素の排出量」等、環境問題に関する統計量表現が表示されている。

図 5 に、簡略化された共起ネットワークの大域図として、統計量表現「リサイクル率」から 7 ホップ先までを表示させた場合の結果を示す。図 5 を見ると、少数のノードから構成されたクリークがハブ型のノードによって連結されていることが分かる。クリークを構成するノードの詳細を観察すると、一つのクリークが一つの新聞記事から抽出されていることが多く、特定のトピックに関する統計量表現の集合となっていた。これらのクリークの間に必ずしも因果関係があるわけではないが、順にたどっていくことによって関連するトピックを把握することができる。

例えば、図 5 において「リサイクル率」から始まって左上に走っている (A) のルートでは、まず、「排気ガス」に関連するクリークが存在し、次に環境に影響を与える化学物質の排出量に関するクリークを観察することができる。ここで一度「受精率」に関するトピックに話題が切り替わり、その後、「手術数」のような医療関係の統計量表現のクリークが存在し、最後に精巣がんに関する統計量表現のクリークに行き着いている。図 5 の (B) のルートでも、(A) と類似したトピックの変遷をたどっている。コーパスの新聞記事が書かれた 1998〜1999 年当時は、環境ホルモンやダイオキシンなどの毒性が騒がれいたため、このような関連性のルートができたのだと考えられる。

また、図 5 のルート (C) では、リサイクル率から消費性向のトピックに遷移し、米の消費量から食料自給率へ、さらには農業に関するトピックへたどり着いている。さらに、ルート (D) では、リサイクル率から失業率・雇用関係をトピックを経て経済成長率、金利、マンション価格と、経済系のトピックへの変遷が見られる。最後に、ルート (E) では、リサイクル率から資源回収に関するトピックに遷移したあと、ビールの売上げやパソコンの売上げに関するトピックへの飛躍が見られ、最終的には市場占有率のようなマーケティング系のトピックにたどり着いた。これらは、必ずしも論理的な関連性が説明付けられるものではないが、世の中の動きを統計量ベースの視点から見た場合の関連性としてとらえることができ、興味深い。

5.2 統計量表現クラスタリング

統計量表現のクラスタリングを行った結果、実際に一般化された統計量の例を表 1 に示す。多くの場合は共通の複合名詞の suffix が共通化して新しいノードになっているが、「温室効果ガスの総排出量」や「野菜の価格」のように、複数の複合名詞の間に助詞「の」が入っていたケースも見られた。この場合、共通 suffix とネットワーク構造に注目したクラスタリングが有効に作用したと言える。一方、あまり好ましくないと思われるクラスタ化の例としては、「ユダヤ人の数」が「人の数」と一般化されてしまったり、「事故時の状況」と「定年時の状況」がクラスタ化されて「時の状況」となってしまったりするケー

