Vol.2009-DBS-149 No.10
2009/11/20

SVO 構造を用いた因果関係ネットワーク構築手法について

石 井 裕 志[†1]　馬　　強[†1]　吉 川 正 俊[†1]

人々のより深いニュース理解を支援するシステム作成のために，事象間の因果関係をネットワーク構造を用いて表現する手法を提案する．我々はこれまで，原因事象を始点ノード，結果事象を終点ノードとしたエッジラベル付き有向グラフとして因果関係ネットワークを表す TEC モデルを提案してきた．本稿では，日本語文法の SVO 構造に着目し，ノードの保持するキーワードのうち因果関係を含む文節から得られるキーワード（事象キーワード）について改良し，因果関係を含む文節から得る主語 (Subject)，動詞 (Verb)，目的語 (Object) の 3 属性を事象キーワードとする手法を提案する．因果関係ネットワークを構築するために類似した事象を表すノード対をマージするが，ノード間の事象キーワードを属性ごとに抽象概念レベルで比較することで，類似した事象のノード対を発見する．また，記事タイトルに含まれる語を利用してトピックの類似する記事集合間だけでマージ計算を行い，計算量を減らす手法を提案する．また，実際の記事から因果関係を抽出してマージを行い，トピックの類似する記事間集合だけでマージ計算を行っても精度や再現率が低下しないことを確かめた．

Causal Network Construction Using SVO Structure

Hiroshi Ishii,[†1] Qiang Ma[†1]
and Masatoshi Yoshikawa[†1]

In this paper, we propose a novel Topic-Event Causal relation model (TEC model) and describe a method to construct a Causal Network in a TEC model to support understanding of news. In the TEC model, causal relations are represented by an edge-labeled directed graph. A source vertex represents the cause of an event, and a destination vertex represents the result of that event. In the model, each vertex includes two types of keywords: topic keywords, which describe topics, and event keywords, which describe events. Using the SVO structure of Japanese, we compose event keywords as three words(Subject, Verb, Object). If each concept of event keywords is equal between two vertcies, we merge the verticies, which represent the similar event. In the merging calculation, Using topic keyword decrease computational complexity. A preliminary experiment of vertices merging to assess the validity of the proposed method demonstrated its usefulness.

1. はじめに

テレビや新聞，Web 上などでニュースとして報道される事象には，事象の展開が速いものや，事象に対して他の事象が複雑に絡み合っているものがある．そのような事象を理解する時には，その事象の起こった前後関係（背景知識）を知らなければ，ニュースを見聞きしても深く理解できない場合や，正しく理解できない場合がある．しかし，利用者が一つ一つの記事を検索してニュースを読み，背景知識を得るには多くの労力が必要であり，また背景知識の見落としなども発生する．よって，利用者に対してニュースの背景知識を提供し，理解支援を行うシステムが必要である．

利用者がニュースの背景知識を得るには，事象が起こったという結果だけでなく，なぜ起こったのかという原因も合わせた事象の因果関係を提示することが有効である．ニュース記事やその他の文書から因果関係を抽出する従来研究[2)–8)] では，因果関係を記事単位で抽出していた．しかし，記事単位の抽出では，得られる因果関係は断片的なものに留まり，ニュースの背景知識を理解するには十分な情報であるとは言えない．

因果関係が背景知識となるには，事象の起こった大元の理由や事象によって波及した結果を辿ることができる因果関係のつながり（因果関係の連鎖）を表現できることが必要である．複数のニュースから抽出された因果関係を結合し，因果関係の連鎖を表現する因果関係ネットワークを構築する必要がある．因果関係ネットワークとは，因果関係における事象やその原因をノードとし，原因事象を表すノードから結果として起こった事象を表すノードへ有効枝を張ったグラフである．因果関係ネットワークは，ある因果関係における結果ノードとある因果関係における原因ノードが同じ事象を表す時，このノード同士をマージすることで因果関係の連鎖を表現できる．このため，どのノードとノードが同じ事象を表したものか判断する必要がある．しかし，文書集合から因果関係ネットワークを構築する手法[4),6),7)] では，因果関係の記述されている文節だけからキーワードを抽出していた．このため，情報の不足や書き手の違いによる表記の揺れが起こり，キーワードの類似度を計算をして類似事象ノードを判断するのは難しく，精度がよくなかった．

†1 京都大学情報学研究科社会情報学専攻 〒 606-8501 京都市左京区吉田本町
Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo, Kyoto 606-8501
ishii@db.soc.i.kyoto-u.ac.jp, {qiang,yoshikawa}@i.kyoto-u.ac.jp

