情報処理学会論文誌　データベース　Vol. 1　No. 3　60–81 (Dec. 2008)

性質継承と概念の再帰的適用に基づくWebからの概念階層抽出

服部　峻[†1]　田中克己[†1]

上位下位関係や部分全体関係といった概念階層は，様々な自然言語処理システムにとって非常に重要な基本的知識である．概念階層の構築が人海戦術的に進められている一方で，Web などの大規模な文書データベースから自動的に知識抽出する研究も数多く行われている．しかしながら，従来の抽出手法の多くは構文パターンに基づいているため，上位下位関係の厳密な構文パターンを用いると適合率は高いが再現率が非常に低くなり，逆に，曖昧な構文パターンを用いると再現率は高くなるが適合率が非常に低くなってしまうという問題があった．これに対して我々は，上位下位関係の構文パターンに合致する文書頻度とは異なる評価軸として，対象概念から下位概念候補への性質継承の度合いに基づく抽出手法を提案する．さらに，注目している 2 つの概念間の直接的な関係を評価するだけでなく，これらの周辺にある概念との関係も考慮することによって，提案手法のロバスト性の向上を図る．具体的には，対象概念の上位概念や下位概念候補の同位概念を厳密な構文パターンを用いて高い適合率で抽出したうえで，対象概念の上位概念から下位概念候補への性質継承の度合い，対象概念から下位概念候補の同位概念への性質継承の度合いなども加味する．また，各概念の典型的な性質を抽出する手法においても，各概念と各性質との間の直接的な関係を評価するだけでなく，対象概念の上位概念からの性質継承や対象概念の下位概念集合からの性質集約も考慮することによって改善を図る．

Extracting Concept Hierarchy Knowledge from the Web by Property Inheritance and Recursive Use of Term Relationships

Shun Hattori[†1] and Katsumi Tanaka[†1]

Concept hierarchies, such as hyponymy and meronymy, are very fundamental for various natural language processing systems. Many researchers have tackled how to mine very large corpora of documents such as the Web for concept hierarchy knowledge. However, their methods are mostly based on lexico-syntactic patterns as not necessary but sufficient conditions of concept hierarchies, so they can achieve high precision but low recall when using stricter patterns or they can achieve high recall but low precision when using looser patterns. In this paper, we propose a method to extract concept hierarchies from the Web based on "Property Inheritance" from a target concept to its subordinate candidate, as a different measure from the document frequency of lexico-syntactic patterns for concept hierarchies. To make our method more robust, we also utilize the other concepts surrounding them, e.g., not only property inheritance from a target concept to its subordinate candidate, but also property inheritance from its superordinate concept to its subordinate candidate and/or from the target concept to a coordinate concept of its subordinate candidate. In addition, we refine a method to extract typical properties for each concept from the Web by utilizing property inheritance from its superordinate concept to the target concept and/or "Property Aggregation" from a set of its subordinate concepts to the target concept.

1. はじめに

近年，モバイルインターネットの整備と Web 検索エンジンの進歩により，携帯電話などのモバイル端末を持ち歩きさえすれば，いつでも，どこでも，Web 検索エンジンを利用することが可能になってきている．実空間を移動中に出遭ったオブジェクトに興味を持ち，そのオブジェクトに関する様々な情報をその場で検索したいという要望は少なくない．しかしながら，ユーザ自身にとっては情報を検索したい対象オブジェクトは具体的なインスタンスであり非常に明確であるにもかかわらず，そのオブジェクトの具体的な名称が不明である場合，Web 検索エンジンに対して入力する検索クエリが曖昧になり，精度の悪い検索結果しか得られないという問題がある．これに対して，対象オブジェクトをより厳密に指定できるはずの具体的な名称は不明であったとしても，まず，より抽象的な上位の概念を表現するクラス名や，特徴に関する記述などをユーザが代替指定することによって，そのオブジェクトの具体的な名称をシステムが特定したうえで，それを Web 検索エンジンに対して検索クエリとして入力すれば，より精度の良い検索結果を得ることができる．したがって，モバイル環境における情報検索を改善するためには，クラス名と特徴記述が入力として与えられた場合に，そのクラスに属するオブジェクトの具体的な名称を網羅的に取得したうえで，その特徴記述にマッチする度合いでランキングし，さらには，各々のオブジェクト名に典型

†1 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

的な特徴記述を添えてランキング結果を返すことで，対象オブジェクトの名称をインタラクティブに特定してゆくことを支援するオブジェクト名サーチが必要であると考える．特に，モバイル環境で出遭ったオブジェクトの特徴を記述する場合，外観や動作などの視覚情報が最も利用されると考えられるため，五感情報に基づくオブジェクト名サーチ[1] の実現を我々は目指している．このようなシステムを実現するためには，上位下位（is-a）関係や部分全体（has-a）関係といった概念階層，および，オブジェクト名と五感情報との相互変換に関する知識が必要不可欠である．

概念間の上位下位関係や部分全体関係といった意味的な階層関係は，我々が目指しているオブジェクト名サーチだけでなく，情報検索における検索質問の拡張・修正[2)–4)]，質問応答[5] 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[6)–8)] など，様々な自然言語処理システムにとっても非常に重要な基本的知識である．

WordNet[9),10)] や Wikipedia[11),12)] において概念階層の構築が人海戦術的に進められている一方で，多種多様なユーザにより文書が記述・蓄積されてゆく 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自動的に構築・拡張する研究も数多く行われている．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，「“x such as y” 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ならば，概念 y は概念 x の下位概念である」という仮説に基づいているが，これ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大量に含まれる」ことは，「概念 y が概念 x の下位概念である」ことの十分条件でしかなく必要条件ではない．したがって，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いると，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ないと誤判定する危険性が増し，再現率を損なってしまう．逆に，高い再現率を実現するために曖昧な構文パターンを用いると，そもそもの仮説を満たさなくなり，適合率が著しく悪化してしまう．

オブジェクト名サーチの基本的知識として利用するためには，対象語（クラス名）に属する下位語をできる限り網羅的に抽出する必要がある．したがって，厳密な構文パターンに基づく従来手法では不十分であり，再現率を改善しつつ，適合率も高く維持するためには，上位下位関係の緩い構文パターンを用いて候補語を収集したうえで，構文パターンに合致する文書頻度とは異なる評価尺度，つまり，上位下位関係の十分条件ではなく，必要（十分）条件を用いてフィルタリングする必要がある．そこで我々は，「概念 y は概念 x の下位概念である」という上位下位関係の必要条件として，「概念 y が概念 x の性質をすべて継承する」という「性質継承」を仮定する．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（振舞い）の継承関係[13)] や，属性分析法における事象階層構造での属性遺伝[14)] に基づく．

本論文で我々は，上位下位関係の必要十分条件として概念間の「性質継承」を仮定し，対象概念（クラス名）が与えられた場合に，その対象概念と下位概念候補各々の典型的な性質を求めたうえで，対象概念から下位概念候補への性質継承の度合いを評価することによって，対象概念に属する下位概念（具体的なオブジェクト名）を Web から抽出する手法を提案する．さらに，2 つの概念間の直接的な関係を評価するだけでなく，これらの周辺にある概念との関係も考慮することで，提案手法のロバスト性の向上を図る．具体的には，対象概念の上位概念や下位概念候補の同位概念を厳密な構文パターンを用いて高い適合率で抽出したうえで，対象概念の上位概念から下位概念候補への性質継承の度合い，対象概念から下位概念候補の同位概念への性質継承の度合いなども考慮する．また，各概念の典型的な性質を抽出する手法においても，各概念と各性質との間の直接的な関係を評価するだけでなく，対象概念の上位概念からの性質継承や対象概念の下位概念集合からの「性質集約」も考慮することで改善を図る．

本論文の以降の構成を示す．まず 2 章では関連研究を紹介し，提案手法との比較も行う．次に 3 章では性質継承に基づく下位語抽出手法を，4 章では周辺概念も考慮した性質継承に基づく下位語抽出手法を，さらに 5 章では下位概念集合からの性質集約も考慮した性質継承に基づく下位語抽出手法を提案する．6 章では各々の提案手法に対する評価実験の結果を考察する．最後に，7 章で本論文をまとめ，今後の課題も述べる．

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 や新聞記事などの大量の文書コーパスからテキストマイニングによって自動的に抽出する従来手法について紹介する．

2.1 上位下位関係

新聞記事や Web ページ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[15)] は，“x such as y” や “such x as y” といった上位下位関係の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係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[16)–18)]．しかしながら，前章でも述べたように，「“x such as y” 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 y が概念 x の下位概念である」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大量に含まれない限り，上位下位関係を持つ概念間に対して上位下位関係がないと誤判定してしまう危険性が根本的に残る．

上位下位関係の構文パターンを利用するだけでなく，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．鶴丸ら[19)]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．桜井ら[20)]は，Web から用語説明を自動生成したうえで上位語を決定している．大石ら[21)]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，森本ら[22)]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抽出の再現率の改善を図っている研究も多数あり，本論文と非常に関連がある．小渕ら[23)]は各語に対して意味素の集合を割り当て，Sanderson ら[24)]や KnowItAll[25),26)]では文書コーパス中での各語の出現の仕方に基づいて素性を割り当て，語概念間で包含関係が認められる場合に上位下位関係があると判定する．山本ら[27)]は，あらゆる形容詞や形容動詞を索引としてあらかじめ定めた共起ベクトルを各抽象名詞を修飾する頻度に基づいて求め，抽象名詞間の共起ベクトルの包含関係をオーバラップ相関係数や補完類似度で評価して，抽象名詞の階層構造を自動構築している．一方，我々の性質継承に基づく抽出手法では，対象概念の典型的な性質の上位 n 件のみを索引として採用し，典型的でない多数の性質については性質継承の度合いを評価する際に関知しないという点が異なる．また，山本らは 1 回でも共起することを重視して各索引に重み付けしているが，Web においては複数回共起しなければ重視すべきでないと我々は考える．新里ら[28)]は，箇条書きや表などの HTML タグの繰返しパターンにより下位語候補を抽出し，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，各名詞が持つ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している．これらの研究と同様に，我々の提案する性質継承に基づく抽出手法も，本質的にはベクトル間の類似度計算に準じているが，索引として用いる性質語を部分全体（has-a）関係や振舞い表現に限定し自動抽出している点，注目している概念間の関係だけでなく，上位概念や同位概念，下位概念といった周辺概念も考慮して性質継承の度合いを評価する点が異なる．

2.2 同位関係

新里ら[29)]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は共通の上位語を持つ下位語の集合であると仮定して同位関係を抽出している．大島ら[30),31)]は，並列助詞を含む構文パターンや，検索クエリのログにおける共起型の共有に基づいて同位関係を発見している．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数多く行われており，相互情報量による意味的な関連の推定[32)]，ベイズ推定を用いたクラスタリング[33)]，係り受け関係の類似度によるクラスタリング[34)]，表の属性と値の関係を利用した類義語抽出[35)]などが提案されている．

2.3 部分全体関係

鶴丸ら[36)]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，および，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[37)]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. 性質継承に基づく下位語抽出

本章では，概念間の上位下位関係の十分条件として構文パターンや文書構造を仮定する従来手法とは異なり，必要十分条件として概念間の「性質継承」を仮定し，対象概念とその下位概念候補との間の性質継承の度合いを評価することによって，上位下位関係を Web から精度良く抽出する手法について提案する．

「概念 y は概念 x の下位概念である」という上位下位関係の必要（十分）条件として，「概念 y が有する性質の集合 $P(y)$ は，概念 x が有する性質の集合 $P(x)$ のすべてを包含する（概念 y が概念 x の性質をすべて継承する）」という概念間の「性質継承」を仮定する．

$$\mathrm{isa}(y,x)=1 \Leftrightarrow P(y) \supseteq P(x) \text{ and } y \neq x,$$
$$P(c)=\{p \in P |\ \mathrm{has}(p,c)=1\}.$$

ただし，P は性質の全体集合を，$\mathrm{has}(p,c)$ は概念 c が性質 p を有するか否かの二値 $\{0,1\}$

を表す．

$$\mathrm{has}(p,c)=\begin{cases}1 & \text{if 概念 } c \text{ が性質 } p \text{ を有する},\\ 0 & \text{otherwise}.\end{cases}$$

