DBSJ Journal, Vol.8, No.1
June 2009

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

Bootstrap Extraction of Similar Relations from the Web by Term Co-occurrences

加藤 誠† 大島 裕明‡
小山 聡‡ 田中 克己‡

Makoto P. KATO Hiroaki OHSHIMA
Satoshi OYAMA Katsumi TANAKA

本論文では，ある関係を満たす語の組をシーズとして与えた場合，その関係と類似するような関係を満たす語の組を，自動的に Web から抽出する手法の提案を行う．我々は，Web 検索エンジンの検索結果として得られるテキストから，与えられた組と同じ関係にある語の組を抽出し，得られた結果を自動的に評価することにより，より精度が高く，多様な抽出手法を生成する．これを繰り返すことで，少ないシーズから正確かつ多くの語の組を取得することを目的とする．

In this paper, we propose a method to extract word-pairs from the Web, which meet relations similar to one met by input word-pairs. Our method first extracts word-pairs in relations similar to given one from Web search results, and then automatically generates a set of more accurate extraction patterns by self-evaluation. Repeating these processes, our method can extract a lot of word-pairs from a few seeds.

1. はじめに

近年，インターネットとWeb検索エンジンの普及により，様々な情報をWeb文書から取得することが可能となった．これらを利用して，Webに存在する多様で非構造な情報から自然言語処理やデータマイニングなどの技術を用いて，有用な情報を抽出することが盛んに行われている．これは辞書やオントロジーの自動構築などに利用され，特定の語の上位語，下位語，同位語などを取得する手法が数多く提案されている．これらの手法は，入力として1語，または，数語を与え，その語と特定の関係にある語をWebから取得することを目的としている．例えば，上位語の取得では，入力として「ペンギン」が与えられれば，出力として「鳥」が得られ，同位語の取得では，入力として「TOYOTA」が与えられれば，出力として「日産」などが得られる．しかし，これらの手法は語間の特定の関係に特化した関係抽出である．

そこで本論文では，任意の関係にある語のペアを入力として与え，それと類似した関係にある語のペアをWebから取得する手法を提案する．例えば，（こころ，夏目漱石），（人間失格，太宰治）などを与えた場合は，タイトルとその著者のペアを，（静岡，茶），（愛媛，みかん）などを与えた場合は，県名とその特産物のペアを出力する．このような，入力と出力によって，我々は多様な関係を一つの手法で取得することを可能とした．また，少ない入力のペアから類似関係にあるペアを大量に取得するために，我々はブートストラップ法[1,2]を用いた．これは近年，知識抽出において用いられており，入力として与えられた少数の語を元にして同位語などを大量に取得する手法である．ブートストラップ法では，入力から取得できた語を使って新しい抽出方法を生成することで，取得する語の数を徐々に増加させていくことが可能である．多くのブートストラップによる知識抽出手法では，言語パターンによる知識抽出が行われているが，本論文では，語の分布の差異を利用した手法を用いる．言語パターンによる抽出は精度が高いが，厳密な一致を求めるために，再現率は低いと考えられている．また，現実世界における多様な関係を考えた際に，言語パターンのみで表現できないものが考えられる．例えば，（京都，八ツ橋）という関係を考えた場合，「八ツ橋は京都の有名なお土産」といったフレーズだけでは表現は難しく，「八ツ橋」は「京都」において，土産である，伝統がある，代表的な菓子である，などといったように，「京都」と「八ツ橋」の間には複雑な関係があることがわかる．これらを考慮した場合，言語パターンによって類似関係抽出を行うことは，条件が厳しく柔軟でないと考えられる．

しかし，ブートストラップに適合率の低い手法を用いれば，得られる結果全体の適合率が大きく低下してしまう．そのため，得られた語のペアが適切であるかどうかを評価し，正解であると確信が得られたもののみを採用して精度を保つ必要がある．提案手法では，ある類似関係を満たすペア(x, y)のxとyはある特定のクラスに属すという性質，すなわち，得られる全てのx (y)は同位語となるという性質を利用して得られたペアの評価を行う．同位語であるかどうかを判定するために，Webを用いて語の共起度を測る指標の一つであるWebPMI[3]を用いた．