これに対し，我々が提案した Topic-Event Causal relation model(TEC モデル)[10] では，次の二種類のキーワード集合を用いる．

事象キーワード： 因果関係を含む文節から抽出される語

トピックキーワード： 記事のタイトルから抽出される語

トピックキーワードによってノードにトピック情報が付加され，全く違うトピック間のノードを間違って類似事象ノードと判断することを防ぎ，類似事象ノードの判断の精度を向上させる．

TEC モデルも，マージするノードの発見にはキーワードの類似度の計算が用いられたが，この方法では，マージをするノードの順番が異なるとネットワークの結果が異なるという問題が存在した．ニュースは日々更新され追加されるので，マージ計算はノードの出現した順（ノードの表す事象の報道された順）に沿って行われる．このため，報道された事象の順番によって，同じ内容のノードでも構築されるネットワークが異なってしまう．そこで本稿では，日本語の SVO 構造に着目し，事象キーワードとして主語，述語，動詞の三属性を設定する改良手法を提案する．この方法では，事象ノード間で各属性のキーワードが一致する場合にノードをマージしていけば良い．語の一致判定にはシソーラスを用いて記述の差異を吸収する．この手法では，三属性のキーワードが一致する時にだけマージを行うので，マージの順序の違いによってネットワークの結果が変わる問題は起こらない．また，TEC モデルでは，ネットワークのエッジに重要度を付加することで，因果関係の重要度を表現し，構築した因果関係ネットワークをリダクションも行う．

本研究の因果関係の構築システムでは，記事集合の「手がかり表現 (『を背景に』や『ため､』など)」を含む文から因果関係を抽出し，事象ノードのマージ，ネットワークのリダクションを行って因果関係ネットワークを構築する．ニュースは日々新しいものが追加されるので，１日単位でネットワークを更新する．まず，1 日分の因果関係ネットワークを構築する．そして１日分の因果関係ネットワークとそれ以前に構築した因果関係ネットワークとを合わせてマージ，リダクションを行い，因果関係ネットワークを更新する．

本稿では，1 日分のネットワークの構築では同じトピック同士の記事集合間だけでマージをして，複数日間でのネットワークの構築ではトピックキーワードを用いて，類似したトピック同士の記事集合間だけでマージを行う方法を提案する．これにより TEC モデルにおけるマージ計算の計算量を削減する．

以下，2 節で関連研究について説明し，3 節で TEC モデルについて，4 節で TEC モデルによる因果関係ネットワークの構築について述べる．5 節で事象ノードのマージの実験について述べ，最後に 6 節で結論について述べる．

2. 関 連 研 究

文書からの因果関係の抽出手法，因果関係ネットワークの構築方法の関連研究について説明する．

2.1 因果関係抽出手法

文書から因果関係を自動抽出する手法として，乾らの接続標識「ため」を用いる手法[5]，格フレームを用いる手法[4),6),7]，坂地らの手がかり表現と因果関係の構文パターンを用いる手法[3),8)] などが提案されている．乾らの手法[5] は，必然性の高い因果関係が成り立つ接続詞として「ため」に注目し,「ため」が使われやすい複文のみから因果関係を抽出している．

佐藤・笠原らの手法[6),7)] では，文が 1 文中に複数の格フレームを持ち，さらに接続関係のある場合にだけ因果関係を抽出し，そこからキーワードを抽出する．佐藤・堀田[4] は，Web 文書から「手がかり標識」（坂地らの手法[8] での「手がかり表現」に相当」）を含む文から因果関係がを抽出し，佐藤・笠原らの方法を元に原因文節（単文）と結果文節（単文）から重要と思われる単語を抽出する．Girju は，WordNet を使用して主語，目的語にあたる単語を概念ごとに抽象化した文書から，因果関係を含みやすい主語と動詞，主語と目的語のパターンを発見することによって因果関係を抽出する[2)]．手法[4)–7)] は因果関係文の抽出を複文・重文のみでしか行えなかったのに対し，坂地らの手法[3),8)] では，複文・重文に限らず「手がかり表現」を含む文全体を対象にして因果関係文節を取得できる．我々は坂地らの手法を利用している．

2.2 因果関係ネットワークの構築

上記で述べた因果関係の抽出法を使用して因果関係ネットワークを構築する手法として，佐藤・笠原ら[6),7)] や佐藤・堀田ら[4] の手法が提案されている．Feng ら[1] は，event threading 手法を用いてニュースをイベントごとに区切り，イベントをクラスタリングすることによって関連するイベント同士をリンクで結んだネットワークを作成した．佐藤・笠原らの手法は，人間が常識的に持っている因果関係知識をデータベース化することを目的として，一般の文書から因果関係ネットワークを構築する．佐藤・堀田の手法は，Web 上の文書から因果関係ネットワークを構築する．両手法ともに，文書の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のキーワードとしている．係受け解析と格フレームを用いて単語が重要かどうかを判断し，単語は重要 (事象データに含める) か冗長 (事象データに含めない) かに分けられる．二つの手法のいずれも，因果関係を表現している文節しかキーワードの対象に