いい換えると，「概念 y は概念 x の下位概念である」ならば「概念 x と概念 y が共有する性質の数が概念 x の有する性質の数と等しい（かつ，概念 y の有する性質の数よりも小さい）」という次の関係が成り立つ．

$$\mathrm{isa}(y,x)=\begin{cases}1 & \text{if } \sum_{p\in P}\mathrm{has}(p,y)\cdot\mathrm{has}(p,x)=\sum_{p\in P}\mathrm{has}(p,x),\\ 0 & \text{if } \sum_{p\in P}\mathrm{has}(p,y)\cdot\mathrm{has}(p,x)<\sum_{p\in P}\mathrm{has}(p,x).\end{cases}$$

以上により，$\sum\mathrm{has}(p,y)\cdot\mathrm{has}(p,x)/\sum\mathrm{has}(p,x)$ の値が 1 と等しいか否かに基づいて上位下位関係の有無を判定できるが，概念 c と性質 p の任意のペアに対して二値 $\{0,1\}$ で $\mathrm{has}(p,c)$ を正確に求められることが必要不可欠である．これは容易ではなく，各概念に対して Web から典型的な性質を抽出する手法を用いると基本的には連続値 $[0,1]$ である $\mathrm{has}^*(p,c)$ しか利用できない．何らかの閾値を境界にして二値 $\{0,1\}$ に射影することは可能であるが，誤りをまったく含まないことは期待できず，下位概念 y の典型的な性質集合 $P(y)$ が上位概念 x の典型的な性質集合 $P(x)$ を完全に包含することは実際には稀有である．

そこで，概念 y が概念 x の下位概念であるか否かを表す二値 $\mathrm{isa}(y,x)$ の近似として，概念 y が概念 x の下位概念である相応しさを表す連続値 isa-PI$^*(y,x)$ を，概念 x から概念 y へ性質が継承されている度合い $\sum\mathrm{has}^*(p,y)\cdot\mathrm{has}^*(p,x)/\sum\mathrm{has}^*(p,x)\cdot\mathrm{has}^*(p,x)$ によって評価する．そして，概念 x と概念 y の任意のペアが与えられた場合，概念 x から概念 y への性質継承度 isa-PI$^*(y,x)$ が閾値 T（$0\ll T<1$）より大きく，かつ，概念 y から概念 x への性質継承度 isa-PI$^*(x,y)$ が閾値 T より小さければ，概念 y は概念 x の下位概念である可能性が高いと判定する．または，対象概念 x と複数の下位概念候補の集合 $C(x)$ が与えられた場合，閾値 T を定めることなく，対象概念 x から各下位概念候補 $y\in C(x)$ への性質継承度 isa-PI$^*(y,x)$ に基づいてランキングする．ただし，性質継承度 isa-PI$^*(y,x)$ の分母を元の $\sum\mathrm{has}^*(p,x)$ ではなく $\sum\mathrm{has}^*(p,x)\cdot\mathrm{has}^*(p,x)$ に変更した理由は，分母の値分布（次数）を分子に合わせ，比が 1 に近い値をとる可能性を残すためである．後者の場合には，下位概念候補間での大小だけが重要であり，分母が下位概念候補 y に依存しない限り対象概念 x が定まれば固定値となるため，$\sum\mathrm{has}^*(p,x)$ であっても，$\sum\mathrm{has}^*(p,x)\cdot\mathrm{has}^*(p,x)$ であっても，あるいは，1 であっても（このとき性質継承度 isa-PI$^*(y,x)$ は内積）ランキング結果は変わらない．しかし，前者の場合には，概念 y が概念 x の下位概念であるときに性質の継承度 isa-PI$^*(y,x)$ が十分に 1 に近い値をとる可能性が高くなる必要があるが，元の $\sum\mathrm{has}^*(p,x)$ のままでは比が 1 に近い値をとる可能性が低くなりすぎる．

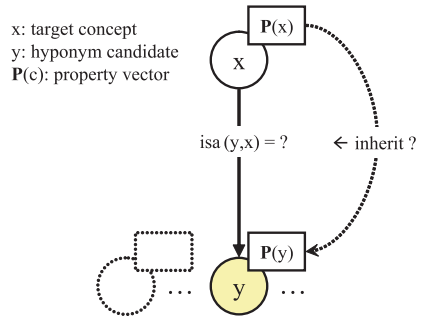


図 1　性質継承に基づく下位概念抽出

Fig. 1　Hyponym extraction based on property inheritance from a target concept to its subordinate candidate.

対象概念 x が与えられた場合に，その下位概念候補集合 $C(x)$ を Web から網羅的に収集したうえで，対象概念 x の典型的な性質を求め，その上位 n 件のみを下位概念候補 y が継承している度合いに基づいてランキングすることで，対象概念 x の下位概念を Web から精度良く抽出する手法（**図 1**）について詳述する．

Step 1．下位概念候補集合の収集：

対象概念 x の下位概念をできる限り洩れなく含む下位概念候補集合 $C(x)$ を収集する必要があるが，あらゆる語概念を候補としてしまうと，候補数およびノイズ割合が大きくなりすぎるため不適切である．構文パターンに基づく「“y は x である” に合致する記述が文書コーパス中に十分に多ければ，概念 y は概念 x の下位概念である」や，文書構造に基づく「概念 y を含む内容に文書のタイトルに上位概念 x が出現しやすい」などの上位下位関係の仮説の中から，厳密すぎず比較的緩い条件を選定することで，ノイズ割合を抑えつつ，対象概念 x の下位概念候補 y を網羅的に収集する．

6 章の評価実験においては，“y は x である”，および，“x である y” という 2 種類の構文パターンを用いて構成した検索クエリを Yahoo!ウェブ検索 API[38] に与え，最大 1,000 件ずつの検索結果から形態素解析で名詞句を切り出し下位概念候補とする．

Step 2. 各概念の典型的な性質の抽出：

各概念 c の有する典型的な性質 p を表す語句として，オブジェクト指向に則り属性名や振舞い表現を想定し，“c の p” という構文パターンで出現することが多いという知見[39),40)]に基づいて Web から抽出する．

- Web 文書中における “c の p” の頻度（document）

  $f_l^d(p,c) := \mathrm{df}([\text{“}c \text{ の } p\text{”}])$,
  $f_g^d(c) := \mathrm{df}([\text{“}c \text{ の”}])$.

  ただし，$\mathrm{df}([q])$ は Yahoo!ウェブ検索 API[38)] で検索クエリ $[q]$ を実行した検索結果の件数を表す．
- 画像の周辺における “c の p” の頻度（image）

  $f_l^i(p,c) := \mathrm{if}([\text{“}c \text{ の } p\text{”}])$,
  $f_g^i(c) := \mathrm{if}([\text{“}c \text{ の”}])$.

  ただし，$\mathrm{if}([q])$ は Yahoo!画像検索 API[41)] で検索クエリ $[q]$ を実行した検索結果の件数を表す．オブジェクトの外観情報抽出[6),40)] で用いたオブジェクトの構成要素名を抽出する手法である．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 c の構成要素名 p を表すことが多いという観測に基づく．
- スニペット中における近接共起度（snippet）

  $f_l^s(p,c) := \mathrm{sf}([c \ \&\ p])$,
  $f_g^s(c) := \mathrm{sf}([c])$.

  ただし，$\mathrm{sf}([c \ \&\ p])$ は Yahoo!ウェブ検索 API[38)] で検索クエリ $[c]$ を実行した最大 1,000 件の検索結果スニペット中における p の頻度を表す．格助詞「の」に基づく手法の比較用で，一般的な特徴語を抽出する手法である．

以上の局所的な共起頻度 $f_l^*(p,c)$ と大局的な頻度 $f_g^*(c)$ を基本項として，概念 c に対する性質 p の相応しさの度合いとして次の 3 種類を定義する．

- 複数回共起することを重要視（proportion）

$$\mathrm{has}_p^*(p,c) := \frac{f_l^*(p,c)}{f_g^*(c)} \in [0,1].$$

- 1 回でも共起することを重要視（once-oriented）

$$\mathrm{has}_o^*(p,c) := \frac{f_l^*(p,c)}{f_l^*(p,c)+1} \in [0,1].$$

- 1 回でも共起することを重要視（binary）

$$\mathrm{has}_b^*(p,c) := \begin{cases} 1 & \text{if } f_l^*(p,c) \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

新聞記事や辞書などとは異なり，Web 文書中でわずか 1 回共起しただけではノイズである可能性は依然として高く，複数回共起するほど適合解である可能性が高くなるという比率による定義を我々は推挙する．

Step 3. 性質継承の度合いの評価：

各下位概念候補 y について対象概念 x の下位概念である相応しさを評価するため，対象概念 x の典型的な性質の上位 n 件を下位概念候補 y が継承している度合いを次式で定義する．

$$\text{isa-PI}_n^*(y,x) := \frac{\sum_{p \in P_n(x)} \mathrm{has}^*(p,y) \cdot \mathrm{has}^*(p,x)}{\sum_{p \in P_n(x)} \mathrm{has}^*(p,x) \cdot \mathrm{has}^*(p,x)}.$$

ただし，$P_n(x)$ は Step 2 で定義した局所的な共起頻度 $f_l^*(p,x)$ に基づいて順序付けした対象概念 x の典型的な性質の上位 n 件だけの集合を表す．最後に，対象概念 x から下位概念候補 y への性質継承度 $\text{isa-PI}_n^*(y,x)$ を，下位概念候補 y が対象概念 x の下位概念である相応しさの度合いを評価する尺度として用い，下位概念候補集合 $y \in C(x)$ のすべてをランキングした結果を返す．ここで，性質継承度 $\text{isa-PI}_n^*(y,x)$ の分母は，下位概念候補 y に依存しない関数である限りは対象概念 x に対して固定値となるため，$\sum \mathrm{has}^*(p,x)$ であっても，あるいは，1 であっても（このとき内積と等価になる）ランキング結果は変わらない．

特徴ベクトル空間における一般的な類似度尺度であるコサイン相関値を参照すると，分子は同一であるが，分母も下位概念候補 y に依存する関数であるため，上述の定義とは異なるランキング結果を示すことになる．

$$\cos_n^*(y,x) := \frac{\sum_{p \in P_n(x)} \mathrm{has}^*(p,y) \cdot \mathrm{has}^*(p,x)}{\sqrt{\sum_{p \in P_n(x)} \mathrm{has}^*(p,y)^2} \sqrt{\sum_{p \in P_n(x)} \mathrm{has}^*(p,x)^2}}.$$

純粋なコサイン相関値の場合，概念 x と概念 y の任意のペアに対して類似度計算に用いる性質集合が固定であり，完全に対称な関数であるため，両者を入れ替えても同じ値とな

る．一方，ここで定義した $\cos^*_n(y,x)$（擬似コサイン相関値と呼ぶことにする）は，片方の概念 x に依存する典型的な性質の上位 n 件のみからなる性質集合を用いて計算されるため非対称な関数であり，性質継承度 isa-PI$^*_n(y,x)$ の定義と同様に，上位下位関係の方向性も識別しうる．

