社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143（22）
2007／7／2

語の同位関連と性質の継承関連を用いた概念階層のWebからの抽出

服部 峻† 大島 裕明† 小山 聡† 田中 克己†

† 京都大学大学院 情報学研究科 社会情報学専攻 〒606-8501 京都府京都市左京区吉田本町
E-mail: †{hattori,ohshima,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 概念間の階層関係は自然言語処理システムにおいて非常に重要な知識であり，Webなどの大量の文書コーパスをテキストマイニングすることで概念階層を自動抽出する研究が盛んに行われている．しかしながら，従来手法の多くは，特定の構文パターンに合致する記述が文書コーパス中に多く含まれるという階層関係の十分条件に基づいており，より厳密な構文パターンを用いると適合率は高いが再現率は低く，より緩い構文パターンを用いると再現率は改善されるが適合率を著しく損なってしまう．本稿で我々は，再現率を改善しつつ適合率も維持するために，上位下位関係の必要十分条件として概念間の性質の継承関連を仮定する．ある概念間に対して，各々の概念の典型的な性質を求めた上で，互いへ性質が継承される度合いを評価することで，上位下位関係の有無を判定する．さらに，一方の概念を上位概念と仮定した場合に，もう一方の概念への性質の継承度を評価するだけでなく，その同位概念への性質の継承度も合わせて評価することで，概念間の上位下位関係をより厳密に抽出する手法を提案する．
キーワード Webマイニング，上位下位関係，部分全体関係，継承関係，同位関係．

Extracting Conceptural Hierarchies from the Web by Term Coordinate and Property Inheritance Relationships

Shun HATTORI†, Hiroaki OHSHIMA†, Satoshi OYAMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
E-mail: †{hattori,ohshima,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract Conceptual hierarchies, such as hyponymy and meronymy relations, are very important for natural language processing systems. Many researchers have tackled how to mine very large corpora of documents such as the Web for conceptual hierarchies. However, their methods are mostly based on lexico-syntactic patterns as not necessary but sufficient conditions of conceptual hierarchies, so they can achieve high precision but low recall using stricter patterns or they can achieve high recall but low precision using looser patterns. In this paper, we assume property inheritance relationships to be necessary and sufficient conditions of hyponymy relations to achieve high recall and not low precision, and propose a method to acquire hyponymy relations from the Web based on property inheritance relationships between a superordinate concept and a subordinate concept with its coordinate concept.
Key words Web Mining, Hyponymy, Meronymy, Inheritance Relations, Coordinate Relations.

1. はじめに

オントロジーにおける概念間の上位下位（is-a）関係や部分全体（has-a）関係といった意味的な階層関係は，情報検索における検索質問の拡張・修正[1]，質問応答[2]や機械翻訳，セマンティックWebなどにおける知識の共有・再利用，テキストマイニングによるオブジェクトの情報抽出など，様々な自然言語処理システムにとって非常に重要な知識である．オブジェクトの外観情報抽出[3]の場合，例えば「カワセミ」の外観情報は，「青い鳥」「美しい鳥」などと「カワセミ」の上位語である「鳥」と，或いは，「青い羽」「長い嘴」などと「カワセミ」の構成要素である「羽」や「嘴」と外観修飾句のペアで記述されることが多く，対象のオブジェクト名の上位語や部分語の知識を利用できれば，より効率的に外観情報を抽出することが可能になる．

法造[4]などのオントロジーエディタの整備が進められてはいるが，オントロジーを人手で構築するには莫大なコストが掛かり，また，構築者の主観の影響も無視できない．さらに，既存概念の変貌や新概念の出現に対して随時保守することも困難である．この問題に対して，一般的な情報源として重要度の高い新聞記事や，多種多様なユーザにより文書が記述され蓄積さ

れて行く Web や Blog などの大量の文書コーパスをテキストマイニングすることで，オントロジーを自動的に構築・拡張できれば非常に有益であり，これまでに多くの研究がなされている．

上位下位関係の定義の一つとして，Miller [5] らは，「英語常用者が "x is a (kind of) y" という表現を受け入れられるならば，概念 y は概念 x の上位概念である（概念 x は概念 y の下位概念である）」という次の定義を与えている．

"x is a (kind of) y" と言って不自然でない $\Rightarrow$ x is-a y

文書コーパスから上位下位関係を抽出する従来手法の多くは，Miller らの定義の近似として，「"y such as x" などの構文パターンに合致する記述が文書コーパス中に多く含まれるならば，概念 y は概念 x の上位概念である」という次の仮説に基づく．