ならないことと，冗長とされた単語は全く考慮されないため，原因・結果の一つの事象ノードは数語の単語からのみ構成される．この状態ではキーワードが不足しており，

- 事象データが何を意味しているのか分からない
- 事象データから実際に起きた事象を特定できない

ということが生じる．例えば，佐藤・堀田[4] の論文では，キーワードとして「慣例」のみを持つノードが取得された例が紹介されている．「慣例」だけでは，元の文が何について述べているのか分からず，また他に「慣例」のみを持つノードが取得されると，2 ノード間の類似度は高くなるが，この二つのノードが類似した内容を意味しているとはいえない．これを解決するために，本研究ではノードの持つキーワードを拡張する．

佐藤・堀田の手法はノードのマージは行っておらず，ノード間の類似度の高さをノード間の距離の近さとして表現している．しかし，類似事象を示すノードが多い場合には因果関係のつながりを理解しづらい．本研究では，キーワードの言葉の揺らぎをシソーラスを用いて同概念をまとめて，ノード同士の持つキーワードが完全一致する時にマージすることによって因果関係のつながりを表現する．

佐藤・堀田，佐藤・笠原らの両手法ともに生成したネットワークを整理していないが，ネットワークが複雑であると利用者が理解できない．本研究では，エッジの重要度を用いた因果関係のリダクションによりネットワークの整理を逐次行いながら，ネットワークの構築を行う．

3. TEC モデル

我々は，因果関係ネットワークを構築するために，TEC モデル (Topic-Event Causal relation model) について研究を行っている．‘T’ は事象ノードが保持するキーワード「トピックキーワード (topic keyword)」，‘E’ は事象ノードが保持するキーワード「事象キーワード (event keyword)」の頭文字に由来している．

3.1 類似度を用いた TEC モデルの問題点

我々が以前提案した手法では，ノードを構成するキーワードとして因果関係を含む文節と記事のタイトルから単語を抽出し，各々のキーワードに対し重要度を設定した．重要度は各キーワードが記事内で使われる頻度とした．4 節で述べるノードのマージ計算では，空間ベクトル法を用いてキーワードとキーワードの重要度からノード間の類似度を計算していた．マージされた後のノードは，マージ前のノードのキーワードの和集合で，キーワードの重要度は元のキーワードの平均値を用いた．この方法では，マージの順序の違いによってマージ結果が異なるという問題が存在した．**図 1** を用いて説明する．ここでは類似度が閾値 0.70

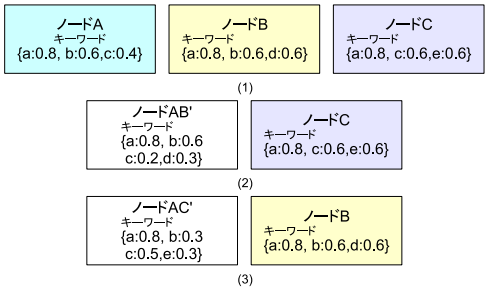

図 1 マージの順序によってネットワークが異なる例
Fig. 1 Example of merging order issue

以上の時にマージする．図 1(1) の A，B，C の 3 ノードがある時に，まずノード A と B からマージ計算すると A と B の類似度は 0.79 なのでマージされ，図 1(2) のようになる．図 1(2) の AB’ と C をマージ計算すると AB’ と C の類似度は 0.61 でマージされない．次に図 1(1) から，まずノード A と C からマージ計算すると A と C の類似度は 0.70 なのでマージされ，図 1(3) のようになる．図 1(3) の AC’ と B をマージ計算すると AC’ と B の類似度は 0.67 でマージされない．このようにマージするノードの順序によって，マージした後のネットワークが異なってしまう．実際，ニュース記事は日々更新されるので，因果関係ネットワークの構築は時系列的に処理を行う．このため，ノードの出現順によってネットワークが異なってしまう．ノード A，B が先に報道されればノード AB’ とノード C のネットワークになり，ノード A，C が先に報道されればノード AC’ とノード B のネットワークになる．このように，同じノード集合でも，ノードの出現する順序の違いによってネットワークの結果が異なってしまう．SVO 構造を用いた TEC モデルでは，因果関係を含む文節から主語，動詞，目的語の三つのキーワードの取得し，4.2.1 の手法によりそれぞれのキーワードが全て一致した時にマージを行う．これにより，マージの順序に依らずに同じネットワークを得ることができる．