山本ら[27] は，語彙の階層構造の自動構築において，二値画像のための補完類似度や多値画像のための補完類似度を用いて，語彙間の出現状況の包含関係を評価している．二値画像のための補完類似度 $\mathrm{csm}(y,x)$ は，元々は劣化印刷文字を認識するために提案された尺度で，下位概念候補（印刷文字）y を表す二値ベクトルが対象概念（テンプレート文字）x を表す二値ベクトルをどの程度包含しているかを表す．

$$\mathrm{csm}(y,x) := \frac{a\cdot d - b\cdot c}{\sqrt{(a+c)(b+d)}},$$

$$a := \sum_{p\in P} \mathrm{has}(p,y)\cdot \mathrm{has}(p,x),$$

$$b := \sum_{p\in P} \mathrm{has}(p,y)\cdot (1-\mathrm{has}(p,x)),$$

$$c := \sum_{p\in P} (1-\mathrm{has}(p,y))\cdot \mathrm{has}(p,x),$$

$$d := \sum_{p\in P} (1-\mathrm{has}(p,y))\cdot (1-\mathrm{has}(p,x)).$$

ここで，a は両者がともに有する性質の数，b は下位概念候補 y は有するが対象概念 x は持たない性質の数，c は下位概念候補 y は持たないが対象概念 x は有する性質の数，d は両者がともに持たない性質の数に相当し，これらの総和はベクトル次元数 N と等しい．一方，多値画像のための補完類似度 $\mathrm{csm}^*(y,x)$ は，グレースケール画像を扱うために拡張された尺度である．

$$\mathrm{csm}^*(y,x) := \frac{a^*\cdot d^* - b^*\cdot c^*}{\sqrt{N\cdot X_2 - X^2}},$$

$$a^* := \sum_{p\in P} \mathrm{has}^*(p,y)\cdot \mathrm{has}^*(p,x),$$

$$b^* := \sum_{p\in P} \mathrm{has}^*(p,y)\cdot (1-\mathrm{has}^*(p,x)),$$

$$c^* := \sum_{p\in P} (1-\mathrm{has}^*(p,y))\cdot \mathrm{has}^*(p,x),$$

$$d^* := \sum_{p\in P} (1-\mathrm{has}^*(p,y))\cdot (1-\mathrm{has}^*(p,x)),$$

$$X := \sum_{p\in P} \mathrm{has}^*(p,x),\quad X_2 := \sum_{p\in P} \mathrm{has}^*(p,x)^2.$$

4. 周辺概念との性質継承に基づく下位語抽出

前章では，対象概念から下位概念候補への性質継承度に基づいて上位下位関係を抽出する手法について述べたが，本章では，2 つの概念間の直接的な関係だけでなく，これらの周辺にある概念との関係も考慮することで抽出手法を改善する．**図 2** のように，周辺概念としては，対象概念の上位概念，対象概念を共通の上位概念として持つ下位概念候補の同位概念，下位概念候補の下位概念の 3 種類があるが，本章では前者 2 つの周辺概念との間の性質継承の活用について述べる．

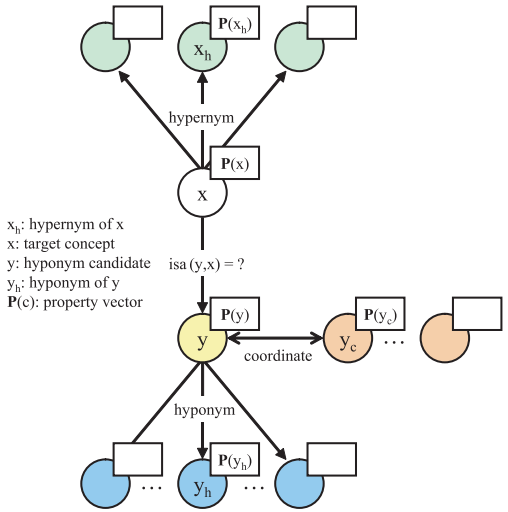


図 2　性質継承に基づく下位概念抽出における周辺概念
Fig. 2　Surrounding concepts in hyponym extraction based on property inheritance.

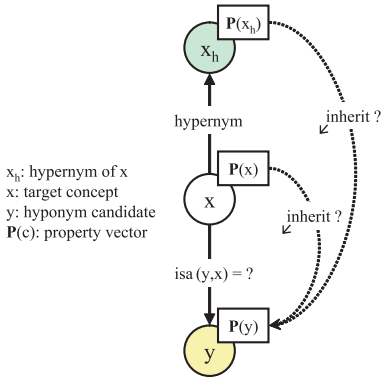


図 3　上位概念も考慮した性質継承に基づく下位概念抽出

Fig. 3　Hyponym extraction based on property inheritances from not only a target concept x but also its superordinate concept x_h to its subordinate candidate y.

上位概念も考慮すると，「概念 y は概念 x の下位概念である」ならば，「概念 y が有する性質の集合 $P(y)$ は，概念 x が有する性質の集合 $P(x)$ だけでなく，概念 x の上位概念 x_h が有する性質の集合 $P(x_h)$ も包含する」という制約条件に拡張できる．

$$\mathrm{isa}(y,x)=1 \Rightarrow P(y) \supseteq P(x) \text{ and } P(y) \supseteq P(x_h),$$
$$\text{where } x_h \in \mathrm{Hypernym}(x).$$

ただし，$\mathrm{Hypernym}(x)$ は概念 x の上位概念の集合を表す．

また，同位概念も考慮すると，「概念 y は概念 x の下位概念である」ならば，「概念 y が有する性質の集合 $P(y)$ だけでなく，概念 y の同位概念 y_c が有する性質の集合 $P(y_c)$ も，概念 x が有する性質の集合 $P(x)$ のすべてを包含する」という制約条件に拡張できる．

$$\mathrm{isa}(y,x)=1 \Rightarrow P(y) \supseteq P(x) \text{ and } P(y_c) \supseteq P(x),$$
$$\text{where } y_c \in \mathrm{Coordinate}(y,x).$$

ただし，$\mathrm{Coordinate}(y,x)$ は概念 x が共通の上位概念である概念 y の同位概念の集合を表す．

対象概念 x が与えられた場合に，その下位概念候補集合 $C(x)$ を Web から網羅的に収集したうえで，対象概念 x から下位概念候補 y への性質継承度に加えて，対象概念 x の上位概念 x_h から下位概念候補 y への性質継承度，あるいは，対象概念 x から下位概念候補 y の同位概念 y_h への性質継承度も考慮してランキングすることで，対象概念 x の下位概念を Web から抽出する手法（**図 3**，あるいは，**図 4**）について述べる．

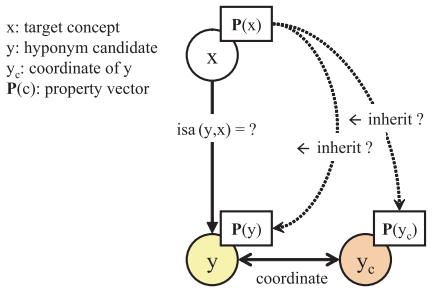


図 4　同位概念も考慮した性質継承に基づく下位概念抽出

Fig. 4　Hyponym extraction based on property inheritances from a target concept x to not only its subordinate candidate y but also a coordinate concept of its subordinate candidate y_c.

Step 1. 下位概念候補集合の収集：

3 章の Step 1 と同様である．

Step 2. 上位概念や同位概念の抽出：

上位概念も考慮した性質継承に基づいて下位概念抽出を行うためには，対象概念 x の上位概念を少なくとも 1 つ抽出できればよく，網羅的に抽出することよりも精度の方が重要である．“x は x_h である”，および，“x_h である x” という 2 種類の構文パターンの文書頻度を統合する場合，相加平均（和），相乗平均（積），調和平均などを用いることが一般的であるが，再現率よりも上位 1 件の適合率を重視するため，両者の文書頻度の最小値を基本項として用い，上位概念候補 x_h に対して，対象概念 x の上位概念としての相応度を次式で定義する．残りの項は，最小値が同じであるペアの間で差を付けるためである[4]．

$$\mathrm{hypernym}(x_h,x) := \min_h + \left(1 - \frac{\min_h}{\max_h}\right),$$
$$\min_h := \min\{\mathrm{df}_h^{\mathrm{A}}, \mathrm{df}_h^{\mathrm{B}}\},$$
$$\max_h := \max\{\mathrm{df}_h^{\mathrm{A}}, \mathrm{df}_h^{\mathrm{B}}\},$$
$$\mathrm{df}_h^{\mathrm{A}} := \mathrm{df}([\text{“}x \text{ は } x_h \text{ である”}]),$$
$$\mathrm{df}_h^{\mathrm{B}} := \mathrm{df}([\text{“}x_h \text{ である } x\text{”}]).$$

同位概念も考慮した性質継承に基づいて下位概念抽出を行うためには，対象概念 x を共通の上位概念として持つ下位概念候補 y の同位概念を少なくとも 1 つ精度良く抽出する必要がある．上述と同様に再現率よりも適合率を重視し，“y や y_c”，および，“y_c や y” とい

う厳密な構文パターンの文書頻度の最小値を基本項として用い，同位概念候補 y_c に対して，対象概念 x を共通の上位概念として持つ下位概念候補 y の同位概念としての相応度を次式で定義する．

$$\mathrm{coordinate}(y_c, y, x) := \mathrm{min}_c + \left(1 - \frac{\mathrm{min}_c}{\mathrm{max}_c}\right),$$
$$\mathrm{min}_c := \min\{\mathrm{df}_c^{\mathrm{A}}, \mathrm{df}_c^{\mathrm{B}}\},$$
$$\mathrm{max}_c := \max\{\mathrm{df}_c^{\mathrm{A}}, \mathrm{df}_c^{\mathrm{B}}\},$$
$$\mathrm{df}_c^{\mathrm{A}} := \mathrm{df}([\text{“}y \text{ や } y_c\text{”} \ \&\ x]),$$
$$\mathrm{df}_c^{\mathrm{B}} := \mathrm{df}([\text{“}y_c \text{ や } y\text{”} \ \&\ x]).$$

Step 3. 各概念の典型的な性質の抽出：

3 章の Step 2 と同様である．

Step 4. 性質継承の度合いの評価：

周辺概念として，上位概念も考慮した性質継承に基づく下位概念抽出では，対象概念 x から下位概念候補 y への性質継承の度合いだけでなく，対象概念 x の上位概念 x_h から下位概念候補 y への性質継承の度合いもパラメータ α で線形結合した次式を用いてランキングする．

$$\text{isa-PIH}_n^*(y, x) := (1-\alpha)\cdot \text{isa-PI}_n^*(y, x) + \alpha \cdot \text{isa-PI}_n^*(y, x_h).$$

同様に，同位概念も考慮した性質継承に基づく下位概念抽出では，対象概念 x から下位概念候補 y への性質継承の度合いだけでなく，対象概念 x から下位概念候補 y の同位概念 y_c への性質継承の度合いもパラメータ β で線形結合した次式を用いてランキングする．

$$\text{isa-PIC}_n^*(y, x) := (1-\beta)\cdot \text{isa-PI}_n^*(y, x) + \beta \cdot \text{isa-PI}_n^*(y_c, x).$$

5. 性質集約に基づく下位語抽出

これまでの章では，対象概念から下位概念候補への性質継承を基本項として，これらの周辺概念も考慮し，対象概念の上位概念から下位概念候補への性質継承，あるいは，対象概念から下位概念候補の同位概念への性質継承も加味して，上位下位関係を Web から抽出する手法について提案してきたが，本章では，上位の概念から下位の概念への性質継承とは逆方向の，下位の概念（の集合）から上位の概念への「性質集約」に着目することで，提案手法の改善を図る．

オブジェクト指向方法論において，下位概念（インスタンスや下位クラス）集合からボトムアップ的に上位概念（上位クラス）を導き定義する場合，下位概念の有する性質の共通部分を上位概念の性質として付与して汎化する．そこで，「概念集合 Y に属する概念 y のすべてが概念 x の下位概念であるならば，概念集合 Y に属する概念 y が有する性質の集合 $P(y)$ のすべての共通集合は，概念 x が有する性質の集合 $P(x)$ のすべてを包含する」という概念間の「性質集約」を仮定する．

$$\forall y \in Y,\ \mathrm{isa}(y, x) = 1 \Leftrightarrow \bigcap_{\forall y \in Y} P(y) \supseteq P(x).$$

いい換えると，少なくとも 1 つ下位概念を持つ（$\sum \mathrm{isa}(c, x) \neq 0$ である）概念 x と性質 p との任意のペアに対して，次の関係が成り立つ．

$$\mathrm{has}(p, x) = \begin{cases} 1 & \text{if } \displaystyle\sum_{c \in C} \mathrm{isa}(c, x)\cdot \mathrm{has}(p, c) = \sum_{c \in C} \mathrm{isa}(c, x), \\ 0 & \text{if } \displaystyle\sum_{c \in C} \mathrm{isa}(c, x)\cdot \mathrm{has}(p, c) < \sum_{c \in C} \mathrm{isa}(c, x). \end{cases}$$