本論文における我々の貢献は以下の2点である．

- これまで提案してきた「語の共起を用いたWebからの関係抽出手法」を用いて，Webから大量の類似関係を抽出する手法を提案し，これを評価した．
- 類似関係の同位語制約に着目し，語の共起度を用いた同位語判定手法を提案した．

2. 関連研究

ブートストラップによって特定の関係にある語のペアを取得する研究は既に存在しているが，その多くが言語パターンを用いたものである．Brin[4]は本のタイトルとその著者のペアを少ないシーズから大量に取得するために，ブートストラップを用いている．これはWeb文書中にシーズが「prefix，著者名，middle，タイトル名，suffix」というパターンで出現している場合，パターンに適合する語列が同様の関係にあるペアであると仮定し，これを繰り返すことで目的とする語のペアを取得している．

Snowball[5]はBrinと同様の処理を行うシステムで，Brinのprefix，suffixなどのパターンに対して重みをつけ精度の向上を図っている．

張ら[6]はレコード抽出を行う文書をユーザの意図に適合した文書に限定することによって，レコード抽出におけ

† 学生会員　京都大学大学院情報学研究科修士課程
kato@dl.kuis.kyoto-u.ac.jp
‡ 正会員　京都大学大学院情報学研究科
{ohshima, oyama, tanaka}@dl.kuis.kyoto-u.ac.jp

る二つの問題を解決している．1つ目の問題は，テキスト処理にかかるコストが大きいことである．多くの文書からレコード抽出を行う場合，この問題は顕著となる．張らはユーザの意図に適合した文書を優先的に処理することによって，抽出効率を向上することに成功している．もう1つの問題は，得られた結果のすべてがユーザの意図に合致するとは限らないことである．これに対して，得られたレコードの正否判定のみでなく，ユーザの意図に合致しているかどうかも踏まえて評価することによって，ユーザの要求により適合したレコードを抽出している．

KnowItAll[1,2]は“such as”や“and other”などの言語パターンを用いて，入力されたシーズと同じクラスに属する語，すなわち，同位語をブートストラップを用いて取得するシステムである．KnowItAllでは，ブートストラップを用いて適合，不適合な語を大量に取得し，候補語がどのようなパターン内で現れる場合には適合であるかを学習している．

最近では，言語パターンのみではなくHTMLデータ構造を利用してブートストラップによる知識抽出が行われている[7,8]．これらは，Web文書中の表や箇条書きなどのHTMLデータ構造にも着目し，パターンにHTMLタグを含めることによって，知識抽出を行っている．

3. 語共起に基づくWebからの類似関係抽出

ブートストラップによる関係抽出の前に，語の共起を用いた関係抽出手法について述べる．我々は，入力として語A, B, Cを与え，AとBの関係に類似するようなCに対する語DをWebから発見する研究を行ってきた[9]．語Dは語Cとの組み合わせの内，語A, Bの関係と類似度が高い順にソートされ出力される．入力$q=(A, B, C)$と，語d_iに対するランク関数Rankは以下のようにして定義している．

$$\mathrm{Rank}(q, d_i) = \mathrm{Sim}(\mathrm{Relation}(A, B), \mathrm{Relation}(C, d_i)). \quad (1)$$

ただし，Relation(A, B)はAとBの間で成り立つような関係の集合を表している．実装では，語AとBを強く結びつけるような語tをWebから発見し，語tとCが出現するときにのみ有意に多く出現するような語が，条件を満たすような語Dである可能性が高いとし，入力A,Bと類似した関係をWebから取得している．以下，語A, Bを強く結びつける語をA, Bの関係接続語と表現する．

例として，入力として語A=静岡，B=お茶，C=愛媛が与えられた場合を考える．「静岡」と「お茶」の関係接続語としては，「産地」「直送」などといった語が抽出できる．関係接続語集合Tを発見した後，「静岡」，「お茶」と関係が類似するような「愛媛」に対する語Dを発見する．最終的に，Tの要素tと「愛媛」が出現するときのみ出現頻度が高くなるような語として「みかん」などが得られる．

入力A, B, Cに対して語Dを発見する手法の概要は以下の通りである．

3.1 関係接続語の発見

(1) 入力A及び入力Bに対して，Aを含みBを含まない文書を検索するクエリと，Bを含みAを含まない文書を検索するクエリでWeb検索を行い，検索結果のタイトルとスニペットを取得する．