3.2 SVO 構造を用いた TEC モデル

本稿では，日本語の SVO 構造を用いて事象キーワードを抽出する方法を提案する．まず，因果関係とは，二つの事柄に対して一方が原因で他方が結果であるような関係があることを

いう．TEC モデルでは，原因部，結果部それぞれに事象ノードを作成して，始点を原因の事象ノード，終点を結果の事象ノードとする．また，エッジは因果関係の重要度を表すラベル (エッジの重要度と呼ぶ) を保持し，因果関係ネットワークをエッジラベル付き有向グラフで表現する．エッジラベル付き有向グラフはグラフ G のノード集合 V，エッジ集合 E，エッジのラベル h に対して，写像 f が式 (1) を満たすように存在し，$G := (V, E, f, h)$ で表わされる．

$$f : E \to V \times V \tag{1}$$

TEC モデルにおける因果関係の表現例を**図 2** に示す．

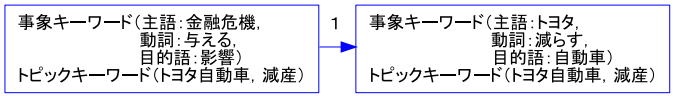

図 2 SVO 構造を用いた TEC モデルにおける因果関係の表現例
Fig. 2 Causal relation in TEC model using SVO structure

グラフの各エッジ e は，始点ノード (原因ノード)v_{s_e}，終点ノード (結果ノード)v_{t_e}，エッジの重要度 h_e に対し，式 (2) で表す．

$$e := (v_{s_e}, v_{t_e}, h_e) \tag{2}$$

エッジの重要度は，因果関係の頻度によって決定する[⋆1]．因果関係の因果関係が抽出された時のエッジの重要度 h の初期値は 1 であり，頻度が高い場合には値が大きくなる．エッジの重要度の計算方法は 4.3 節で述べる．

TEC モデルでは各事象ノードに，因果関係を表す事象キーワード集合と記事のトピックを表すトピックキーワード集合を保持する．事象キーワードは主語を表すキーワード k_{e_s}，動詞を表すキーワード k_{e_v}，目的語を表すキーワード k_{e_o} の 3 種類で構成される．トピックキーワードは関連記事のタイトルから抽出されるた単語で，頻度の高かった 2 個のキーワード k_{t_1}，k_{t_2} で構成される．よって，事象ノード v は，

$$v := \{(k_{e_s}, k_{e_v}, k_{e_o}), (k_{t_1}, k_{t_2})\} \tag{3}$$

として表現できる．次に，事象キーワード k_e，トピックキーワード k_t について述べる．

(1) 事象キーワード k_e

事象キーワードは取得した因果関係を表現するキーワードである．因果関係を含むとして取得した文節を日本語構文解析し，文節の主語，動詞，目的語を事象キーワードとする．

(2) トピックキーワード k_t ノードが持つ因果関係のトピックを表すキーワードである．元の記事の関連記事集合のタイトルに含まれる語から，頻度の高い単語 2 個をトピックキーワードとする．

4. SVO 構造を用いた TEC モデルによる因果関係ネットワークの構築

SVO 構造を用いた TEC モデルにおける因果関係ネットワークの構築手法について述べる．記事から TEC モデルに沿うように因果関係を抽出し，そこから類似ノードの結合や不要因果関係の削除を行い，因果関係ネットワークとして構築する．因果関係の抽出は，手がかり表現を用いて文書から因果関係を抽出し，事象ノードと因果関係のエッジを作成する．ノードのマージは，事象キーワードを用いて類似事象を示すノードを結合し，ネットワークのリダクションは，エッジの重要度を利用してネットワークを簡略化する．

4.1 因果関係の抽出

ニュース記事から因果関係を抽出し，事象ノードを作成するまでの手順を以下に示す．

(1) 記事から因果関係を含んだ文節の抽出
(2) 事象ノード，エッジの作成
　　(a) 因果関係の事象キーワードの抽出
　　(b) 記事のトピックキーワードの抽出

図 3 に記事から因果関係を抽出し，事象ノードを作成する例を示す．

4.1.1 記事からの因果関係文節の抽出

本研究では，文書から因果関係を抽出するのに坂地ら[7] の手法を用いた．まず,「を背景に」や「ため､」などの因果関係の存在する文を示す「手がかり表現」を含む文を記事から抽出し，因果関係を含む文と判断する．抽出された文を cabocha[9] によって係受け解析し，次の 4 種の構文パターンに分類する．