ただし，C は概念の全体集合を表す．

3 章で概念の性質継承を仮定した際にも述べたが，以上の理想的な性質集約に基づいて性質抽出手法を構成する場合，概念 c と性質 p の任意のペアに対する $\mathrm{has}(p, c)$，かつ，任意の概念間に対して $\mathrm{isa}(c, x)$ を二値 $\{0, 1\}$ で正確に求められることが必要不可欠であるが，これは容易なことではなく，これまでの章で提案した性質抽出，および，下位語抽出を用いる場合，連続値 $[0, 1]$ しか利用できない．したがって，概念 x が性質 p を有するか否かを表す $\mathrm{has}(p, x)$ の近似として，性質 p が概念 x の典型的な性質である相応しさの度合いを，$\sum \mathrm{isa}(c, x)\cdot \mathrm{has}(p, c) / \sum \mathrm{isa}(c, x)$ によって評価する．

対象概念 x が与えられた場合，Step 1 で下位概念候補集合 $C(x)$ を網羅的に収集したうえで，まず，対象概念 x から下位概念候補 y への性質継承の度合いに基づいてランキングする．次に，性質継承の度合いが大きい下位概念候補が有する性質を重視して集約し，対象概念 x の典型的な性質の相応度を再評価する．同様に Step 2 と Step 3 を繰り返し，対象概念 x の典型的な性質を再帰的に修正してゆくことで，対象概念 x の下位概念を Web から抽出する手法（**図 5**）である．

Step 1. 下位概念候補集合の収集：

3 章の Step 1 と同様である．

Step 2. 各概念の典型的な性質の（再）抽出：

各概念 c に対する性質 p の 0 次（初期）の相応度 $\mathrm{has}^{(0)}(p, c)$ は，3 章の Step 2 で定義した 3 種類の評価尺度により与える．また，対象概念 x に対する性質 p の $m \in \{1, 2, ...\}$ 次

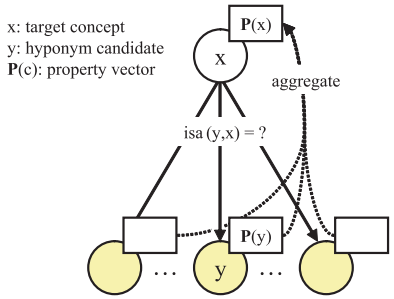


図 5　再帰的な性質集約を行う性質継承に基づく下位概念抽出
Fig. 5　Hyponym extraction based on property inheritance and recursive property aggregation.

の相応度を次式で定義する．

$$\mathrm{has}^{(m)}(p,x) := (1-\gamma)\cdot \mathrm{has}^{(m-1)}(p,x) + \gamma \cdot \frac{\sum_{y\in C(x)} \text{isa-PIA}_n^{(m-1)}(y,x)\cdot \mathrm{has}^{(0)}(p,y)}{\sum_{y\in C(x)} \text{isa-PIA}_n^{(m-1)}(y,x)}.$$

Step 3. 性質継承の度合いの（再）評価：

対象概念 x から下位概念候補 y への m 次の性質継承の度合いを，3 章の Step 3 に対応させて次のように定義する．

$$\text{isa-PIA}_n^{(m)}(y,x) := \frac{\sum_{p\in P_n^{(m)}(x)} \mathrm{has}^{(m)}(p,y)\cdot \mathrm{has}^{(0)}(p,x)}{\sum_{p\in P_n^{(m)}(x)} \mathrm{has}^{(m)}(p,x)\cdot \mathrm{has}^{(m)}(p,x)}.$$

ただし，$P_n^{(m)}(x)$ は Step 2 で求めた m 次の相応度 $\mathrm{has}^{(m)}(p,x)$ に基づいてランキングした対象概念 x の上位 n 件の典型的な性質を表す．（Step 2 へ戻る）

6. 評 価 実 験

本章では，提案手法の有効性を評価するため，従来の構文パターンに基づく下位概念抽出手法をベースラインとして比較実験を行う．まず，対象概念から下位概念候補への性質継承の度合いだけに基づく下位概念抽出手法に関して，性質継承という新しい評価軸の有効性を検証し，各概念に対する典型的な性質を抽出するための評価尺度による精度の違い，性質継承の度合いを評価する際に用いる典型的な性質の数による精度の違いについても考察する．次に，対象概念から下位概念候補への性質継承の度合いだけでなく，対象概念の上位概念から下位概念候補への性質継承の度合い，あるいは，対象概念から下位概念候補の同位概念への性質継承の度合いといった周辺概念との関係も考慮することによる精度の改善を検証する．最後に，下位概念候補集合から対象概念へ性質集約することによって，対象概念の典型的な性質の重みを再帰的に修正することによる精度の改善を検証する．

表 1　各対象概念 x に対して収集された下位概念候補集合 $C(x)$ の分析
Table 1　Analysis of hyponym candidates $C(x)$ for each target concept x.

対象概念 x	下位概念候補数	適合候補数	EDR 辞書との一致数	EDR 辞書での総数
鳥	252	105	60	500
魚	259	102	51	408
昆虫	202	124	43	610
花	226	80	4	147
野菜	227	128	41	229
果物	321	67	18	175
俳優	575	327	0	0
お笑い芸人	163	83	—	—
漫画家	375	185	0	0
家電	132	78	2	12
楽器	303	160	66	427
乗り物	203	108	46	1,434

6.1 Web から収集した下位概念候補の分析

12 種類の対象概念 x に対して，“y は x である”，および，“x である y” という 2 種類の構文パターンを用いて構成した検索クエリを Yahoo!ウェブ検索 API[38] に与え，最大 1,000 件の検索結果のスニペットから形態素解析で名詞句を切り出し下位概念候補 $y \in C(x)$ とする．**表 1** は，各対象概念 x に対して Web から収集された下位概念候補集合 $C(x)$ の要素数，下位概念候補集合中の適合解の数，さらに，既存の辞書として EDR 電子化辞書[42] の概念辞書を用い，EDR 辞書に登録されている下位概念との一致数，EDR 辞書に登録されている下位概念の総数を示している．各下位概念候補の適合・不適合の判断は著者自身で行ったが，EDR 辞書や Wikipedia，一般の Web ページなどを参考にして，できる限り客観的な判断になるように努めている．また，EDR 辞書は，レコード番号，概念識別子，概念見出し（日本語・英語），概念説明（日本語・英語），および，管理情報からなる概念見出しレコード（概念ノード）の集合と，レコード番号，上位概念識別子，下位概念識別子，および，

管理情報からなる概念体系レコード（概念間の二項関係）の集合を備えている．EDR 辞書に登録されている下位概念（語）とは，概念説明が “対象語” と完全一致するか，なければ概念見出しが “対象語 [カナ読み]” と完全一致する概念見出しレコードを始点として下位をすべて展開してゆき，得られた概念見出しレコードの概念見出しから “[カナ読み]” の部分を除いたものである．

既存の EDR 辞書に登録されている下位概念の総数に比べると，下位概念候補集合中の適合解の数は劣っている場合が多いが，最大 1,000 件以上の検索結果のスニペットも解析したり[43]，下位概念候補のさらに下位概念も展開したりすることで S/N 比を維持しつつ改善できる．また，EDR 辞書では，「俳優」や「漫画家」には下位概念が 1 つも登録されておらず，さらに，「お笑い芸人」に至っては語概念としての登録すらされていないが，構文パターンを用いて Web から収集する手法では多数の適合解を獲得できている．動物や植物の種名など，あまり頻繁には下位概念が追加されない概念の場合には，人海戦術による維持管理が可能（有効）であるが，構文パターンを用いて Web から収集できた適合な下位概念の約 4 割以上が EDR 辞書に登録されておらず，対象概念に対して網羅的に下位概念を収集するという目的のためには後者を併用する価値も高い．一方，人名や地名，製品名など，より頻繁に下位概念が追加されてゆく概念の場合には，人海戦術による維持管理は困難になり，増大し続ける Web から自動的に抽出する手法の方が有効である．

6.2　構文パターンに基づく下位概念抽出の検証

我々が提案した性質継承に基づく下位概念抽出手法と比較するため，従来の構文パターンに基づく下位概念抽出をベースラインとして定義する．Web から下位概念候補を収集するために用いた “y は x である”，および，“x である y” という 2 種類の構文パターンの文書頻度を基本項とする．4 章の Step 2 でも述べたが，2 種類の構文パターンの文書頻度を同等に統合する場合，相加平均，相乗平均，調和平均，最小値・最大値に基づく重み付け[4] などが使える．下位概念候補 $y \in C(x)$ が対象概念 x の下位概念である相応しさの度合いを評価する尺度として次の 4 種類を定義する．

- 相加平均（和）

$$\text{isa-SP}^{a}(y,x) := \frac{\text{df}^{\text{A}} + \text{df}^{\text{B}}}{2}.$$

- 相乗平均（積）

$$\text{isa-SP}^{g}(y,x) := \sqrt{\text{df}^{\text{A}} \cdot \text{df}^{\text{B}}}.$$

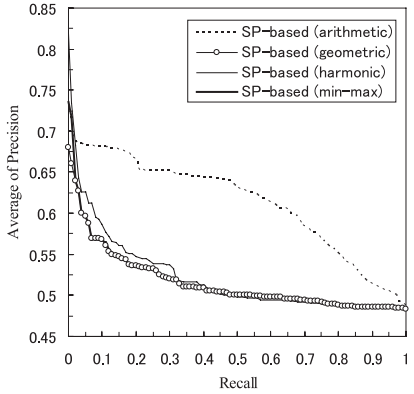


図 6　構文パターンに基づく下位概念抽出の平均 PR 曲線
Fig. 6　Average of PR curve by SP-based hyponym extraction.

- 調和平均

$$\text{isa-SP}^{h}(y,x) := \frac{2}{\frac{1}{\text{df}^{\text{A}}} + \frac{1}{\text{df}^{\text{B}}}}.$$

- 最小値・最大値に基づく重み付け

$$\text{isa-SP}^{m}(y,x) := \min + \left(1 - \frac{\min}{\max}\right),$$
$$\min := \min\{\text{df}^{\text{A}}, \text{df}^{\text{B}}\},$$
$$\max := \max\{\text{df}^{\text{A}}, \text{df}^{\text{B}}\},$$
$$\text{df}^{\text{A}} := \text{df}([\text{“}x \text{ は } y \text{ である”}]),$$
$$\text{df}^{\text{B}} := \text{df}([\text{“}y \text{ である } x\text{”}]).$$

図 6 は，12 種類の対象概念 x に対して，構文パターンに基づく下位概念抽出（4 種類の統合関数）を適用して，下位概念候補集合 $C(x)$ をランキングした結果の適合率・再現率（PR: Precision-Recall）曲線を比較している．明らかに，2 種類の文書頻度を相加平均で統合した重み付けが最良であり，これを以降のベースラインとして選定する．PR 曲線下面積[44] は 0.573 である．

4 章の上位概念や同位概念抽出において，上位 1 件適合率を重視し，最小値・最大値に基づく重み付けを採用したが，この妥当性についても検証する．**図 7** は図 6 の再現率が低い

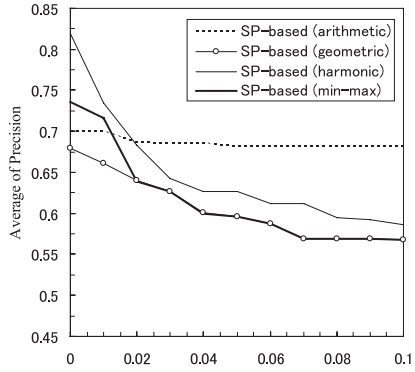


図 7　構文パターンに基づく下位概念抽出の平均 PR 曲線（拡大）
Fig. 7　Average of PR curve by SP-based hyponym extraction (scale-up).