(2) 入力Aと入力Bを含む文書を検索するクエリでWeb検索を行い，検索結果のタイトルとスニペットを取得する．

(3) タイトルとスニペットに対して形態素解析を行い，形態素ごとに分割する．また，ストップワードリストを用いて不要な語を除去する．

(4) 品詞が「名詞」であると推定された各語t_iに対して，“語Aを含み語Bを含まない文書と語AとBを共に含む文書において，語t_iの出現確率が等しい"という帰無仮説に対してχ^2検定を行う．同様に，語Bを含み語Aを含まない文書と語AとBを共に含む文書に対しても検定を行う．

(5) 有意水準αの検定において棄却された語t_iのうち，語A, Bが出現したときに出現確率が高くなるものを，入力された語AとBの関係接続語として採用し，これを関係接続語集合Tとする．

(1)～(5)までの手法は，語AとBが現れたときのみに有意に多く出現するような関係接続語集合TをWebから発見するものであり，以下，入力として語AとB, パラメータαが与えられたとき，関係接続語集合Tを出力するプロセスをRelationalConnectingTerm$_\alpha$(A, B) (RCT$_\alpha$(A, B)と略す)と表記する．

3.2 類似関係にある語の発見

(6) 関係接続語集合Tの全ての語t_iに対して，入力Cを含み語t_iを含まない文書を検索するクエリと，t_iを含みCを含まない文書を検索するクエリでWeb検索を行い，検索結果のタイトルとスニペットを取得する．

(7) 関係接続語集合Tの全ての語t_iに対して，入力Cと語t_iを含む文書を検索するクエリでWeb検索を行い，検索結果のタイトルとスニペットを取得する．

(8) タイトルとスニペットに対して形態素解析を行い，形態素ごとに分割する．また，ストップワードリストを用いて不必要な語を除去する．

(9) 品詞が「名詞」であると推定された各語d_jに対して，“語Cを含み語t_iを含まない文書と語Cとt_iを共に含む文書において，語d_jの出現確率が等しい”という帰無仮説に対してχ^2検定を行う．同様に，語t_iを含み語Cを含まない文書と語Cとt_iを共に含む文書に対しても検定を行う．両者の検定の結果，帰無仮説が発生する確率をそれぞれ$P_C(d_j)$, $P_{ti}(d_j)$とする．

(10) 有意水準βの検定において棄却された語d_jのうち，語C, t_iが出現したときに出現確率が高くなるものに対して，$P_C(d_j)$, $P_{ti}(d_j)$の積を$P_{C, ti}(d_j)$とする．

(11) 関係接続語集合Tの全ての語t_iに対する，$P_{C, ti}(d_j)$の全ての積を語d_jのスコアScore$_C(d_j)$とする．

(6)～(11)までの手法は関係接続語集合Tを利用した語Cに対する語Dの発見であり，以下，入力として関係接続語集合Tと語C, パラメータβが与えられたとき，順序付き語集合の上位k件を出力するプロセスをSimilarRelationTerm$_{\beta,k}$(C, T) (SRT$_{\beta,k}$(C, T)と略す)と表記する．

この手法は，語集合TによってAとBが共起するような文脈を表現し，その文脈下において語Cとよく共起する語が

AとBの関係と類似するような，Cに対するDであるという仮定を利用している．語の出現分布を用いることで，言語パターンよりも複雑な関係に対して適応しやすいと考えている．

4. ブートストラップ法による Web からの類似関係抽出

4.1 プロセス概要

3節で用いたプロセスRelationalConnectingTerm, SimilarRelationTermを用いて，少数のシーズ(seeds)から類似関係を抽出する手法を提案する．本論文では前ステップで得られた出力を入力として与え，再帰的操作によって徐々に解を増やしていく手法であるブートストラップ法を用いる．ブートストラップ法を用いる際には得られる出力の精度を高く保つことが必要である．前のステップで得られた出力にノイズが含まれている場合，次のプロセスではノイズが入力として与えられるため，連鎖的に精度が低下してしまう危険性がある．そのために，解の選択は慎重に行い，抽出手法に関しても精度のよいものだけを利用することが望ましい．そこで，提案手法ではWebPMIにより得られた解の候補及び関係接続語集合を評価し，確実に正解であると判断できる解のみを次回のプロセスにおける入力として与える．