"y such as x" などの記述を大量に含む $\Rightarrow$ x is-a y

この仮説が成り立つ構文パターンを用意すれば，任意の概念間における上位下位関係の有無を高い適合率で判定可能である．一方で，この仮説が真であるとしても，「構文パターンに合致する記述が文書コーパス中に多く含まれる」ことは，「概念 y が概念 x の上位概念である」ことの十分条件でしかなく必要条件ではないため，たとえ「概念 y は概念 x の上位概念である」ことが真であったとしても，必ずしも「構文パターンに合致する記述が文書コーパス中に多く含まれる」とは限らない．従って，高い適合率を実現するために厳密な構文パターンだけを用意すると，その構文パターンに合致する記述が文書コーパス中で少なくなってしまうため，本当は上位下位関係を持つ概念間に対して上位下位関係が無いと誤判定する危険性が増し，再現率を損なうというトレードオフの問題が付き纏う．ある程度の再現率を確保するために緩い構文パターンを使用し，適合率が悪化することは妥協する場合も多い．Hearst [6] らによって，上位下位関係の構文パターンを発見する手法も提案されてはいるが，上位下位関係の十分条件を用いるという根本的な問題は残る．

文書コーパスから上位下位関係を抽出する際，再現率を改善しつつ，適合率も維持するためには，緩い構文パターンを十分条件として使用した上で，さらに必要条件で絞る必要がある．そこで，我々は，「概念 y は概念 x の上位概念である」という上位下位関係の必要条件として，「概念 x が概念 y の性質を全て継承する」という性質の継承関連を仮定する．

x is-a y $\Rightarrow$ x は y の性質を全て継承する

この仮説は，オブジェクト指向方法論におけるクラス間でのデータメンバ（属性）とメソッド（振る舞い）の継承関係 [7] や，属性分析法における事象階層構造での属性遺伝 [8] に基づく．

本稿で我々は，上位下位関係の必要十分条件として概念間の性質の継承関連を仮定し，ある概念間に対して，各々の概念の典型的な性質を求めた上で，互いへ性質が継承される度合いを評価することで，上位下位関係の有無を判定する．さらに，一方の概念を上位概念と仮定した場合に，もう一方の概念への性質の継承度を評価するだけでなく，その同位概念への性質の継承度も合わせて評価することで，概念間の上位下位関係をより厳密に抽出する手法を提案する．

2. 関連研究

本章では，関連研究として，概念間の上位下位関係や同位関係，部分全体関係を，Web などの大量の文書コーパスからテキストマイニングする従来手法について紹介する．

2.1 上位下位関係

上位下位関係とは，概念間の包含関係である．概念 y が概念 x の上位概念であるとは，概念 y が概念 x を含み，より抽象的である場合である．同時に，概念 x を概念 y の下位概念であると呼び，概念 x は概念 y に含まれ，より具体的である．

新聞記事や Web などの大量の文書コーパスをテキストマイニングすることで，概念間の上位下位関係を自動抽出する手法がこれまでに数多く提案されている．Hearst [6] は，"A such as B" や "such A as B" といった構文パターンを用意しておき，文書コーパスから構文パターンに合致する記述を収集することで，概念間の上位下位関係を獲得する手法を提案し，新しい構文パターンを発見する手法についても述べている．従来研究の多くはこの流れを汲み，概念間の上位下位関を抽出するための様々な構文パターンが提案されている [9]〜[11]．しかしながら，前章でも述べたように，「"A such as B" などの構文パターンに合致する記述が文書コーパス中に大量に含まれる」ことは，「概念 A が概念 B の上位概念である」ことの必要条件ではないため，構文パターンを網羅的に用意したとしても，構文パターンに合致する記述が文書コーパス中に運良く十分に含まれない限り，上位下位関係のある概念間に対して上位下位関係なしと誤判定してしまう危険性が高いという問題がある．

国語辞典や百科事典における見出し語とその説明文の構造をモデル化し，構文解析などによって上位下位関係を獲得する手法も提案されている．鶴丸 [12] らは，国語辞典を利用し，見出し語とその語義文に現れる定義語との間に階層関係を付けることで，シソーラスを自動構築している．桜井 [13] らは，Web から用語説明を自動生成した上で上位語を決定している．大石 [14] らは，Web を事典的に利用するために構築された Cyclone コーパスを用いて，見出し語とその説明文の方向性を考慮した確率的な出現頻度モデルと局所的な構文情報に基づく統計モデルによって，単語の階層関係を統計的に自動識別している．一方，森本 [15] らは，専門用語の構成規則に基づいて，複合用語を基本構成用語（語基）に分解し，用語の各語基の包含関係を比較することで，専門用語間の階層関係を解析している．