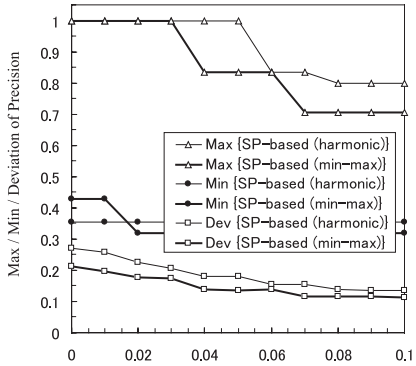


図 8　構文パターンに基づく下位概念抽出の PR 曲線の最大値・最小値・標準偏差（拡大）
Fig. 8　Max, min and deviation of PR curve by SP-based hyponym extraction (scale-up).

区間を拡大したものであり，提案手法は調和平均よりも劣っているが，これは 12 種類の対象概念の PR 曲線の平均である．**図 8** により，再現率が非常に低い区間では提案手法の方が適合率の最小値が大きく標準偏差も小さいため，上位 1 件適合率を重視する目的に対して最良な選定であるといえる．

6.3　性質継承に基づく下位概念抽出の検証

性質継承に基づく下位概念抽出として，概念 c と性質 p との共起頻度を求める手法を 3 種類（document-，image-，snippet-frequency），これら 2 つの文書頻度を統合して性質 p の概念 c に対する典型度を求める手法を 3 種類（proportion，once，binary），下位概念候補 y が対象概念 x の下位概念である相応しさの度合いを測る尺度を 3 種類（性質継承度 isa-PI，擬似コサイン相関値 cos，補完類似度 csm）を 3 章で定義した．

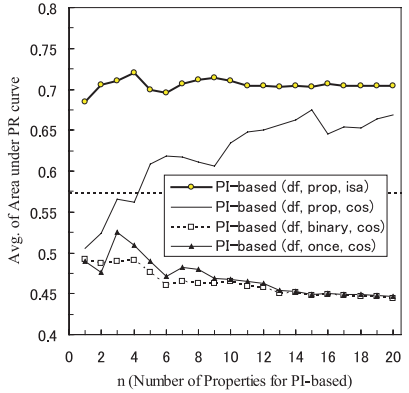


図 9　性質継承に基づく下位概念抽出の平均 PR 曲線下面積（Web 文書頻度 df を元にした場合）
Fig. 9　Average of AuPR by PI-based hyponym extraction using document frequency (df).

まず，**図 9**，**図 10**，**図 11** を用いて，概念 c と性質 p との共起頻度を求める 3 種類の手法ごとに考察を行う．ただし，1 回でも共起することを重要視した連続値（once）または二値（binary）で統合した性質の典型度を用いた場合，性質継承度（isa-PI）で上位下位関係を評価しても擬似コサイン類似度（cos）で評価しても，ほぼ同じ変化を示したため，性質継承度（isa-PI）との組合せは一部割愛している．

図 9 は，Web 文書中における “c の p” の頻度（df）を元にした性質継承に基づく下位概念抽出の PR 曲線下面積を比較している．複数回共起することを重要視した比率（prop）で統合し，性質継承度（isa-PI）で上位下位関係を評価した場合に限り，典型的な性質の件数 n に依存せずつねにベースラインを上回っており，$n = 4$ で最良値 0.722 を記録した後もほぼ 0.70 以上を維持し変動も小さい．一方，性質抽出における統合関数は比率（prop）のまま，上位下位関係の尺度を擬似コサイン相関値（cos）に変更した場合，$n = 4$ まではベースラインよりも低いが，n の増加にともなって上回り，$n = 15$ をピークに改善してゆく．他の組合せの場合，ベースラインの下を減衰していっている．以上により，df は，性質抽出における統合関数として比率（prop）との相性が良く，他との相性は悪い．

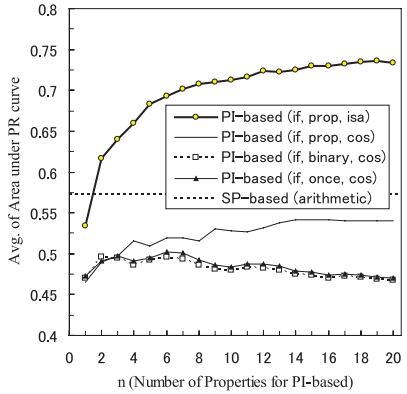


図 10　性質継承に基づく下位概念抽出の平均 PR 曲線下面積（Web 画像頻度 if を元にした場合）
Fig. 10　Average of AuPR by PI-based hyponym extraction using image frequency (if).

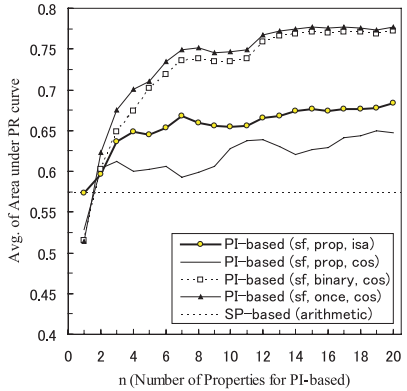


図 11　性質継承に基づく下位概念抽出の平均 PR 曲線下面積（スニペット近接共起頻度 sf を元にした場合）
Fig. 11　Average of AuPR by PI-based hyponym extraction using snippet frequency (sf).

図 10 は，画像の周辺における “c の p” の頻度（if）を元にした性質継承に基づく下位概念抽出の PR 曲線下面積を比較している．df と同様に比率（prop）で統合し，性質継承度（isa-PI）で上位下位関係を評価した場合に限り，$n = 1$ を除いてベースラインを上回り，n の増加にともなって改善してゆき，$n = 19$ で最良値 0.737 を記録している．一方，性質抽出における統合関数は比率（prop）のまま，上位下位関係の尺度を擬似コサイン相関値（cos）

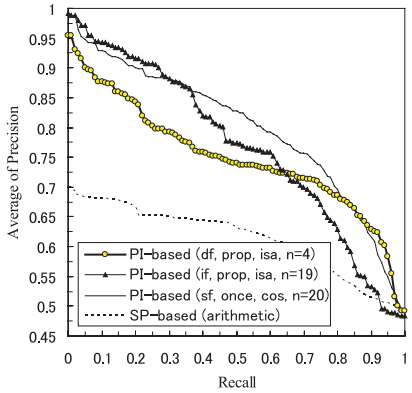


図 12　性質継承に基づく下位概念抽出の平均 PR 曲線
Fig. 12　Average of PR curve by PI-based hyponym extraction.

に変更した場合，n の増加にともなって改善していってはいるが，df とは異なりベースラインを大きく下回っている．以上により，if も，性質抽出における統合関数として比率（prop）との相性が良く，他とは悪い．また，上位下位関係の尺度として擬似コサイン相関値（cos）との相性も良くない．

図 11 は，スニペット中における近接共起度（sf）を元にした性質継承に基づく下位概念抽出の PR 曲線下面積を比較している．1 回でも共起することを重要視した連続値（once）または二値（binary）のどちらで統合しても，上位下位関係の尺度によらず，n の増加にともなって急激に改善し約 0.780 を上界に収束するという同様の変化を示している．性質抽出における統合関数を連続値（once）に，上位下位関係の尺度を擬似コサイン相関値（cos）にした場合の $n = 20$ で最良値 0.777 を記録している．一方，比率（prop）で統合した場合も，n の増加にともなって緩やかに改善していっている．また，$n = 1$ を除いて，どんな組合せにおいてもベースラインを上回っている．以上により，PR 曲線下面積の観点では，スニペット中における近接共起度（sf）を，1 回でも共起することを重要視した連続値（once）で統合し，擬似コサイン相関値（cos）で上位下位関係を評価する手法が全体で最良である．

図 12 は，共起頻度を求める 3 種類の手法ごとにベストな組合せをとった場合の PR 曲線を比較している．約 0.35 までの再現率が低い区間では if が，約 0.85 までの再現率の中間では sf が，残りの再現率が高い区間では df が最良の適合率である．手法によって PR 曲線の特徴が分かれており，各々の良い特徴を保持するように組み合わせられれば，全区間に

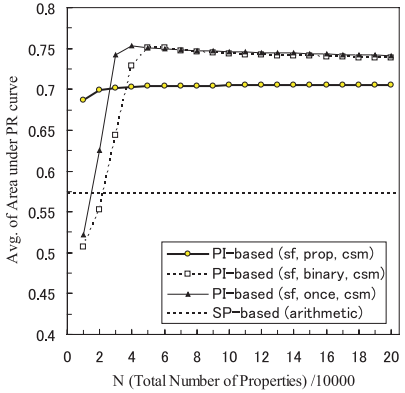


図 13　補完類似度による下位概念抽出における全性質数 N 依存性

Fig. 13　Average of AuPR for total number of properties N in PI-based hyponym extraction using complementary similarity measure.

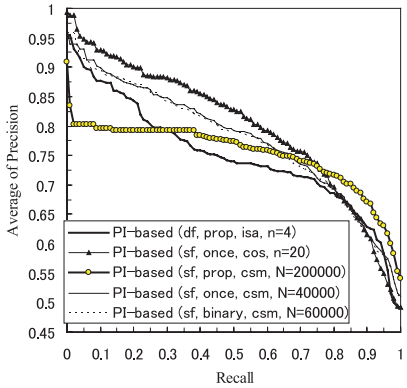


図 14　補完類似度による下位概念抽出の平均 PR 曲線

Fig. 14　Average of PR curve by PI-based hyponym extraction using complementary similarity measure.

わたって最良な PR 曲線を実現できる可能性がある．オブジェクト名サーチを実現するため，下位概念抽出において再現率を重視しつつ適合率も高く維持したいという目的からすれば，Web 文書中における “c の p” の頻度（df）を，複数回共起することを重要視した比率（prop）で統合し，性質継承度（isa-PI）で上位下位関係を評価する提案手法が最も優れている．また，df のベストな組合せでは評価に必要な性質数（検索エンジンへの問合せ回数）が少なく，計算時間と実現精度というコスト・パフォーマンスの観点からも優れている．

図 13 は，スニペット中における近接共起度（sf）をもとにして上位下位関係を補完類似度（csm）で評価する手法の PR 曲線下面積を比較している．ここで，補完類似度を計算するためには，対象概念 x も下位概念候補 y も持たない性質の数 d が必要である．したがって，各々の最大 1,000 件スニペット中での共起状況に加えて，あらゆる性質（語）の総数 N が必要であるが，明ではないために依存性を調べている．**図 14** は，補完類似度による下位概念抽出に関して，sf を統合する手法ごとにベストな組合せをとった場合の PR 曲線を比較しており，sf を比率（prop）で統合し補完類似度で評価した方が，再現率が高い区間で最良の適合率となっている．

表 2，**表 3** は，12 種類の対象概念の典型的な性質の上位 n 件を抽出した結果を共起頻度を求める 3 種類の手法ごとに示している．sf と比べ，df や if では部分全体（has-a）関係や振舞い表現といった性質を表す語をより多く抽出できており，歴史漫画の作品名である「花の慶次」は明らかにノイズであるが他はあまり目立たない．また，df や if による性質抽出の結果は互いに類似しており，性質継承に基づく下位概念抽出の PR 曲線下面積の性質数 n 依存性に関する図 9 および図 10 を参照すると，比率（prop）によって統合し性質継承度で評価する場合を除いて，類似した変化を示している．比率（prop）によって統合した場合に異なる変化を示すのは，df や if で抽出された性質集合は互いに類似していても，各性質の占める割合が df と if とでは異なるためである．一方，sf では，対象概念の上位概念や同位（等）概念，下位概念を表す語が多く含まれており，性質抽出としての精度は良くないが，これらの語は上位下位関係を評価する際には有効に働いている．なぜならば，下位概念は対象概念から部分全体（has-a）関係だけでなく，4 章で述べたように上位（hypernym）関係や同位（coordinate）関係，そして，下位（hyponym）関係の一部も継承するため，これらの語が典型的な性質の上位 n 件の集合に含まれているだけで，明に周辺概念を考慮するように改良した性質継承度の関数 isa-PIH/$\mathrm{C}^*_n(y,x)$ を使わなくても暗に考慮する形になっていると考えられる．

表 4，**表 5** は，共起頻度を求める 3 種類の手法ごとにベストな組合わせをとった場合の性質継承に基づく下位概念抽出によって，各対象概念の下位概念候補集合をランキングした結果の上位 k 件を示している．著者自身が適合解と判断した候補は「**太字**」に，また，EDR 概念辞書で対象概念の下位概念として登録されていた候補には「ϵ」を付している．各々の