提案手法で行うブートストラップの概要を図1に示す．また，図2に$Seeds$ = {(静岡，お茶), (愛媛，みかん)}を入力したときの動作例，及び，各変数の変化の過程を示す．

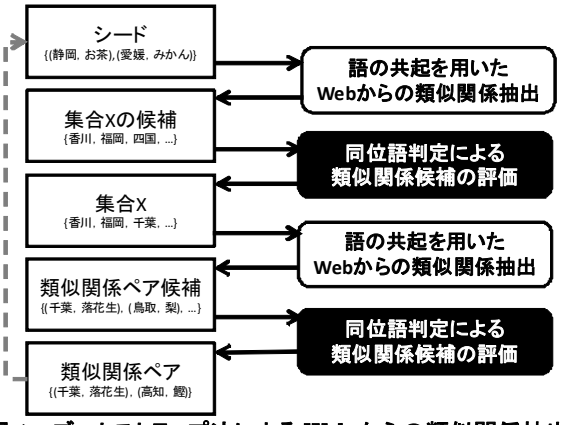

図 1　ブートストラップ法による Web からの類似関係抽出
Fig.1　Bootstrap Extraction of Similar Relations from the Web

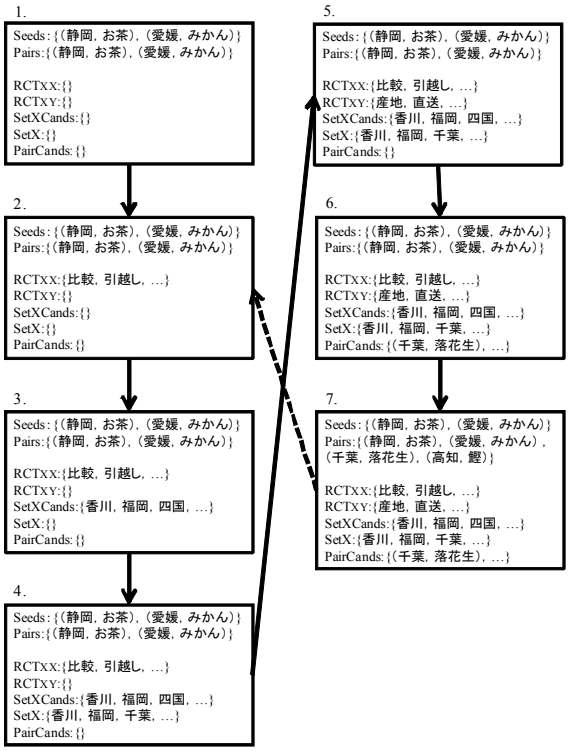

図 2　類似関係抽出の動作例
Fig.2　Example of Extraction of Similar Relations

最初に，$Seeds$を入力として与える．$Seeds$の定義は$Seeds \subseteq X \times Y$となる．すなわち，$Seeds$は集合$X$と集合$Y$の要素のペアになる．例えば，$Seeds$ = {(静岡，お茶), (愛媛，みかん)}であれば，集合Xは県名であり，集合Yが特産物であることが暗に与えられる．これは，$Pairs$の定義とも等しく，$p = (x, y) \in Pairs$のとき$p[X] = x$, $p[Y] = y$, また，$Pairs[X] = \{p[X] \mid p \in Pairs\}$, $Pairs[Y] = \{p[Y] \mid p \in Pairs\}$と定義する．初期操作として，与えられた$Seeds$は正解と推定されたペア集合$Pairs$に追加される．以上の操作が図2のステップ1にあたる．

$Pairs$が十分な量に達していなければ，以下のプロセス(2-7) を繰り返し行う．ただし，これ以降，プロセスで用いる$Pairs$の要素は評価値の高い上位n件のみに限定する．