構文パターンに依存しない抽出手法により，概念間の上位下位関係の自動抽出の再現率の改善も図られている．小渕 [16] らは，各語に対して意味素の集合を割り当て，その包含関係によって上位下位関係を定義している．Sanderson [17] や山本 [18] らは，文書コーパス中での二語の出現の仕方に包含関係が認められる場合に，その概念間に上位下位関係があると判定する手法を提案している．新里 [19] らは，箇条書きや表などの HTML タグの繰り返しパターンにより下位語候補の集合を抽出し，DF や IDF などの統計量や表題に基づいて上位語の候補を絞り，名詞が持つ主として動詞との係り受け関係を特徴ベクトル化し類似度を計算することで，上位下位関係の獲得を試みている．

2.2 同位関係

同位関係とは，共通の上位概念を持ち，かつ，その上位概念からの具体的である程度もほぼ等しいという概念間の無向関係である．概念 x と概念 z が互いに同位概念であるとは，共通の上位概念 y を持ち，一方が他方の上位概念でも下位概念でもなく，同一概念でもない場合である．

新里 [20] らは，HTML の文書構造に着目し，同じレベルに列挙されている語句集合に対して，共通の上位語を持つ下位語の集合であると仮定することで，同位関係を抽出している．大島 [21], [22] らは，並列助詞を含む構文パターンや，クエリログにおける共起型の共有に基づいて同位語を発見する手法を提案している．同位関係ではないが，類義関係の抽出に関する研究は非常に多数行われており，相互情報量による意味的な関連の推定 [23]，ベイズ推定を用いたクラスタリング [24]，係り受け関係の類似度によるクラスタリング [25]，表の属性と値の関係を利用した類義語抽出 [26] などが提案されている．

2.3 部分全体関係

部分全体関係とは，概念間の集約関係である．概念 y が概念 x の全体概念であり，概念 x が概念 y の部分概念であるとは，概念 y が概念 x を所有する場合である．

鶴丸 [27] らは，国語辞典に基づくシソーラスの構築に関して，同義関係によるグループ化を利用した極大語の処理，及び，上位下位関係や同義関係との融合による部分全体関係の拡張可能性について論理的に考察している．Sundblad [28] らは，自然言語の質問文コーパスに対して単純なパターンマッチングを行うことで，上位下位関係だけでなく部分全体関係を収集している．

3. 性質の継承関連に基づく上位下位関係の抽出

本章では，概念間の上位下位関係の必要十分条件として，概念間の性質の継承関連を仮定することで，構文パターンや文書構造を十分条件として仮定する従来手法よりも高い再現率で上位下位関係を Web から抽出する手法について提案する．

3.1 性質の継承関連に基づく上位概念の抽出

ある概念 x が与えられた場合に，その上位概念をできる限り洩れなく含むような上位概念候補集合 Y を収集した上で，上位概念候補 y_i から概念 x への性質の継承度に基づいてランキングすることで，概念 x の上位概念を Web から精度良く抽出する手法について述べる．以下の四つのステップから成る．

Step 1. 概念 x の上位概念候補集合の収集：
「“x は y である” に合致する記述が多いほど，概念 y は概念 x の上位概念である」といった構文パターンに基づく仮説や「概念 x を含む文書のタイトルに上位概念 y が出現し易い」といった文書構造に基づく仮説の中で，やや緩い条件を仮定して，ある程度の適合率を維持しつつ再現率が高い概念 x の上位概念候補集合 Y を得る．任意の概念を概念 x の上位概念の候補とすることも可能ではあるが，あらゆる概念と概念 x のペアに対して，以下のステップを実行するのは現実的でない．

Step 2. 概念 x および上位概念候補の性質の抽出：
各概念の持つ性質として，オブジェクト指向に則り，属性名と振る舞いを想定する．オブジェクトの外観情報の抽出 [3] で用いたオブジェクトの構成要素名の抽出手法を本稿でも利用する．概念 x の典型的な性質を求めるには，まず，[x の] という検索クエリを画像検索エンジンで実行し，検索された画像の周辺テキスト中で “x の” に続く名詞や動詞を候補 p_j とし，概念 x の性質としての典型度 $\mathrm{property}_x(p_j)$ を次式により評価する．

$$\mathrm{property}_x(p_j) := \frac{\mathrm{if}([x\ \text{の}\ p_j])}{\mathrm{if}([x\ \text{の}])}$$

但し，if([q]) は，画像検索エンジンで検索クエリ [q] を実行した検索結果の件数を表す．文書検索エンジンではなく，画像検索エンジンを用いるのは，写真のテーマを記述する語句が概念 x の性質を表していることが多いという観測に基づく．