Pattern A: 結果表現の主部と述部が存在する場合
　「＜原因表現＞のため，＜結果表現主部＞が＜結果表現述部＞した.」
Pattern B: 手がかり表現の直後に結果表現が存在する場合
　「＜原因表現＞のため，＜結果表現＞した.」
Pattern C: 結果表現が抽出文の前文である場合

⋆1 将来的に，サポート度と確信度を考慮したエッジの重要度の決定手法を検討していく．

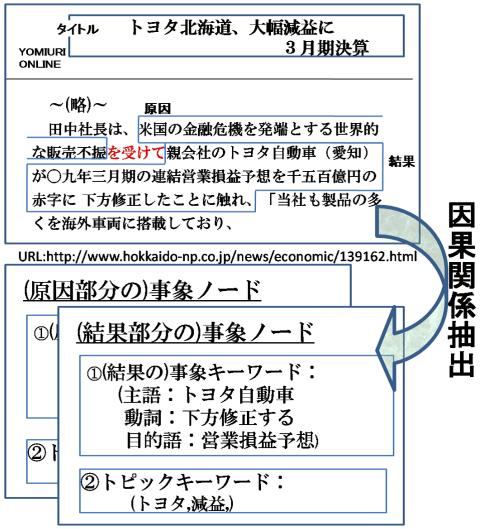

図 3 記事から事象ノードを抽出する例
Fig. 3 Extracting event vertices

「＜結果表現＞した．＜原因表現＞のためだ.」

Pattern D: 結果表現が根拠表現より前に出現する場合

「＜結果表現＞は，＜原因表現＞のためだ」

構文パターン別の文法的特徴により，元の文書から根拠表現文節と結果表現文節を抽出する．また，主な手がかり表現を**図 4** に示す．

手がかり表現
「を背景に」,「を挙げる」,「ため、」,「に伴う」,「に加え、」,「ためだ。」,「により」,「を受けて、」,「の効果が」,「の影響も」,「の影響が」,「によって」,「から、」

図 4 手がかり表現
Fig. 4 Clue phrases

図 3 の記事の例では，手がかり表現「を受けて」を検索することによって因果関係を含む文が抽出され，係受け解析によって Pattern A として判定され，原因表現「米国の金融危機を発端とする世界的な販売不振」と結果表現「親会社のトヨタ自動車 (愛知) が○九年三月期の連結営業損益予想を千五百億円の赤字に下方修正したことに触れ」が抽出される．

4.1.2 事象ノードの作成

得られた因果関係のそれぞれについて原因の事象ノードと結果の事象ノード，それを結ぶエッジを作成する．この時，エッジの重要度は 1 とする．そして，事象ノードごとにキーワードを抽出し挿入していく．以降，ここで因果関係の文節を抽出した元の記事のことを注目記事と呼ぶ．

(a) 事象キーワード k_e の抽出

事象ノードの事象キーワード k_e となる語を抽出する．因果関係文節の抽出で抽出された文節を Syncha[⋆1] を用いて述語項構造を解析し，主語 k_{e_s}，動詞 k_{e_v}，目的語 k_{e_o} を事象キーワード k_e とする．

(b) トピックキーワード k_t の抽出

事象ノードのトピックキーワード k_t となる語を抽出する．トピックキーワードは，ニュース記事を抽出する時に注目記事の関連記事集合内の記事タイトルを集めて解析し，頻度の高い単語をトピックキーワードとする．具体的には，タイトルの部分を茶筌[⋆2] を用いて形態素解析し，名詞と未知語を抽出する．抽出された語を頻度に高い順に並べ，頻度の高い二つをトピックキーワード k_t とする．

図 3 の事象ノードの例では，抽出された結果文節から結果の事象キーワード「主語：トヨタ自動車，動詞：下方修正する，目的語：営業損益予想」が抽出される．また，記事タイトルからトピックキーワード「北海道」や「減益」が抽出され，頻度が高ければトピックキーワードとなる．

4.2 ノードのマージ

マージでは類似した事象ノードと事象ノードの結合を行う．4.1 節で一つ一つの因果関係の取得について述べたが，このマージによって複数の因果関係を結びつけ，因果関係の連鎖を表現することができる．**図 5** に示すマージの例では，事象キーワードが「トヨタ，減らす，　自動車」と「トヨタ，削減，車」の 2 ノードが，類似ノードとしてマージされ，因果関係の連鎖が取得できる．

⋆1 Syncha：http://sourceforge.jp/projects/syncha/
⋆2 茶筌：http://chasen-legacy.sourceforge.jp/