まず，集合Xに含まれると予想される要素を，語の共起を用いたWebからの類似関係抽出手法によって取得する．$Pairs$ = {(静岡，お茶), (愛媛，みかん)}の例の場合，A=静岡，B=愛媛，C=静岡という入力を与えることによって，集合Xに含まれるような要素，すなわち，静岡や愛媛の同位語を取得することが出来る．ブートストラップによる類似関係抽出では，あらかじめ関係接続語集合RCT_{XX}を$Pair$に含まれるXの全ての組み合わせから求める．RCT_{XX}の部分集合から得られる上位k件のX要素を求め，これを$SetXCands$とする．$SetXCands$の各要素をWebPMIによって評価し，有効な関係接続語集合及び新しいX要素集合$SetX$を決定する．以上の操作が図2のステップ2から4 にあたる．

次に，得られた集合$SetX$の要素から，シーズと同じような関係にある集合Yの要素を発見する．$Pairs$ = {(静岡，お茶), (愛媛，みかん)}，$SetX$ = {香川，福岡，千葉，…}の例の場合，A=静岡，B=お茶，C=香川という入力を与える

ことによって，香川に対する集合Yの要素を発見する．$Pair$に含まれる全ての組から得られた関係接続語集合RCT_{XY}の部分集合から得られる上位k件のY要素を求め，これとX要素の組集合を$PairCands$とする．$PairCands$の各要素をWebPMIによって評価し，有効に働く関係接続語集合と$Seeds$と類似した関係を持つ組集合$Pairs$を得る．以上の操作が図2のステップ5から7にあたる．

以上がブートストラップによる類似関係発見手法の概要である．

4.2 抽出手法及びペア候補の評価

ブートストラップにより多様な抽出手法を生成してその精度を保つためには，抽出手法の評価が必要となる．提案手法における抽出手法とはすなわち，新たに語を発見するのに用いる関係接続語の種類を指している．また，語の共起を用いたWebからの類似関係抽出手法は精度の面で不十分であるため，ブートストラップ法で多くの類似関係を抽出するためには，正解と判断できたものだけを選択する必要がある．そこで，類似関係を抽出した後にペアの候補を評価し，以下の2つの仮定に基づいて抽出手法及びペア候補の評価値を求める．

- 良いペアを得られた抽出手法は良い抽出手法である．
- 良い抽出手法から得られたペアは良いペアである．

提案手法では，ある類似関係にあるペア(x, y)のxとyはある特定のクラスに属すという性質，すなわち，得られる全てのx (y)は同位語となるという性質を利用する．得られたペアを評価したあとに上の仮定に従い「良い」抽出手法を評価する．最終的には，「良い」抽出手法から得られた「良い」ペアが解として選択される．

ある同位語集合Tに対して語tがその同位語であるかどうかを判定するために，我々はWebを用いて語の共起度を測る指標の一つであるWebPMIを用いた．WebPMIは二語の意味的類似度を判定するためにBollegalaら[3]によって用いられており，その定義は以下のようになる．

$$\mathrm{WebPMI}(P, Q) = \begin{cases} 0 & \text{if } \mathrm{H}(P, Q) \le c \\ \log_2\left(\frac{\frac{\mathrm{H}(P \cap Q)}{N}}{\frac{\mathrm{H}(P)\mathrm{H}(Q)}{N\ N}}\right) & \text{otherwise} \end{cases} \quad (2)$$

ただし，$\mathrm{H}(x)$はクエリxで検索を行ったときのヒット件数であり，Nは検索エンジンの全インデックスページ数である．

語t'が語集合Tと同位関係にあるかを判定するために，このWebPMIをある語t'と語集合Tの各要素間で求めたときの平均を利用した．もし語t'が語集合Tの同位語であれば，この平均値が高いことが期待される．提案手法の中ではこれを利用し，シーズから得られる語集合$Seeds[Y] = \{p[Y] \mid p \in Seeds\}$の全要素とある語$t'$の共起度をWebPMIによって計算し，その共起度の平均を集合$Seeds[Y]$の要素間の平均で正規化した値がある程度大きければ，語t'は集合$Seeds[Y]$と同じクラスに属する，すなわち，同位語であると判定した．

ある同位語集合Tに対して語t'がその同位語であるかどうかを，WebPMIを用いて評価する関数$\mathrm{WE}(t', T)$を以下のように定義する．

$$\mathrm{WE}(t', T) = \frac{1}{\mathrm{InnerWE}(T)} \frac{1}{|T|} \sum_{t \in T} \mathrm{WebPMI}(t, t') \quad (3)$$