Step 3. 上位概念候補から概念 x への性質の継承度の評価：
上位概念候補 y_i から概念 x への性質の継承度 $\mathrm{inherit}_{y_i}(x)$ を評価するため，上位概念候補 y_i の上位 n 件の典型的な性質 p_j と，概念 x と上位概念候補 y_i それぞれに対する典型度とを元に生成した性質ベクトルの内積値によって定義する．

$$\mathrm{inherit}_{y_i}(x) := \sum_{p_j \in P_n(y_i)} \mathrm{property}_x(p_j) \cdot \mathrm{property}_{y_i}(p_j)$$

但し，$P_n(y_i)$ は，上位概念候補 y_i の上位 n 件の典型的な性質の集合を表す．

Step 4. 上位概念候補の概念 x に対する相応度の評価：
上位概念候補 y_i に対して，概念 x の上位概念としての相応度 $\mathrm{hypernym}_x(y_i)$ を次式で定義し，概念 x の上位概念候補集合 Y 全体を相応度に基づいてランキングする．

$$\mathrm{hypernym}_x(y_i) := \mathrm{inherit}_{y_i}(x)$$

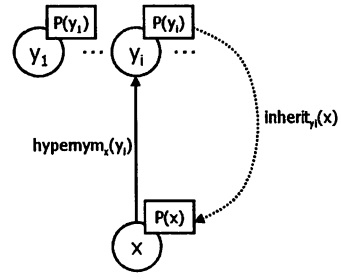

図 1 性質の継承関連に基づく上位概念の抽出

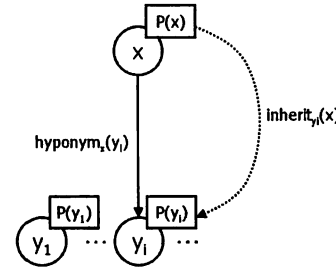

図 2 性質の継承関連に基づく下位概念の抽出

3.2 性質の継承関連に基づく下位概念の抽出

ある概念 x が与えられた場合に，下位概念候補 Y を網羅的に収集した上で，概念 x から下位概念候補 y_i への性質の継承度に基づいてランキングし，概念 x の下位概念を Web から抽出する手法について述べる．以下の四つのステップから成る．

Step 1. 概念 x の下位概念候補集合の収集：

「"y は x である" に合致する記述が多いほど，概念 y は概念 x の下位概念である」といった構文ターンに基づく仮説や「概念 x をタイトルに含む文書中に下位概念 y が出現し易い」といった文書構造に基づく仮説の中で，やや緩い条件を仮定して，概念 x の下位概念候補集合 Y を網羅的に得る．

Step 2. 概念 x および下位概念候補の性質の抽出：

3.1 節の Step 2. と同様にして求める．

Step 3. 概念 x から下位概念候補への性質の継承度の評価：

概念 x から下位概念候補 y_i への性質の継承度 $\mathrm{inherit}_x(y_i)$ を評価するには，まず，概念 x の上位 n 件の典型的な性質 p_j と，概念 x と下位概念候補 y_i それぞれに対する典型度とを元に，各々の性質ベクトルを生成し，その内積値によって定義する．

$$\mathrm{inherit}_x(y_i) := \sum_{p_j \in P_n(x)} \mathrm{property}_x(p_j) \cdot \mathrm{property}_{y_i}(p_j)$$

Step 4. 下位概念候補の概念 x に対する相応度の評価：

下位概念候補 y_i に対して，概念 x の下位概念としての相応度 $\mathrm{hyponym}_x(y_i)$ を次式で定義し，概念 x の下位概念候補集合 Y 全体を相応度に基づいてランキングする．