Vol.2009-DBS-149 No.10
2009/11/20

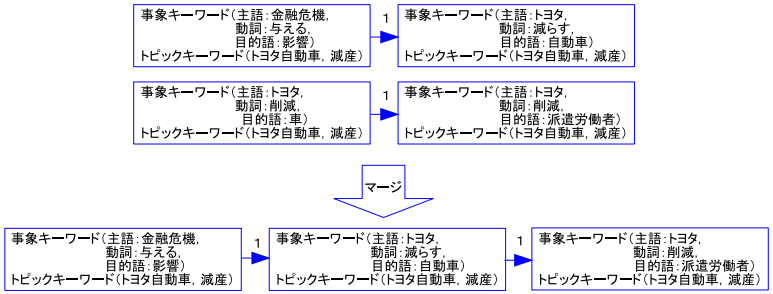

図 5 ノードのマージ
Fig. 5 Merging vertices

4.2.1 マージ手法

マージは，直接はエッジで結ばれていない相異なるノード間で行う．ノードのキーワードを比較し，主語，動詞，目的語のキーワードがそれぞれノード間で一致する時に，ノードのマージを実行する．主語，動詞，目的語のそれぞれのキーワードは，概念辞書を用いることによって表記の揺れを吸収し，同義の単語を同じものとして扱う．図 5 の例では,「減らす」と「削減」,「自動車」と「車」が同概念なので,「主語：トヨタ，動詞：減らす，　目的語：自動車」と「主語：トヨタ，動詞：削減する，目的語：車」のノードはキーワードが全て一致すると判断でき，マージを実行する．

マージ後のノードのキーワードは，マージされたノードペアのどちらかのキーワードを用いて構成される．図 5 では，マージされた図 5 上段の結果事象ノードのキーワードをマージ後のキーワードとする．

ノード v_a，v_b において，事象キーワードがすべて一致する時に，マージを実行し，v_{a+b} を作成，v_a，v_b を削除する．v_{a+b} のキーワードは，v_a もしくは v_b の事象キーワード，トピックキーワードとなる．

4.3 ネットワークのリダクション

リダクションでは重複するエッジをまとめ，まとめたエッジの重要度 (因果関係の重要度を表す) を求める．これを利用して重要度の低い因果関係の削除や利用者へ提示時する因果関係の判断を行う．エッジが重複しているとは，ノードのマージの結果，式 (4) のように二つのノード間に同じ方向のエッジが二つ以上存在していることをいう．

$$\left.\begin{aligned} e_{k_1} &= (v_1, v_2, h_{e_{k_1}}) \\ e_{k_2} &= (v_1, v_2, h_{e_{k_2}}) \\ &\vdots \\ e_{k_n} &= (v_1, v_2, h_{e_{k_n}}) \end{aligned}\right\} \tag{4}$$

リダクションでは，まず各ノード間で重複したエッジが存在するかどうかを調べる．存在する場合には，重複したエッジの重要度の和を重要度とするエッジを作成し，元のエッジを削除する．式 (4) の場合はエッジをまとめたエッジ $e_{k_1+k_2+\cdots+k_n}$ を式 (5) のように作成し，エッジ $e_{k_1}, e_{k_2}, \cdots, e_{k_n}$ を削除する．**図 6** にエッジがまとめられる例を示す．

$$e_{k_1+k_2+\cdots+k_n} = (v_1, v_2, h_{e_{k_1}} + h_{e_{k_2}} + \cdots + h_{e_{k_n}}) \tag{5}$$

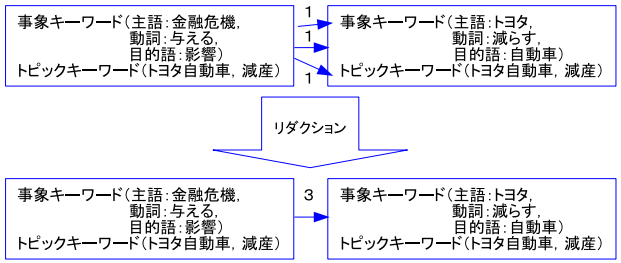

図 6 ノードのリダクション
Fig. 6 Reducing the causal network

4.4 ネットワークのインクリメンタル構築

ニュースは毎日新しいものが報道されるため，因果関係ネットワークを更新する必要がある．追加された記事で構築する因果関係ネットワークとそれまでに構築したネットワークを逐次構築することによってネットワークの更新をする．