ただし，正規化を行うための同位語集合内でのWebPMI平均，$\mathrm{InnerWE}(T)$は以下の通りである．

$$\mathrm{InnerWE}(T) = \frac{1}{m} \sum_{t_1, t_2 \in T} \mathrm{WebPMI}(t_1, t_2) \quad (4)$$

式中のmは同位語集合の任意の2語の組み合わせ数であり，$m = {}_{|T|}\mathrm{C}_2$となる．

以上で定義された指標を用いて抽出手法及びペア候補の評価関数を定義する．抽出手法，すなわち，関係接続語集合Tが語集合Xに対して語集合$SetY = \cup_{x \in X} \mathrm{SRT}_{\beta,k}(x, T)$を得たとき，語集合$SetY$は$Seeds[Y]$ の同位語でなくてはならない．そこで抽出手法の評価関数$\mathrm{RCTE}(T, X)$には各$y \in SetY$と$Seeds[Y]$の同位語評価値の平均を採用する．

$$\mathrm{RCTE}(T, X) = \frac{1}{|SetY|} \sum_{y \in SetY} \mathrm{WE}(y, Seeds[Y]) \quad (5)$$

良いと判断された抽出手法の評価に基づいてペアの評価は行われる．そのため，ある閾値γ以上の評価値を得た抽出手法から得られたペアにのみ評価値を与える．与えられたペア$p = (x, y)$の評価関数$\mathrm{SRTE}_{\gamma}(y, T, X)$は以下のように定義される．

$$\mathrm{SRTE}_{\gamma}(y, T, X) = \begin{cases} \mathrm{WE}(y, Seeds[Y]) & \text{if } \mathrm{RCTE}(T, X) > \gamma \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

評価関数SRTEがある閾値δ以上ならば解として妥当であると判断し採用する．精度良く抽出するためには手法中で用いられるγ，δを適切に設定する必要がある．

5. 実験

ブートストラップによる類似関係抽出手法の精度を確認するため，我々は県名と特産物のペアを与え，その関係と類似したペアを抽出する実験を行った．使用したシーズは{(静岡，お茶), (愛媛，みかん)}の2組であり，類似関係検索に利用されるパラメータα及びβは事前実験より両方共に0.01に設定した．本実験の検索で取得するWeb文書数は100件であり，類似関係検索には得られた結果の上位5件を採用した ($k = 5$)．ブートストラップで用いられる評価値の閾値γ_1，δ_1 (集合Xの抽出で用いる)，γ_2，δ_2(集合Yの抽出で用いる)は，反復1回の事前試行によりそれぞれ0.9，0.9，0.9，1.2に設定した．また，反復試行に用いる$Pairs$の要素は評価値の高い上位n (= 5)件のみに限定する．

ブートストラップによる反復を5回行い，その結果に対して人手で評価を行った．5回目に得られた一意な解を全正解集合と仮定したときの適合率-再現率グラフを図3に示す．また，反復回数と得られた総正解数の増加の様子を図4に示す．

1回目の反復試行では適合率を90%弱に保ち，再現率を初期シーズの数十倍に増加させている．2回目の反復試行では適合率を65%程度まで低下させている一方，正解ペアを多く獲得することに成功している．3回目，5回目の試行では正解数を大きく増大させることはないものの適合率は60%台を保ち，4回目の試行においても多くの正解を得ていることがわかる．

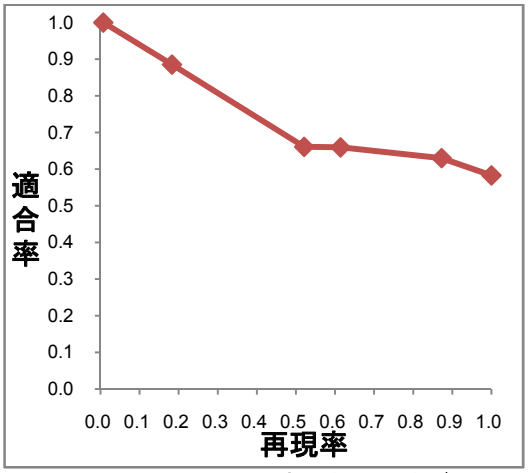

図 3　類似関係抽出の適合率-再現率グラフ
Fig.3　Precision-Recall Curve for Extraction of Similar Relations