$$\mathrm{hyponym}_x(y_i) := \mathrm{inherit}_x(y_i)$$

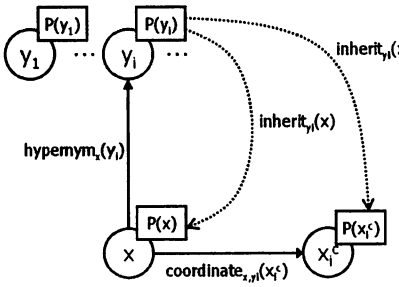

図 3　語の同位関連と性質の継承関連に基づく上位概念の抽出

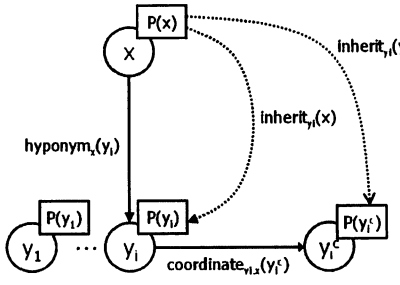

図 4　語の同位関連と性質の継承関連に基づく下位概念の抽出

4. 語の同位関連に基づく上位下位関係の抽出

本章では，概念間の性質の継承関連に加えて，語の同位関連も考慮することで，前章で提案した上位下位関係の Web からの抽出手法よりもさらに適合率の改善を図る．

4.1 語の同位関連に基づく上位概念の抽出

「概念 y_i が概念 x の上位概念であるならば，概念 y_i を上位概念とする概念 x の同位概念 x_i^c が存在し，概念 x だけでなく同位概念 x_i^c も上位概念 y_i の性質を継承する」という仮説に基づき，ある概念 x が与えられた場合に，上位概念候補 y_i から概念 x および x の同位概念 x_i^c への性質の継承度の総和に基づいてランキングすることで，概念 x の上位概念をより精度良く抽出する手法について述べる．以下の五つのステップから成る．

Step 1. 概念 x の上位概念候補集合の収集：

3.1 節の Step 1. と同様にして求める．

Step 2. 上位概念候補における概念 x の同位概念の抽出：

概念 x の同位概念のうち，上位概念候補 y_i を上位概念として持つ同位概念 x_i^c を求める．大島 [21] らの同位語の Web マイニング手法を利用して，上位概念候補 y_i を上位概念として持つ同位概念としての相応度 $\mathrm{coordinate}_{x,y_i}(x_i^c)$ の値が最大となる概念 x の同位概念 x_i^c を採用する．

$$\mathrm{coordinate}_{x,y_i}(x_i^c) := \sqrt{\mathrm{ya}_{y_i}(x, x_i^c) \cdot \mathrm{ya}_{y_i}(x_i^c, x)}$$
$$\mathrm{ya}_{y_i}(x, x_i^c) := \mathrm{df}([\text{“}x \text{ や } x_i^c\text{”} \text{ AND } y_i])$$
$$\mathrm{ya}_{y_i}(x_i^c, x) := \mathrm{df}([\text{“}x_i^c \text{や } x\text{”} \text{ AND } y_i])$$

但し，$\mathrm{df}([q])$ は，文書検索エンジンで検索クエリ $[q]$ を実行した検索結果の件数を表す．

Step 3. 概念 x，同位概念，上位概念候補の性質の抽出：

3.1 節の Step 2. と同様にして求める．

Step 4. 上位概念候補から下位概念への性質の継承度の評価：

上位概念候補 y_i から概念 x および同位概念 x_i^c への性質の継承度を，3.1 節の Step 3. と同様にして求める．

Step 5. 上位概念候補の概念 x に対する相応度の評価：

上位概念候補 y_i に対して，概念 x の上位概念としての相応度 $\mathrm{hypernym}_x(y_i)$ を次式で定義し，概念 x の上位概念候補集合 Y 全体を相応度に基づいてランキングする．

$$\mathrm{hypernym}_x(y_i) := \mathrm{inherit}_{y_i}(x) + \mathrm{inherit}_{y_i}(x_i^c)$$

4.2 語の同位関連に基づく下位概念の抽出

「概念 y_i が概念 x の下位概念であるならば，概念 x を上位概念とする概念 y_i の同位概念 y_i^c が存在し，概念 y_i だけでなく同位概念 y_i^c も上位概念 x の性質を継承する」という仮説に基づき，ある概念 x が与えられた場合に，概念 x から下位概念候補 y_i および y_i の同位概念 y_i^c への性質の継承度の総和に基づいてランキングすることで，概念 x の下位概念をより精度良く抽出する手法について述べる．以下の五つのステップから成る．

Step 1. 概念 x の下位概念候補集合の収集：

3.2 節の Step 1. と同様にして求める．

Step 2. 概念 x における下位概念候補の同位概念の抽出：

4.1 節の Step 2. と同様にして，下位概念候補 y_i の同位概念の

うち，概念 x を上位概念として持つ同位概念 y_i^c を求める．

$$\mathrm{coordinate}_{y_i,x}(y_i^c) := \sqrt{\mathrm{ya}_x(y_i, y_i^c) \cdot \mathrm{ya}_x(y_i^c, y_i)}$$

$$\mathrm{ya}_x(y_i, y_i^c) := \mathrm{df}([\text{“}y_i\text{や }y_i^c\text{”} \text{ AND } x])$$

$$\mathrm{ya}_x(y_i^c, y_i) := \mathrm{df}([\text{“}y_i^c\text{や }y_i\text{”} \text{ AND } x])$$