その様子を**図 7** に示す．図 7(a) では，一日単位で新たに収集された記事から因果関係の抽出し ((1)，(2))，抽出された因果関係からマージ・リダクションを行い一日分の因果関係ネットワークを構築 ((3)，(4)) し，そして，一日分のネットワークとそれまでに構築したネットワークのマージ (図 7(b))，リダクション (図 7(c)) を行う．つまり，図 7 の (a)，(b)，

Vol.2009-DBS-149 No.10
2009/11/20

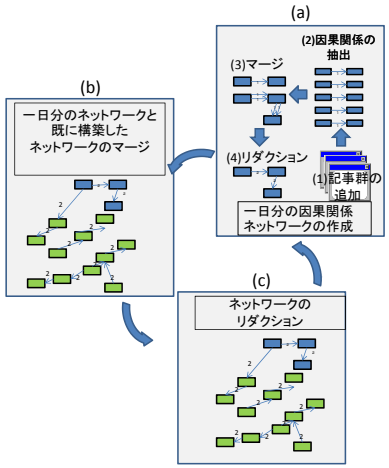

図 7 ネットワークのインクリメンタル構築
Fig. 7 Incremental construction of network

(c) を (a) → (b) → (c) → (a) →…と一日単位で繰り返し行うことによって，ネットワークを更新する． 本稿では，1 日単位のネットワーク作成をする時には同トピック内の記事集合 (関連記事として Web 上でまとめられていた記事集合) のみでマージ計算を行い，複数日のネットワーク同士を結合する時には，トピックキーワードの一致する記事集合間のみでマージ計算を行う．つまり，トピック a とトピック b のトピックキーワードが一致しない時には，そのトピック間ではマージ計算を行わない．この手法を用いることによって，マージ計算の計算量を削減することができる．全てのノード間でマージ計算を実行するとき，計算回数 N は次の式 (6) として計算できる．

$$N = \frac{(T \cdot a) \cdot (T \cdot a - 1)}{2} \tag{6}$$

ここで，T はトピックの数であり，a は 1 トピックあたりの記事数の平均である．類似したトピックのノード間のみでマージを計算をする時，計算回数 N' は次の式 (7) として計算できる．

$$N' = \frac{(T \cdot a) \cdot (s \cdot T \cdot a - 1)}{2} \tag{7}$$

ここで，s は 2 トピックが一致している平均確率である．一般に，s の値は十分に小さいので，N' は N よりも大幅に小さくなる．

5. 実　　験

提案手法を用いて，ノードをマージする精度と再現率，マージ計算量の削減量を評価する実験を行った．今回の実験では，SVO 構造を用いた TEC モデル手法ではなく，類似度を用いた TEC モデルの手法を使用して事象ノードを抽出し，ノードをマージした．SVO 構造を用いた TEC モデル手法の評価実験は今後行っていく．

5.1 実験環境

以下の 7 つのトピック，60 の日本語記事を用いて実験を行った．

- 「トヨタ自動車が減産する」ことについて 8 記事
- 「トヨタ自動車の労働組合が賃上げ要求をした」ことについて 5 記事
- 「トヨタ自動車の利益が減少した」ことについて 16 記事
- 「トヨタ自動車の北アメリカ工場が操業停止した」ことについて 4 記事
- 「アメリカの景気刺激策が大幅に修正される見通しである」ことについて 2 記事
- 「オバマ大統領が減税政策を行った」ことについて 10 記事
- 「米民主党が景気刺激策を発表した」ことについて 15 記事

これらは，2009 年 1 月 11 日から 20 日までの GoogleNews[⋆1]の経済記事から収集した記事である．

5.2 実験内容

まず，実験セットから因果関係の抽出をシステムによって行った．その結果，86 個の事象ノードを抽出し，そのうち 52 個が因果関係ネットワークとして正しく表現できた．ここで抽出した事象ノードを用いて，マージ計算の評価を行った．この 86 個の事象ノードにおいて，事象ノード同士の類似度を計算し，閾値 (threshold) を超える類似した事象ノード対をマージする．86 個の事象ノード中には，実際に類似していると判断できる事象ノード対は 107 個あった．類似していると判断できる事象ノード対全体を，全ての正解マージ (all correct merging) と呼ぶ．また，全てのの判断で類似しているといえる事象ノード対のうち，システムがマージできたものを正解マージ (correct merging) と呼ぶ．精度 (precision) は式 (8) のように定義する．

$$precision = \frac{number\ of\ correct\ merging}{number\ of\ system\ merging} \tag{8}$$