適合率・再現率（PR）曲線の平均を比較した図 12 から考察されたとおり，再現率が低い区間である上位数件においては，if や sf と比べて，df の精度は悪く，対象概念の上位概念や同位概念がノイズとして現れている．また，df では，EDR 辞書に登録されていない下位概念が上位数件にランキングされており，一方，if や sf では EDR 辞書に登録されている下位概念が上位数件に多くランキングされている傾向が観察できる．これらの違いは典型的な性質の件数 n に依存しており，n が小さいほど EDR 辞書に登録されていないマイナーな下位概念も洩らさず，n が大きくなるほど EDR 辞書に登録されているメジャーな下位概念が上位を占めてしまいマイナーな下位概念が洩れやすくなると考えられる．

6.4　周辺概念として上位概念や同位概念との性質継承も考慮した下位概念抽出の検証

12 種類の対象概念 x に対して，対象概念 x から下位概念候補 y への性質継承の度合いだけでなく，対象概念 x の上位概念 x_h から下位概念候補 y への性質継承の度合い，あるいは，対象概念 x から下位概念候補 y の同位概念 y_c への性質継承の度合いといった周辺概念も考慮した下位概念抽出手法を適用した．

まず，対象概念 x から下位概念候補 y への性質継承の度合いだけでなく，対象概念 x の上位概念 x_h から下位概念候補 y への性質継承の度合いも考慮した下位概念抽出手法（4 章で定義した isa-PIH$^*_n(y,x)$）について考察する．**図 15** は，2 つの性質継承の度合いを線形結合するためのパラメータ α を 0.0 から 1.0 まで動かし，各 α ごとに $n \in \{1,...,20\}$ で下位概念抽出した 20 個の PR 曲線下面積の平均の変化を表している．$\alpha = 0.0$ のとき，対象概念 x の上位概念 x_h から下位概念候補 y への性質継承の度合いはまったく考慮されず，対象概念 x から下位概念候補 y への性質継承の度合いだけに基づく基本的な下位概念抽出そのものである場合に最良値である．α をより大きくし，対象概念 x の上位概念 x_h から下位概念候補 y への性質継承の度合いを考慮してゆくほど，精度が悪化してしまっており，対象概念の上位概念を考慮することは有効でないことが分かった．「昆虫」「魚」の上位概念として適合する「変温動物」が抽出されたが，「変温動物」の下位概念には他に「爬虫類」「両生類」やその下位概念が多数あり，これらをノイズとしてすくったり，「鳥」に対して「恐竜」，「花」に対して「生殖器」という誤った上位概念も抽出されたりしていたことが原因である．

次に，対象概念 x から下位概念候補 y への性質継承の度合いだけでなく，対象概念 x から下位概念候補 y の同位概念 y_c への性質継承の度合いも考慮した下位概念抽出手法（4 章で定義した isa-PIC$^*_n(y,x)$）について考察する．**図 16** は，2 つの性質継承の度合いを線形結合するためのパラメータ β を 0.0 から 1.0 まで動かし，各 β ごとに $n \in \{1,...,20\}$ で下位概念抽出した 20 個の PR 曲線下面積の平均の変化を表している．$\beta = 0.0$ のとき，対象

表 2　各対象概念 x の典型的な性質の上位 n 件
Table 2　Top n typical properties of concept x.

x = 鳥			
n	df	if	sf
1	巣	声	鳥インフルエンザ
2	声	巣	感染
3	さえずり	さえずり	情報
4	鳴き声	鳴き声	日本
5	詩	羽	人
6	夏	餌	TEL
7	写真	唐揚げ	鶏肉
8	羽	丸焼き	鳥料理
9	歌	名前	焼き鳥
10	名前	糞	発生
x = 魚			
n	df	if	sf
1	骨	棚	販売
2	種類	フライ	新鮮
3	名前	種類	旬
4	写真	王様	海
5	活性	群れ	ホームページ
6	形	名前	魚料理
7	煮付け	すり身	鮮魚
8	味	餌	紹介
9	数	骨	中心
10	干物	水槽	味
x = 昆虫			
n	df	if	sf
1	写真	森	自然
2	森	世界	カブトムシ
3	世界	観察	クワガタ
4	生態	標本	世界
5	標本	生態	紹介
6	幼虫	幼虫	身近
7	観察	家	生態
8	名前	名前	昆虫館
9	種類	体	昆虫類
10	体	種類	日本

x = 花			
n	df	if	sf
1	写真	名前	販売
2	色	色	花束
3	名所	都	アレンジメント
4	名前	形	紹介
5	香り	香り	フラワーギフト
6	慶次	写真館	季節
7	形	名所	写真
8	季節	山	あなた
9	種	図鑑	運営
10	名	寺	花屋
x = 野菜			
n	df	if	sf
1	サラダ	宅配	旬
2	苗	サラダ	果物
3	スープ	花	販売
4	栽培	スープ	有機野菜
5	味	苗	野菜ソムリエ
6	煮物	収穫	紹介
7	収穫	種	トマト
8	種	煮物	京野菜
9	販売	カレー	運営
10	甘み	栽培	新鮮
x = 果物			
n	df	if	sf
1	王様	王様	野菜
2	皮	花	販売
3	香り	女王	旬
4	女王	話	フルーツ
5	種類	木	季節
6	名前	皮	桃
7	木	生産	りんご
8	味	老舗	新鮮
9	摂取	味	さくらんぼ
10	栽培	香り	産地直送

表 3　各対象概念 x の典型的な性質の上位 n 件
Table 3　Top n typical properties of concept x.

x = 俳優			
n	df	if	sf
1	演技	仕事	プロフィール
2	名前	一人	公式サイト
3	一人	清水	映画
4	プロフィール	プロフィール	舞台
5	道	名前	声優
6	仕事	演技	日本
7	顔	イ	女優
8	方々	顔	タレント
9	出演	柳生	活躍
10	紹介	道	紹介

x = お笑い芸人			
n	df	if	sf
1	ネタ	ネタ	プロフィール
2	名前	合コン	お笑い
3	陣内智則	陣内智則	ブログ
4	動画	人	公式サイト
5	ブログ	道	芸人
6	大田	トークラジオ	掲示板
7	木村祐一	テレビ	ネタ
8	ラジオ	トーク	本人
9	合コン	劇団ひとり	今
10	話	皆さん	吉本興業

x = 漫画家			
n	df	if	sf
1	犬	先生	公式サイト
2	作品	松本零士	プロフィール
3	先生	卵	イラスト
4	アシスタント	作品	本人
5	卵	サイン	作品紹介
6	一人	本	作品リスト
7	人	私	日記
8	名前	水木しげる	掲示板
9	松本零士氏	一人	ブログ
10	場合	横山	ファンサイト

x = 家電			
n	df	if	sf
1	未来	タンタンショップ	販売
2	価格	レシピィ	家電リサイクル法
3	価格比較	販売	家電製品
4	タンタンショップ	買取	生活家電
5	安値屋本舗	専門店	パソコン
6	買取	紹介	テレビ
7	販売	デンマート	エアコン
8	普及	リサイクル	家電リサイクル
9	選び方	商品	冷蔵庫
10	購入	商品一覧	運営

x = 楽器			
n	df	if	sf
1	音	演奏	販売
2	演奏	専門店	音楽教室
3	音色	音	ピアノ
4	練習	ページ	ギター
5	種類	紹介	管楽器
6	販売	音色	楽譜
7	事	販売	修理
8	購入	練習	音楽
9	紹介	楽譜	中古楽器
10	話	説明	楽器販売

x = 乗り物			
n	df	if	sf
1	写真	話題	乗り物酔い
2	名前	一部	車
3	運転	旅	飛行機
4	絵	おもちゃ	バイク
5	おもちゃ	種類	バス
6	話	名前	旅
7	数々	一つ	その他
8	旅	運転	自転車
9	話題	本	自動車
10	絵本	写真素材	ここ

表 4　性質継承に基づいて順序付けられた下位概念候補の上位 k 件
Table 4　Top k hyponym candidates ranked by PI-based extraction.

x = 鳥			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	**雁** ϵ	**雁** ϵ	**家禽** ϵ
2	**ツバメ** ϵ	祖先神	**ニワトリ** ϵ
3	**カッコウ** ϵ	**ツバメ** ϵ	**トキ** ϵ
4	**鷲**	**鷲**	**ブロイラー** ϵ
5	**小鳥** ϵ	**カッコウ** ϵ	**ウズラ** ϵ
6	**燕**	**サイチョウ** ϵ	**ハト** ϵ
7	**ホトトギス** ϵ	**ジュウイチ**	身
8	**シロバト**	**コマドリ** ϵ	**カモ類**
9	**ジュウイチ**	**オオミズナギドリ** ϵ	**キジ** ϵ
10	**鷹**	**ヒバリ** ϵ	**ペリカン** ϵ

x = 魚			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	主要貨物	とのこ	**ひらめ**
2	馬	**グルクン**	**金目鯛**
3	武器	**マス類**	**いさき** ϵ
4	動物	**メジナ** ϵ	**アンコウ** ϵ
5	犬	**エソ** ϵ	**しらす**
6	**マンタ**	**白身魚** ϵ	**さわら**
7	料理	タンパク源	**鮃**
8	具	**ディスカス**	**鰆**
9	食べ物	**チョウザメ** ϵ	**青魚** ϵ
10	**淡水エイ**	**青魚** ϵ	肴

x = 昆虫			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	**ハマヒョウタンゴミムシダマシ**	**かぶと虫**	**クワガタムシ** ϵ
2	**ガロアムシ** ϵ	**クロオオアリ** ϵ	**カミキリムシ** ϵ
3	**シマアカネ**	**カブト虫**	**タマムシ** ϵ
4	**エゾカミキリ**	**トビゲラ**	**カマキリ** ϵ
5	**ルイスハンミョウ**	**タイコウチ** ϵ	**ナミテントウ**
6	**ガガンボカゲロウ**	**サシガメ** ϵ	**タガメ**
7	花	**マツノマダラカミキリ**	**アリ類**
8	動物	**クサカゲロウ** ϵ	**タイコウチ** ϵ
9	**トラマルハナバチ**	**アオマツムシ**	**コオイムシ** ϵ
10	**コルリクワガタ**	**ユスリカ** ϵ	**かぶと虫**

x = 花			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	**桜** ϵ	**桜** ϵ	**カーネーション** ϵ
2	左	**ハナショウブ**	**胡蝶蘭**
3	様子	**スカーレット**	**バラ**
4	**ツツジ**	**はまなす**	百花
5	**ハナカツミ**	これら	**ヒマワリ**
6	**つつじ**	**県花** ϵ	**シクラメン**
7	右	**笹ゆり**	**テッポウユリ**
8	**ヒマワリ**	**ササユリ**	**ひまわり**
9	**コスモス**	**ひまわり**	**アマリリス**
10	**サクラ**	**テッポウユリ**	**ブーゲンビレア**

x = 野菜			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	**水菜**	**白オクラ**	**地場野菜**
2	**玉造黒門白瓜**	**貝割大根**	**加賀野菜**
3	パパイヤ	**万願寺甘とう**	**金時草**
4	**ミニトマト**	**赤なす**	**アシタバ** ϵ
5	**レタス** ϵ	**ソラマメ**	**ゴボウ** ϵ
6	**レタス系**	**アシタバ** ϵ	**みょうが**
7	**島かぼちゃ**	**水菜**	**根菜類** ϵ
8	**剣崎なんば**	**モロヘイヤ**	**じゃがいも**
9	**五郎島金時さつまいも**	**タアサイ**	**ほうれんそう**
10	**リーキ**	**リーキ**	**大根** ϵ

x = 果物			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	**ベルガモット**	**ラムヤイ**	**いちご**
2	**グレープフルーツ** ϵ	**桃** ϵ	**巨峰**
3	**レモン**	九州産	**キウイフルーツ**
4	**ミカン** ϵ	山形産	**梨**
5	**ラムヤイ**	**ベルガモット**	**渋柿**
6	**柑橘類**	**スターフルーツ**	**いちじく**
7	**バナナ** ϵ	**柿** ϵ	**デコポン**
8	**ライチ**	**晩白柚**	**柑橘類**
9	**マンゴー** ϵ	**渋柿**	**すいか**
10	**みかん** ϵ	**パパイア**	**ネーブル**