Step 3. 概念 x，同位概念，上位概念候補の性質の抽出：
3.1 節の Step 2. と同様にして求める．
Step 4. 概念 x から下位概念への性質の継承度の評価：
概念 x から下位概念候補 y_i および同位概念 y_i^c への性質の継承度を，3.2 節の Step 3. と同様にして求める．
Step 5. 下位概念候補の概念 x に対する相応度の評価：
下位概念候補 y_i に対して，概念 x の下位概念としての相応度 $\mathrm{hyponym}_x(y_i)$ を次式で定義し，概念 x の下位概念候補集合 Y 全体を相応度に基づいてランキングする．

$$\mathrm{hyponym}_x(y_i) := \mathrm{inherit}_x(y_i) + \mathrm{inherit}_x(y_i^c)$$

5. 実　　験

本章では，語の同位関連と性質の継承関連を用いた概念階層の Web からの抽出手法のうち，「鳥」という概念の下位概念を抽出した実験結果について考察する．

表 1 は，“（下位概念）は鳥である”という構文パターンに合致した語句を件数によってランキングした結果，及び，構文パターンに合致した語句の全てを「鳥」の下位概念の候補集合とした上で，「鳥」の上位 $n(=5, 10)$ 件の典型的な性質を継承している度合いに基づいてランキングした結果を比較している．「鳥」の典型的な性質の上位は以下のように求まり，上位 n 件の性質と典型度とを用いて，「鳥」および下位概念候補の性質ベクトルを生成し，「鳥」から下位概念候補への性質の継承度を算出している．表 1 において，ランキングされている各概念に続く数値が「鳥」からの性質の継承度を表している．

> 声 (0.158)，さえずり (0.114)，鳴き声 (0.093)，巣 (0.083)，名前 (0.041) 5，唐揚げ (0.038)，夏 (0.038)，姿 (0.037)，糞 (0.032)，種類 (0.023) 10，群れ (0.022)，歌 (0.021)，絵 (0.018)，足 (0.018)，...

また，次頁の図 5 は，「鳥」の下位語抽出の上位 k 件の平均適合率を比較している．上位 2 件までは適合率 1.0 で差はないが，その後大きな差へと広がって行っている．構文パターンのみ用いる場合に適合率を維持する方法として合致件数に閾値条件を設けることが考えられるが，例えば閾値条件を 5 件以上と設定した場合，構文パターンによる「鳥」の下位概念抽出の再現数は 5 個，適合率は 0.71 となる．継承関連による「鳥」の下位概念抽出では，$n = 5$ の場合，上位 k 件の適合率が 0.71 よりも低くなる上位 13 件目までの再現数は 9 個であり，大きく改善されているが，$n = 10$ の場合，上位 9 件目までの再現数は 6 個であり，$n = 5$ の場合に比べて悪化してしまっている．これは，「鳥」の典型的な性質として相応しくない「唐揚げ」や「夏」が性質ベクトルに加わってしまったためであり，性質ベクトルとして用いる典型的な性質を適切に選択する必要がある．

表 1　継承関連に基づく「鳥」の下位語抽出の上位 20 件

	構文パターン	継承関連 (n=5)	継承関連 (n=10)
1	**ペンギン** (17)	**うぐいす** (0.0712)	**うぐいす** (0.0717)
2	**カラス** (9)	**コマドリ** (0.0544)	**コマドリ** (0.0552)
3	動物 (7)	**ツバメ** (0.0379)	**ツバメ** (0.0392)
4	**カナリア** (7)	**海猫** (0.0261)	**海猫** (0.0264)
5	我輩 (7)	**カラス** (0.0205)	**カラス** (0.0214)
6	**スズメ** (5)	天馬 (0.0143)	天馬 (0.0153)
7	**ハト** (5)	**スズメ** (0.0129)	**スズメ** (0.0144)
8	もの (4)	**カナリヤ** (0.0111)	たち (0.0123)
9	私 (4)	アリ (0.0109)	自分 (0.0122)
10	兎 (3)	自分 (0.0097)	アリ (0.0115)
11	自分 (3)	**カナリア** (0.0094)	**カナリヤ** (0.0113)
12	恐竜 (3)	**鴉** (0.0092)	彼女 (0.0104)
13	双子 (3)	たち (0.0089)	**カナリア** (0.0097)
14	**カナリヤ** (3)	彼女 (0.0088)	**鴉** (0.0096)
15	**ニワトリ** (2)	**鷲** (0.0087)	**鷲** (0.0093)
16	これ (2)	**トキ** (0.0081)	**トキ** (0.0086)
17	**ツバメ** (2)	絵鳥 (0.0072)	吾輩 (0.0085)
18	部分 (2)	私 (0.0068)	私 (0.0083)
19	編鳥 (2)	**始祖鳥** (0.0064)	**ハト** (0.0081)
20	**ダチョウ** (2)	**ニワトリ** (0.0062)	コウモリ (0.0075)