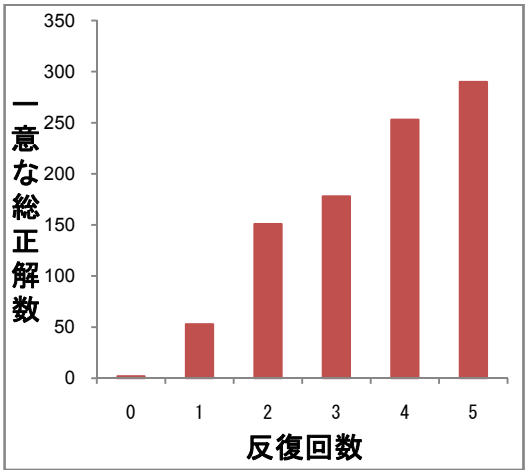

図 4　反復回数と一意な総正解数
Fig.4　Number of Iterations and Unique Relevant Pairs

2回目の試行以降に適合率を損なう要因はいくつか考えられる．第一に，1回目の反復にはシーズが用いられているが，2回目以降は新たに得られた解からシーズを生成するため，最初に入力したシーズとは徐々にその関係が異なってくるためであると考えられる．しかし，シーズを変えることを行わなければ，同じような解が多く得られ再現率の向上につながらない．そのため，反復試行のシーズに初期シーズを用いるか新しいシーズを用いるかは，適合率と再現率のトレードオフの関係に対応している．第二に，1回目の試行で適切であったパラメータが2回目以降は適切でなくなる可能性がある．提案手法ではパラメータが多く用いられるが，これらを正確に決定することは困難である．

また，2回目の試行において大きく適合率を低下させつつも，3回目においてその適合率を下げなかったのは，反復試行に用いる*Pairs*の数を評価値の高い上位n (= 5)件に限定しているため，信頼度の低い解が反復プロセスに影響しなかったためと考えられる．

表 1　類似関係抽出により得られた正解例
Table 1　Example of Relevant Similar Relations

X（県名）	Y（特産物）	X（県名）	Y（特産物）
愛媛	伊予柑	愛媛	温州みかん
岡山	桃	岡山	ピオーネ
沖縄県	ドラゴンフルーツ	沖縄県	パッションフルーツ
宮崎	マンゴー	岐阜	富有柿
京都府	宇治茶	宮崎	パパイヤ
熊本	馬刺し	宮崎県	すいか
広島	牡蠣	京都府	京野菜
香川県	ポンカン	熊本	晩白柚
高知	かつお	熊本	塩トマト
佐賀	ほのか	広島県	レモン
三重	松阪牛	高知	土佐文旦
山口	下関うに	高知県	生姜
山梨県	ぶどう	三重	伊勢茶
滋賀県	赤こんにゃく	山口	車えび
鹿児島	黒豚	山口	長門ゆず
静岡県	マスクメロン	滋賀	鮒寿司
大分	椎茸	鹿児島	さつま揚げ
長崎県	びわ	鹿児島	黒酢
長野	リンゴ	大分県	カボス
鳥取	らっきょう	長崎	ザボン
徳島	鳴門金時	鳥取	二十世紀梨
奈良	富有柿	島根県	いちじく
福岡県	明太子	徳島	すだち
兵庫県	いちご	福岡県	苺

表1にブートストラップによる類似関係発見で得られた正解の一部を示す．不適合であるペアの多くはY(特産物)の誤りによるもので，都道府県名の誤りは数例しか見られなかった．また，シーズとして{(静岡，お茶), (愛媛，みかん)}を指定したために，中部以西の地域のみしか都道府県名を得ることができなかった．これは，2個のペアを与えるだけでは正確には同位語の粒度，すなわち，どの上位語の下位語にあたる同位語であるかということを暗に与えることができなかったためであると考えられる．シーズとして，(青森，リンゴ)などを加えれば全ての都道府県名が得られると予想される．

また，シーズ数が2個のみであったために「みかん」とWebPMIが高いような柑橘系の特産物が多く得られるという結果になった．これも同様にシーズ数を適当な数に増やすことによって解決できる問題であると考えられる．