表 5　性質継承に基づいて順序付けられた下位概念候補の上位 k 件
Table 5　Top k hyponym candidates ranked by PI-based extractio.

x = 俳優			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	チャ・ホンニョ	吉田鋼太郎	中村獅童
2	チャン・チンホイ	大竹浩一	西村雅彦
3	ショーン・ペン	佐藤佐吉	長塚京三
4	ロビン・ウイリアムス	ユ・オソン	ゲスト
5	ユ・スンホ	パク・コニョン	阿部寛
6	李京源	トム・ホフマン	小栗旬
7	マイケル・ケーン	キム・ヒョンジュ	東山紀之
8	Ethan Hawke	米倉斉加年	中村俊介
9	ムン・ソリ	アン・ソンギ	田辺誠一
10	キム・サンギョン	パク・チュンフン	金城武

x = お笑い芸人			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	笑い飯さん	有野氏	2 丁拳銃
2	波田陽区さん	若井おさむ	スマイリーキクチ
3	タカアンドトシさん	松本美香	吉本興業所属
4	我が家さん	小島よしおさん	鳥居みゆき
5	小島よしお	インスタントジョンソン	アンガールズ
6	ホーキング青山氏	バカリズム	南海キャンディーズ
7	主人公	小島よしお	ココリコ
8	藤崎マーケット	陣内智則	松本人志
9	バカリズム	鳥居みゆき	笑い飯
10	柳原可奈子	柳原可奈子	小島よしお

x = 漫画家			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	ラズウェル氏	ラズウェル氏	西原理恵子
2	木戸あいらく	安部慎一	野間美由紀
3	キム・オクスン	語シスコ	伊藤理佐
4	志水圭	道原かつみ	和月伸宏
5	日本橋ヨヲコさん	雷句誠	作者
6	フレデリック・ボワレ	赤松健	わじゅん
7	山田花子先生	岡本太郎氏	麻生
8	中条智子	細川貂々	松本零士
9	ツージィQ 氏	高橋ヒロシ	井上
10	李志清	杉浦茂	高橋留美子

x = 家電			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	Miele	DVD プレーヤー	4 品目
2	LF-PK1	LF-PK1	情報家電 $^{\epsilon}$
3	ゲームコンソール	カラーテレビ	白物家電 $^{\epsilon}$
4	液晶テレビ	商品	デバイスタイル
5	パソコン	ネット家電	コンポ
6	高精細液晶テレビ	加湿器	洗濯乾燥機
7	ネットワーク対応家電	ビデオデッキ	プラズマテレビ
8	商品	シェーバー	ネット家電
9	HDD ビデオレコーダ	ゲーム機	エアコン
10	ブルーレイ DVD	プラズマテレビ	家庭用

x = 楽器			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	口琴テミルコムズ	フィーデル	電子ピアノ
2	フールースー	編鐘 $^{\epsilon}$	金管楽器 $^{\epsilon}$
3	笛 $^{\epsilon}$	リードオルガン $^{\epsilon}$	和楽器
4	ハープシーコード	馬頭琴	打楽器 $^{\epsilon}$
5	シンギング・リン	大正琴 $^{\epsilon}$	エレキベース
6	編鐘 $^{\epsilon}$	日本胡弓	大正琴 $^{\epsilon}$
7	胡弓 $^{\epsilon}$	薩摩琵琶 $^{\epsilon}$	鍵盤楽器 $^{\epsilon}$
8	馬頭琴	箏	管楽器 $^{\epsilon}$
9	琴 $^{\epsilon}$	ガイタ	弦楽器 $^{\epsilon}$
10	和太鼓	ブズーキ	ピアノ $^{\epsilon}$

x = 乗り物			
k	df, prop, isa-PI, n:4	if, prop, isa-PI, n:19	sf, once, cos, n:20
1	原動機付自転車	チャリンコ	トゥクトゥク
2	ロバ車	気球船	自転車タクシー
3	AT 上	龍馬号	リニアモーターカー $^{\epsilon}$
4	機関車	列車 $^{\epsilon}$	ケーブルカー $^{\epsilon}$
5	車	御料車 $^{\epsilon}$	新幹線 $^{\epsilon}$
6	馬	宇宙戦艦	牛車 $^{\epsilon}$
7	列車 $^{\epsilon}$	鉄道 $^{\epsilon}$	飛行機 $^{\epsilon}$
8	動物達	チャリ	トラム
9	自動車 $^{\epsilon}$	ケムケム	機関車
10	三輪車 $^{\epsilon}$	本アイテム	列車 $^{\epsilon}$

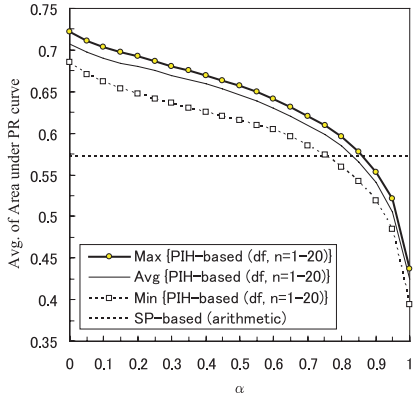


図 15　上位概念も考慮した性質継承に基づく下位概念抽出における線形結合パラメータ α 依存性
Fig. 15　Average of AuPR for parameter α in PIH-based hyponym extraction.

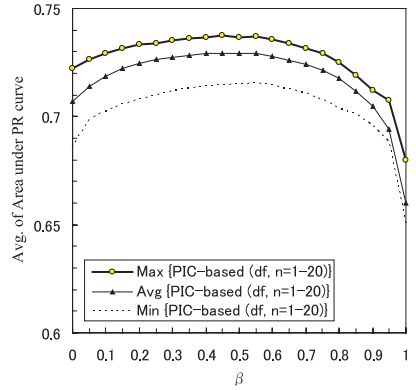


図 16　同位概念も考慮した性質継承に基づく下位概念抽出における線形結合パラメータ β 依存性
Fig. 16　Average of AuPR for parameter β in PIC-based hyponym extraction.

概念 x から下位概念候補 y の同位概念 y_c への性質継承の度合いはまったく考慮されず，対象概念 x から下位概念候補 y への性質継承の度合いだけに基づく基本的な下位概念抽出そのものである．$\beta = 0.5$ の近辺でベストになっており，対象概念 x から下位概念候補 y への性質継承の度合いだけでなく，対象概念 x から下位概念候補 y の同位概念 y_c への性質継承の度合いも同等に考慮した場合に最も精度が良くなっている．また，$\beta = 1.0$ のとき，対象

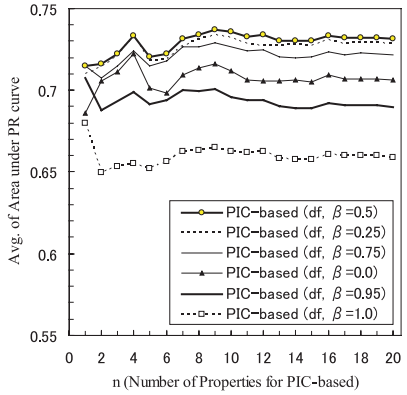


図 17　同位概念も考慮した性質継承に基づく下位概念抽出における典型的な性質の数 n 依存性
Fig. 17　Average of AuPR for number of properties n in PIC-based hyponym extraction.

概念 x から下位概念候補 y の同位概念 y_c への性質継承の度合いだけに基づいた下位概念抽出の方が，$\beta = 0.0$ のとき，対象概念 x から下位概念候補 y への性質継承の度合いだけに基づいた下位概念抽出よりも約 0.04 ほど下がっている．これは，下位概念候補の同位概念を Web から抽出する精度が完全ではないためと考えられる．具体的には，厳密すぎる構文パターンを用いたため，同位概念が 1 つも抽出されないケースがあった．

図 17 は，代表的な β に対して，性質継承の度合いの評価に用いる典型的な性質の数 n による PR 曲線下面積の変化を表している．対象概念 x から下位概念候補 y の同位概念 y_c への性質継承だけに基づいた下位概念抽出では $n = 4$ の場合に最良値 0.722 となるが，同位概念への性質継承も同等に（$\beta = 0.5$）考慮した下位概念抽出では $n = 9$ の場合に最良値 0.737 となり，同位概念も考慮することで改善されている．

図 18 は，同位概念を考慮するか否かによる PR 曲線を比較している．概念 c と 1 回でも近接共起することを重要視して重み付けした特徴語 p による擬似コサイン相関値 $\cos_n^{s,o}(y, x)$ に基づいて上位下位関係を評価する従来手法に対して，Web 文書中において “c の p” という構文パターンが複数回出現することを重要視して重み付けした性質語 p による性質継承度 isa-PI$_n^{d,p}(y, x)$ に基づいて上位下位関係を評価する提案手法によって，さらに，これら両手法に対して，周辺概念として同位概念も考慮した性質継承度 isa-PIC$_n^{d,p}(y, x)$ に基づいて上位下位関係を評価する改良手法によって，本論文の最大の目的である再現率の高い区間での適合率を改善することに確かに成功している．

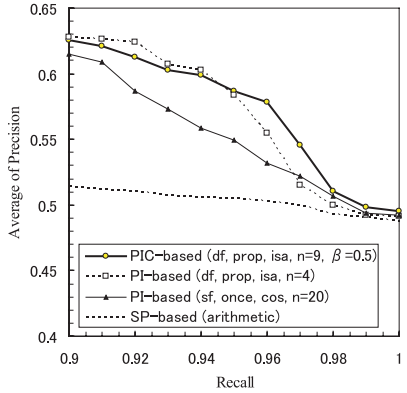


図 18　同位概念を考慮するか否かによる平均 PR 曲線の比較
Fig. 18　Average PR curve by PI-based vs. PIC-based hyponym extraction.

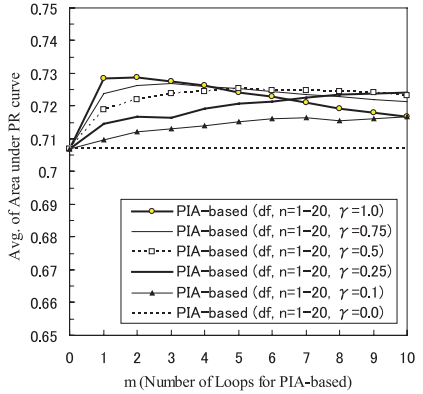


図 19　性質集約に基づく下位概念抽出におけるループ数 m 依存性
Fig. 19　Average of AuPR for number of loops m in PIA-based hyponym extraction.

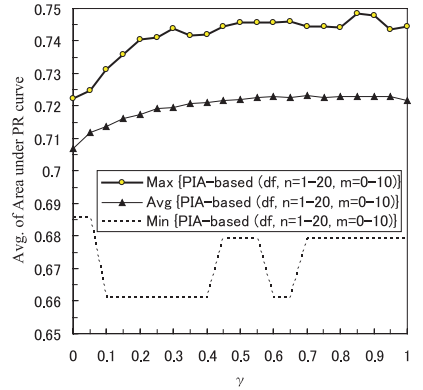


図 20　再帰的な性質集約を行う性質集約に基づく下位概念抽出における線形結合パラメータ γ 依存性
Fig. 20　Average of AuPR curve for parameter γ in PIA-based hyponym extraction.