表 2　同位関連に基づく「鳥」の下位語抽出の上位 20 件

	同位関連 (n=5) [同位概念]	同位関連 (n=10) [同位概念]
1	**コマドリ** (0.1107) [オオルリ]	**コマドリ** (0.1156) [オオルリ]
2	**うぐいす** (0.0828) [ひばり]	**うぐいす** (0.0851) [ひばり]
3	**ツバメ** (0.0508) [スズメ]	**ツバメ** (0.0549) [スズメ]
4	**海猫** (0.0357) [カモメ]	**海猫** (0.0387) [カモメ]
5	**スズメ** (0.0333) [カラス]	**スズメ** (0.0368) [カラス]
6	**カラス** (0.0333) [スズメ]	**カラス** (0.0368) [スズメ]
7	アリ (0.0271) [ハチ]	**ハト** (0.0302) [カラス]
8	**ハト** (0.0264) [カラス]	アリ (0.0290) [ハチ]
9	**カナリヤ** (0.0169) [文鳥]	彼女 (0.0207) [彼]
10	天馬 (0.0166) [飛竜]	もの (0.0202) [人]
11	**カナリア** (0.0162) [文鳥]	**カナリヤ** (0.0201) [文鳥]
12	彼女 (0.0161) [彼]	**鴉** (0.0188) [鳩]
13	**鴉** (0.0161) [鳩]	**カナリア** (0.0187) [文鳥]
14	もの (0.0148) [人]	天馬 (0.0185) [飛竜]
15	相手 (0.0143) [自分]	相手 (0.0180) [自分]
16	**トキ** (0.0141) [コウノトリ]	自分 (0.0159) [家族]
17	**鷹** (0.0132) [鷲]	たち (0.0158) [牛]
18	**鷲** (0.0132) [鷹]	**トキ** (0.0152) [コウノトリ]
19	自分 (0.0118) [家族]	自分たち (0.0145) [自分]
20	自分たち (0.0109) [自分]	**鷹** (0.0143) [鷲]

6. おわりに

Web などの大量の文書コーパスをテキストマイニングすることで概念間の上位下位関係を抽出する従来手法の多くは，特定の構文パターンに合致する記述が多く含まれるという十分条件に基づいており，より厳密な構文パターンを用いると適合率は高いが再現率は低く，より緩い構文パターンを用いると再現率は改善されるが適合率を著しく損なってしまう．本稿で我々は，再現率を改善しつつ適合率も維持するために，上位下位関係の

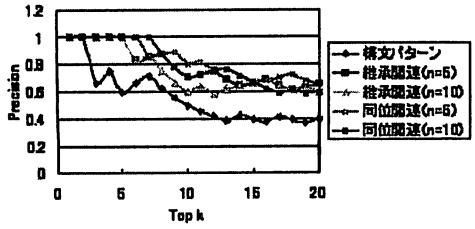

図 5 「鳥」の下位語抽出の上位 k 件平均適合率

必要十分条件として概念間の性質の継承関連を仮定した．ある概念間に対して，一方の概念を上位概念と仮定した場合に，もう一方の概念への性質の継承度を評価するだけでなく，その同位概念への性質の継承度も合わせて評価することで，概念間の上位下位関係をより厳密に抽出する手法を提案した．

謝　辞

本研究の一部は，文部科学省研究委託事業「異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己）」，及び，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」における計画研究「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），及び，計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」（研究代表者：安達淳，Y00-01，課題番号：18049073），文部科学省科学研究費補助金若手研究（B）「Web からの履歴情報の発見とその呈示方式の研究」（研究代表者：小山聡，課題番号：19700091）による．ここに記して謝意を表します．

文　献

[1] Mandala, R., Tokunaga, T., and Tanaka, H., "The Use of WordNet in Information Retrieval," *Proceedings of the COLING ACL Workshop on Usage of WordNet in Natural Language Processing*, pp.31–37 (1998).

[2] Fleischman, M., Hovy, E. and Echihabi, A., "Offline Strategies for Online Question Answering:. Answering Questions Before They Are Asked," *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, pp.1–7 (2003).

[3] 服部 峻, 手塚太郎, 田中克己, "オブジェクトの外観情報の Web マイニング," 電子情報通信学 会第 18 回データ工学ワークショップ（DEWS'07）論文集, L4-6 (2007).

[4] 古崎晃司, 來村徳信, 池田 満, 溝口理一郎, "「ロール」および「関係」に関する基礎的考察に基づくオントロジー記述環境の開発," 人工知能学会論文誌, vol.17, no.3, pp.196–208 (2002).

[5] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J., "Introduction to WordNet: An On-line Lexical Database," *International Journal of Lexicography*, vol.3, no.4, pp.235–312 (1993).