表2に各反復試行で用いた関係接続語集合（RCT_{XY})の評価値が高い上位3件を示す．1回目から3回目の試行では特産物関係を結びつけるような語集合が得られているが，2回目，3回目からは果物に限定されるようになり，4回目，5回目の試行では柑橘系の固有名詞が出現しているため，これがノイズとなって精度の悪化に関与していると考えられる．

表 2　各反復回の関係接続語集合（上位 3 件）
Table 2 Relational Connecting Terms for Each Iteration (Top 3)

	関係接続語集合	評価値 (RCTE)
1	直送	1.17
	産地 直送	1.11
	ギフト 直送	1.09
2	果物	1.24
	特産品	1.17
	果物 特産品	1.15
3	果物	1.26
	果物 産地	1.21
	産地	1.2
4	果物 産地 清見オレンジ	1.25
	果物 清見オレンジ	1.25
	果物	1.25
5	雲仙レモネード 果物 清見オレンジ	1.24
	果物	1.21
	レモン果汁 雲仙レモネード 果物	1.21

6. まとめ

本論文では，任意の関係にある語のペアを入力として与え，それと類似した関係にある語のペアを Web から取得する手法を提案した．類似関係の抽出には語の分布の差異を利用した手法を用い，精度の低下を防ぐために抽出されたペア及び抽出手法を WebPMI を用いて評価した．都道府県名と特産物の関係にあるシーズを入力として与えた実験において，1 回目の試行では高い適合率での類似関係抽出に成功し，2 回目以降の反復では大きく適合率を損なうことなく大量の関係を取得することに成功した．

[謝辞]

本研究の一部は，京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」，および，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，課題番号 18049041）によるものです．ここに記して謝意を表します．

[文献]

[1] O. Etzioni, M. J. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld and A. Yates: “Web-Scale Information Extraction in KnowItAll: (Preliminary Results)”, Proceedings of the 13th International World Wide Web Conference (WWW 2004), pp. 100-110 (2004).

[2] S. Soderland, O. Etzioni, T. Shaked and D. Weld: “The Use of Web-based Statistics to Validate Information Extraction”, Proceedings of AAAI 2004 Workshop on Adaptive Text Extraction and Mining (ATEM’04) (2004).

[3] D. Bollegala, Y. Matsuo and M. Ishizuka: “Measuring Semantic Similarity between Words Using Web Search Engines”, Proceedings of the 16th International World Wide Web Conference (WWW 2007), pp. 757-766 (2007).

[4] S. Brin: “Extracting Patterns and Relations from the World Wide Web”, The World Wide Web and Databases, International Workshop (WebDB’98), pp. 172-183 (1998).

[5] E. Agichtein and L. Gravano: “Snowball: Extracting Relations from Large Plain-Text Collections”, Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 85-94 (2000).

[6] 張建偉，石川佳治，北川博之：“トピックを考慮した大規模文書情報源からのレコード抽出”，情報処理学会論文誌 vol.48, no.SIG 14 (TOD 35), p.107-123 (2007).

[7] T. Hokama and H. Kitagawa: “Extracting Mnemonic Names of People from the Web”, Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL 2006), pp. 121-130 (2006).

[8] 楠村幸貴，土方嘉徳，西田正吾：“テンプレートの交叉と DOM 構造の解析による情報抽出手法の提案”，電子情報通信学会第 17 回データ工学ワークショップ論文集 (DEWS2006) (2006).

[9] 加藤誠，大島裕明，小山聡，田中克己：“語の共起を用いた Web の類似関係検索”，Web とデータベースに関するフォーラム (WebDB Forum 2008) (2008).

加藤 誠 Makoto P. KATO
京都大学大学院情報学研究科社会情報学専攻修士課程在学中．2008 年京都大学工学部情報学科卒業．主に情報検索の研究に従事．日本データベース学会学生会員．

大島 裕明 Hiroaki OHSHIMA
京都大学大学院情報学研究科社会情報学専攻特定助教．2007 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主にウェブ，情報検索，データベースの研究に従事．情報処理学会，電子情報通信学会，日本データベース学会，ACM 各会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助教．2002 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に機械学習，データマイニング，情報検索の研究に従事，電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．京大工博．主にデータベース，マルチメディアコンテンツ処理の研究に従事. IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．