⋆1 GoogleNews:http://news.google.co.jp/

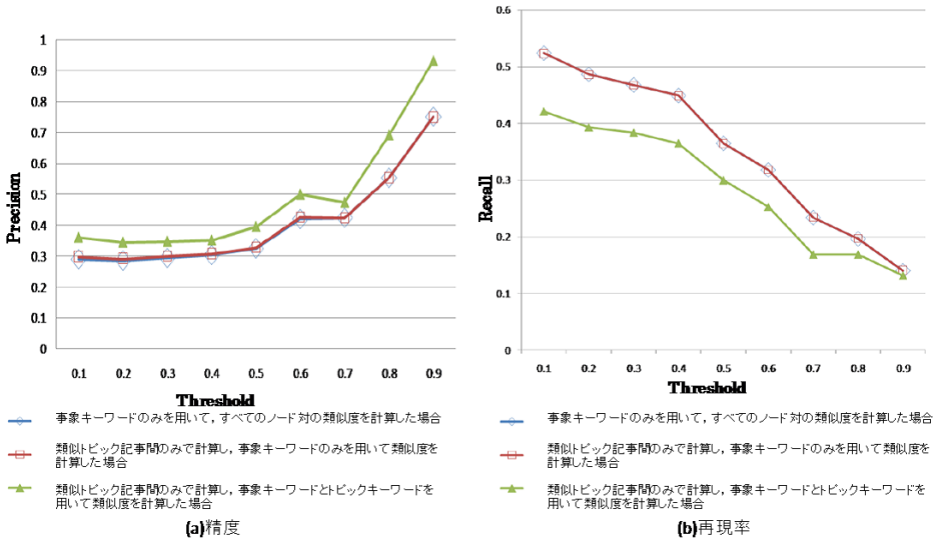

図 8 ノードをマージするシステムの精度と再現率
Fig. 8 Accuracy and recall of merging vertices

再現率 (recall) は次の式 (9) のように定義する．

$$recall = \frac{number\ of\ correct\ merging}{number\ of\ all\ correct\ merging} \tag{9}$$

図 8 に，それぞれの閾値における精度と再現率の実験結果を示す．この実験では，3 種類の類似度計算によるマージを行った．

- 事象キーワードだけを使用し，全てのノード対の類似度を計算する．
- トピックキーワードを用いてトピックの類似している記事間だけでノード対の類似度計算を行い，その時，事象キーワードを用いて類似度を計算する．
- トピックキーワードを用いてトピックの類似している記事間だけでノード対の類似度計算を行い，その時，事象キーワードとトピックキーワードを用いて類似度を計算する．

図 8 の結果から，全てのノード対の類似度を計算する場合と，トピックの類似している記事間だけでノード対の類似度を計算する場合では，ほとんど精度，再現率に差がみられないことが分かる．この結果から，マージ計算の始めにトピックキーワードを用いることによって，精度と再現率を低下させることなくマージ計算量を削減できることが分かった．

6. おわりに

本稿では，SVO 構造を用いた TEC モデルの提案を行い，モデルを用いた因果関係ネットワーク構築手法について提案した．また，類似したトピックの記事間だけでマージ計算する手法を提案した．実験により，類似したトピックの記事間だけでマージ計算する手法は，精度と再現率を悪化させずに，マージ計算量を減少させることができることが確かめられた．今後は，SVO 構造を用いた TEC モデルの因果関係ネットワーク構築の評価実験を行っていく．また，時間軸をモデルの中に取り入れることや，エッジの重要度に確信度や支持度の概念を導入し，因果関係ネットワークの構築手法を改善していく．

謝辞 本研究の一部は，科研費（20700084 と 20300042）の助成を受けたものである．

参考文献

1) Feng, A. and Allan, J.: Finding and linking incidents in news, *Proceedings of the 16th ACM Conference on information and knowledge management*, pp. 821–830 (2007).
2) Girju, R.: Automatic detection of causal relations for Question Answering, *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, Vol.12, pp.76–83 (2003).
3) Sakaji, H., Sekine, S. and Masuyama, S.: Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns, *Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management (PAKM2008)*, pp.111–122 (2008).
4) 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，pp.66–74 (2006).
5) 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).
6) 佐藤浩史，笠原　要，松澤和光：表層的因果知識ベースによる事象推移予測方式，情報処理学会 全国大会講演論文集， Vol.第 56 回平成 10 年前期, No.2, pp.251–252 (1998).
7) 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，電子情報通信学会技術研究報告. TL, 思考と言語， Vol.98, pp.27–32 (1999).
8) 坂地泰紀，竹内康介，関根　聡，増山　繁：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).
9) 工藤　拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌， Vol.43, No.6, pp.1834–1842 (2002).
10) 石井裕志，馬　強，吉川正俊：因果関係ネットワークの構築によるニュースの理解支援，*DEIM Forum 2009* (2009).