[6] Hearst, M. A., "Automatic Acquisition of Hyponyms from Large Text Corpora," *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, vol.2, pp.539–545 (1992).

[7] 青木利晃, 片山卓也, "オブジェクト指向方法論のための形式的モデル," 日本ソフトウェア科学会学会誌コンピュータソフトウェア, vol.16, no.1, pp.12–32 (1999).

[8] 王 凱軍, 池田 満, 國藤 進, "属性分析法に基づく類似性の分析," 第 18 回人工知能学会全国大会, 2F3-02 (2004).

[9] Caraballo, S. A., "Automatic construction of a hypernym-labeled noun hierarchy from text," *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp.120–126 (1999).

[10] 安藤まや, 関根 聡, 石崎 俊, "定型表現を利用した新聞記事からの下位概念単語の自動抽出," 情報処理学会研究報告「自然言語処理」, vol.2003, no.98, pp.77–82 (2003).

[11] Emmanuel, M., Christian, J., "Automatic Acquisition and Expansion of Hypernym Links," *Computer and the Humanities*, vol.38, no.4, pp.363–396 (2004).

[12] 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田 将: "国語辞典情報を用いたシソーラスの作成について," 情報処理学会研究報告「自然言語処理」, vol.1991, no.37, pp.121–128 (1991).

[13] 桜井 裕, 佐藤理史: "ワールドワイドウェブを利用した用語説明の自動生成," 情報処理学会論文誌, vol.43, no.5, pp.1470–1480 (2002).

[14] 大石康智, 伊藤克亘, 武田一哉, 藤井 敦, "単語の共起関係と構文情報を利用した単語階層関係の統計的自動識別," 情報処理学会研究報告「音声言語情報処理」, vol.2006, no.40, pp.25–30, (2006).

[15] 森本貴之, 藤原 譲, "例外処理を考慮した用語間の階層・関連関係の抽出," 情報知識学会第 8 回研究報告会講演論文集, no.8, pp.17–22 (2000).

[16] 小淵洋一, 斉藤 隆, "意味の分割によるシソーラスの自己組織," 情報処理学会研究報告「情報学基礎」, vol.1992, no.54, pp.17–23 (1992).

[17] Sanderson, M., and Croft, B., "Deriving Concept Hierarchies from Text," *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.206–213 (1999).

[18] 山本英子, 神崎享子, 井佐原均: "出現状況の包含関係による語彙の階層構造の構築," 情報処理学会論文誌, vol.47, no.6, pp.1872–1883 (2006).

[19] 新里圭司, 鳥澤健太郎, "HTML 文書からの単語間の上位下位関係の自動獲得," 自然言語処理, vol.12, no.1, pp.125–150, (2005).

[20] 新里圭司, 鳥澤健太郎 "HTML 文書中の箇条書きとその表題に注目した下位語の自動獲得," 情報処理学会研究報告「自然言語処理」, vol.2004, no.93, pp.29–36 (2004).

[21] 大島裕明, 小山 聡, 田中克己, "Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見," 情報処理学会論文誌（トランザクション）データベース, vol.47, no.SIG19(TOD32), pp.98–112 (2006).

[22] 大島裕明, 山口雅史, 小山 聡, 田中克己, "Web 検索エンジンのインデックスとクエリログを用いた同位語発見," 情報処理学会データベースと Web 情報システムに関するシンポジウム（DBWeb'06）論文集, pp.305–312 (2006).

[23] Church, K. W., and Hanks, P., "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol.16, no.1, pp.22–29 (1990).

[24] Ghahramani, Z. and Heller, K., "Bayesian Sets," *Advances in Neural Information Processing Systems 18 (NIPS'05)*, pp.435–442 (2006).

[25] Lin, D., "Automatic Retrieval and Clustering of Similar Words," *Proceedings of the 17th International Conference on Computational Linguistics and of the 36th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'98)*, pp.768–774 (1998).

[26] 関恒仁, 鳩田和孝, 遠藤 勉, "表の属性と属性値の関係を利用した類義語抽出," 電子情報通信学会論文誌, vol.J89-D, no.9, pp.2087–2100 (2006).

[27] 鶴丸弘昭, 前田英幸, 山本和博, 日高 達, 吉田 将, "国語辞典に基づくシソーラスの構築に関する一考察," 電子情報通信学会技術研究報告「言語理解とコミュニケーション」, vol.93, no.367, pp.29–36 (1993).

[28] Sundblad, H., "Automatic Acquisition of Hyponyms and Meronyms from Question Corpora," *Proceedings of the ECAI'02 Workshop on Natural Language Processing and Machine Learning for Ontology Engineering*, (2002).