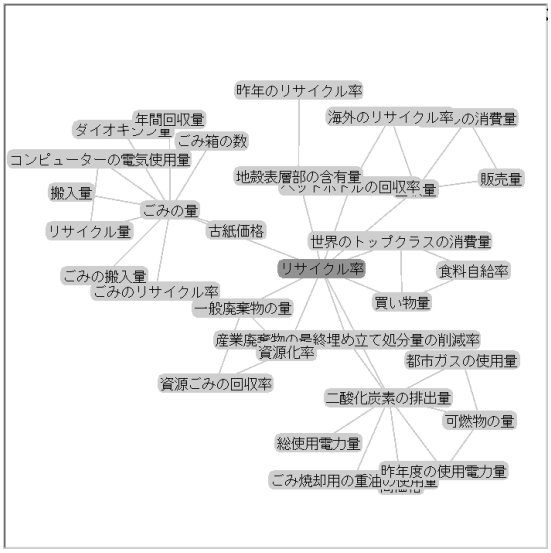

図 4: 簡略化された共起ネットワークの例

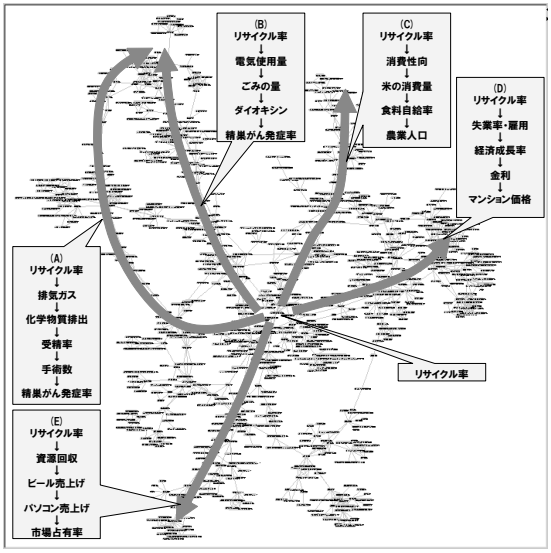

図 5: 簡略化された共起ネットワークの大域図

表 1: 統計量表現のクラスタリングの例

クラスタリング前	クラスタリング後
わが国の温室効果ガスの総排出量	温室効果ガスの総排出量
96年の温室効果ガスの総排出量	
日本の温室効果ガスの総排出量	
8月の新規住宅着工戸数	新規住宅着工戸数
今年度の新規住宅着工戸数	
参院選比例代表での自民党の得票率	自民党の得票率
比例代表の自民党の得票率	
12月のパソコン販売台数	パソコン販売台数
秋葉原の電気街のパソコン販売台数	
埼玉県所沢市の野菜の価格	野菜の価格
すべての野菜の価格	

スが見られた。これらのケースは、「複合名詞の途中ではクラスタ化しない」というルールを付け加えることによって対応が可能と考えている。

6. おわりに

本稿では、グローバルダイナミクス可視化のためのネットワーク構造簡略化手法として、ノードの次数制限とネットワーク構造を考慮した統計量表現のクラスタリング方法を提案した。実験では、提案手法を新聞記事データに適用し、簡略化されたネットワーク構造の上で、特定のトピックに関する統計量表現のクリークと、それを結びつけるハブ型の統計量表現が存在することが分かった。また、クリーク同士の関係を観察することによって、統計量ベースでのトピックの関連性を確認できた。今後は、コーパスの大規模化や因果関係の方向性に着目した抽出・可視化手法に取り組んでいきたい。

参考文献

[乾 05] 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.

[加藤 04] 加藤恒明, 松下光範, 平尾努: 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会研究報告 2004-NL-164, 情報処理学会, 2004.

[加藤 07] 加藤恒明, 松下光範, 神門典子: 動向情報の要約・可視化から情報編纂へ, 第 21 回人工知能学会全国大会論文集, 2H5-11, 人工知能学会, 2007.

[斉藤 98] 斉藤公一, 迫田昭人, 中江富人, 岩井禎広, 田村直良, 中川裕志: 数値情報をキーとした新聞記事からの情報抽出, 情報処理学会研究報告 1998-NL-125, 情報処理学会, 1998.

[齋藤 07] 齋藤悠, 河合英紀, 土田正明, 水口弘紀, 久寿居大: 新聞記事コーパスからの統計量表現自動抽出と共起関係ネットワーク構築, 動向情報の要約と可視化に関するワークショップ第 2 回成果進捗報告会予稿集, 2007.

[Saito 07] Haruka Saito, Hideki Kawai, Masaaki Tsuchida, Hironori Mizuguchi, Dai Kusui: Extraction of Statistical Terms and Co-occurrence Networks from, Proceedings of NTCIR-6 Workshop Meeting on Evaluation of Information Access Technologies, 2007.

[佐藤 06] 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.

[佐藤 99] 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, 1999.

[藤畑 01] 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, 情報処理学会研究報告 2001-NL-145, 情報処理学会, 2001.

[森 07] 森辰則, 藤岡篤史, 村田一郎: 動向情報編纂のためのテキストからの統計量の自動抽出, 第 21 回人工知能学会全国大会論文集, 3H9-4, 人工知能学会, 2007.