6.5　性質集約に基づく下位概念抽出の評価

性質集約に基づく下位概念抽出では，まず，基本的な性質継承に基づく下位概念抽出によって各下位概念候補に重み付けしたうえで，その重みに応じて下位概念候補集合から対象概念へ性質集約することにより，対象概念の典型的な性質の重みが再帰的に修正され，対象概念から下位概念候補への性質継承の度合いも再帰的に修正されてゆく．これまでと同じ 12 種類の対象概念 x に対して，性質集約に基づく下位概念抽出を適用し，性質集約を制約条件として追加することによって，性質継承に基づく下位概念抽出の精度が改善されるか否かを検証する．

図 19 は，対象概念に対する典型的な性質としての相応度を再計算する式中における $m-1$ 次の相応度との結合パラメータ γ を固定してループさせた場合に，各ループ数 $m \in \{0, ..., 10\}$ ごとに性質数 $n \in \{1, ..., 20\}$ で下位概念抽出した 20 個の PR 曲線下面積の平均の変化を表している．$m = 0$ の場合の結果は，性質集約を用いない，性質継承に基づく基本的な下位概念抽出とまったく同じである．ループ数 m が増加するほど単調に精度が良くなるということではなく，あるループ数で最良値をとり，それ以降もループを続けてしまうと精度が減衰してゆく傾向が観察できる．また，$\gamma = 1.0$ では $m = 1$ 次に最良値 0.728 をとり，一方，$\gamma = 0.5$ では $m = 5$ 次に最良値 0.725 をとっており，結合パラメータ γ を大きくするほど早いループ数 m で（より大きな）最良値をとっていることも分かる．

図 20 は，結合パラメータ γ を 0.0 から 1.0 まで動かし，各 γ ごとに性質数 $n \in \{1, ..., 20\}$ およびループ数 $m \in \{0, ..., 10\}$ で下位概念抽出した 220 個の PR 曲線下面積の平均の変化を表している．全体では $n = 4$，$m = 7$，$\gamma = 0.85$ のとき最良値 0.749 を記録している．対象概念から下位概念候補への性質継承だけに基づく基本的な下位概念抽出の最良値が 0.722 で，対象概念から下位概念候補の同位概念への性質継承も考慮した下位概念抽出の精度の最大値が 0.737 であり，いずれよりも精度が改善されている．

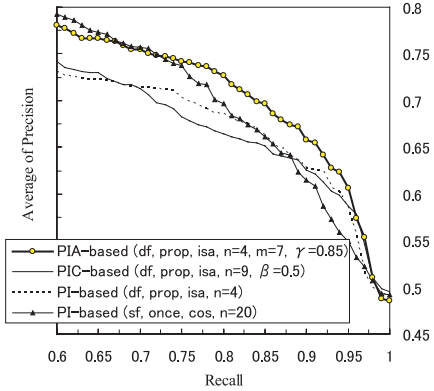


図 21　再帰的な性質集約を行うか否かによる平均 PR 曲線の比較
Fig. 21　Average PR curve by PI-based vs. PIA-based hyponym extraction.

図 21 は，再帰的な性質集約を行うか否かによる PR 曲線を比較している．1 回でも近接共起することを重要視して重み付けした特徴語による擬似コサイン相関値に基づいて上位下位関係を評価する従来手法に対して，再帰的な性質集約を行わない性質継承に基づく基本的な下位概念抽出ですでに約 0.85 以上の再現率が高い区間で適合率を上回っていたが，再帰的な性質集約を行うことによって約 0.70 以上の再現率が高い区間で適合率を上回ることができるように改善されている．

7. おわりに

概念階層に関する知識は，我々が目指している五感情報に基づくオブジェクト名サーチだけでなく，様々な自然言語処理システムにとって非常に重要な基本的知識である．従来の抽出手法の多くは構文パターンに基づいているため，上位下位関係の十分条件である厳密な構文パターンを用いると適合率は高いが再現率が非常に低くなり，逆に，曖昧な構文パターンを用いると再現率は高くなるがノイズばかりになり適合率が非常に低くなってしまうという問題があった．特にオブジェクト名サーチというアプリケーションにおいては，対象概念の下位概念をできる限り網羅的に精度良く抽出できる機能が不可欠であるため，再現率が高い区間での適合率を改善する必要がある．これに対して我々は，上位下位関係の構文パターンに合致する文書頻度とは異なる評価軸として，対象概念から下位概念候補への性質継承の度合いに基づく抽出手法を提案した．概念間の類似度計算において一般的に用いられる単なる特徴語ではなく，各概念の部分全体（has-a）関係や振舞い表現といった性質語に限定する点が従来手法とは異なる．さらに，2 つの概念間の直接的な関係を評価するだけでなく，対象概念の上位概念から下位概念候補への性質継承や，対象概念から下位概念候補の同位概念への性質継承といった，周辺にある概念との関係も考慮することによって，提案手法のロバスト性の向上を図った．また，各概念の典型的な性質を抽出する手法においても，各概念と各性質との間の直接的な関係を評価するだけでなく，対象概念の上位概念からの性質継承や対象概念の下位概念集合からの性質集約も考慮することによって改善を図った．

評価実験の結果，約 0.35 までの再現率が低い区間では，画像の周辺テキストにおいて構文パターン “c の p” が複数回出現することを重要視して重み付けした各概念 c の典型的な性質語 p による性質継承度に基づいて上位下位関係を評価する提案手法が最良の適合率となり，約 0.85 までの再現率の中間では，概念 c と 1 回でも近接共起することを重要視して重み付けした特徴語 p による擬似コサイン相関値に基づいて上位下位関係を評価する従来手法が最良の適合率となった．本論文での最大の改善対象である残りの再現率が高い区間では，Web 文書中において構文パターン “c の p” が複数回出現することを重要視して重み付けした各概念 c の典型的な性質語 p による性質継承度に基づいて上位下位関係を評価する提案手法が最良の適合率となった．したがって，概念間の類似度計算に準ずる従来の下位概念抽出手法の改善の余地であった再現率が低い区間および再現率が高い区間における適合率の改善に対して，一般的な特徴語ではなく性質語に限定して概念間の性質継承の度合いを計算し，上位下位関係の有無の評価尺度とする提案手法が有効であることが確認された．つまり，概念間の上位下位関係を精度良く網羅的に抽出する（再現率が高い区間での適合率を改善する）ために不可欠な必要（十分）条件として，「概念間に上位下位関係があれば互いに類似した特徴語を持つ」という既存の仮説では弱く，「概念間に上位下位関係があれば上位概念の性質語のすべてを下位概念が継承する」という我々の仮説がより適しているといえる．

周辺概念として対象概念の上位概念も考慮してしまうと精度が悪化してしまったが，一方，周辺概念として下位概念候補の同位概念を考慮することによって，再現率が高い区間での適合率をわずかではあるがさらに改善できることも確認された．

最後に，下位概念候補集合から対象概念に対して再帰的に性質集約することによって，概念間の類似度計算に準ずる従来の下位概念抽出手法よりも適合率を高く保つことが可能な再現率が高い区間の幅を，約 0.85 以上から約 0.70 以上へと大幅に改善できた．

今後の研究課題としては，本論文では周辺概念として，対象概念の上位概念や，下位概念候補の同位概念など，対象概念および下位概念候補と直接的な関係が見出されている概念だ

けを利用したが，対象概念の上位概念の上位概念や，対象概念の同位概念の同位概念など，対象概念および下位概念候補と間接的に関係がある概念の活用も考えられる．また，提案手法によって抽出された対象概念（クラス名）の下位概念（具体的なオブジェクト名）集合を基本的知識として利用した五感情報に基づくオブジェクト名サーチの開発も行ってゆく．

謝辞　本研究は，科学研究費補助金特別研究員奨励費「モバイル・ユビキタス環境における空間情報アクセスに関する研究」（研究代表者：服部峻，課題番号：1955301，平成 19〜20 年度），および，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」（研究代表者：田中克己，平成 19〜23 年度），および，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）の助成を受けたものである．ここに記して謝意を表す．

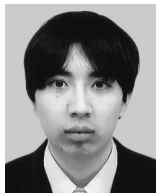
参　考　文　献

1) 服部　峻，田中克己：外観・状況表現を用いたオブジェクト名検索，iDB フォーラム 2008，情報処理学会研究報告「データベースシステム」，Vol.2008, No.88, pp.109–114 (2008).
2) Mandala, R., Tokunaga, T. and Tanaka, H.: The Use of WordNet in Information Retrieval, *Proc. COLING ACL Workshop on Usage of WordNet in Natural Language Processing*, pp.31–37 (1998).
3) Hattori, S., Tezuka, T. and Tanaka, K.: Activity-based Query Refinement for Context-aware Information Retrieval, *Proc. 9th International Conference on Asian Digital Libraries* (*ICADL'06*), LNCS, Vol.4312, pp.474–477 (2006).
4) Hattori, S., Tezuka, T., Hiroaki, O., Oyama, S., Kawamoto, J., Tajima, K. and Tanaka, K.: ReCQ: Real-world Context-aware Querying, *Proc. 6th International and Interdisciplinary Conference on Modeling and Using Context* (*CONTEXT'07*), LNAI, Vol.4635, pp.248–262 (2007).
5) Fleischman, M., Hovy, E. and Echihabi, A.: Offline Strategies for Online Question Answering:. Answering Questions Before They Are Asked, *Proc. 41st Annual Meeting of the Association for Computational Linguistics* (*ACL'03*), pp.1–7 (2003).
6) 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).
7) Hattori, S., Tezuka, T. and Tanaka, K.: Mining the Web for Appearance Description, *Proc. 18th International Conference on Database and Expert Systems Applications* (*DEXA*), LNCS, Vol.4653, pp.790–800 (2007).
8) 服部　峻，田中克己：コンテキストに依存する外観情報の Web からの抽出，電子情報通信学会第 19 回データ工学ワークショップ（DEWS'08）論文集，A2-1 (2008).
9) WordNet. http://wordnet.princeton.edu/
10) 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).
11) Wikipedia. http://www.wikipedia.org/
12) Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H. and Studer, R.: Semantic wikipedia, *Proc. 15th International Conference on World Wide Web* (*WWW'06*), pp.585–594 (2006).
13) 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).
14) 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).
15) Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, *Proc. 14th International Conference on Computational Linguistics* (*COLING'92*), Vol.2, pp.539–545 (1992).
16) Caraballo, S.A.: Automatic Construction of a Hypernym-Labeled Noun Hierarchy from Text, *Proc. 37th Annual Meeting of the Association for Computational Linguistics* (*ACL'99*), pp.120–126 (1999).
17) 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).
18) Emmanuel, M. and Christian, J.: Automatic Acquisition and Expansion of Hypernym Links, *Computer and the Humanities*, Vol.38, No.4, pp.363–396 (2004).
19) 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).
20) 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).
21) 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).
22) 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).
23) 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).
24) Sanderson, M. and Croft, B.: Deriving Concept Hierarchies from Text, *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.206–213 (1999).

25) Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A.: WebScale Information Extraction in KnowItAll (Preliminary Results), *Proc. 13th International World Wide Web Conference* (*WWW'04*), pp.100–110 (2004).
26) Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M. and Etzioni, O.: Open Information Extraction from the Web, *Proc. 20th International Joint Conference on Artificial Intelligence* (*IJCAI'07*), pp.2670–2676 (2007).
27) 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).
28) 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).
29) 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).
30) 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).
31) 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).
32) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol.16, No.1, pp.22–29 (1990).
33) Ghahramani, Z. and Heller, K.: Bayesian Sets, *Advances in Neural Information Processing Systems 18* (*NIPS'05*), pp.435–442 (2006).
34) Lin, D.: Automatic Retrieval and Clustering of Similar Words, *Proc. 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics* (*COLING/ACL'98*), pp.768–774 (1998).
35) 関恒　仁，嶋田和孝，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).
36) 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).
37) Sundblad, H.: Automatic Acquisition of Hyponyms and Meronyms from Question Corpora, *Proc. ECAI'02 Workshop on Natural Language Processing and Machine Learning for Ontology Engineering*, (2002).
38) Yahoo!ウェブ検索 API. http://api.search.yahoo.co.jp/WebSearchService/V1/webSearch
39) 野田武史，大島裕明，小山　聡，田島敬史，田中克己：主題語からの話題語自動抽出とこれに基づく Web 情報検索，日本データベース学会 Letters，Vol.5, No.2, pp.69–72 (2006).
40) 服部　峻，手塚太郎，田中克己：文書中の地物画像を言語的記述で代替するための地物の外観情報の Web からの抽出，情報処理学会論文誌：データベース，Vol.48, No.SIG11(TOD34), pp.69–82 (2007).
41) Yahoo!画像検索 API. http://api.search.yahoo.co.jp/ImageSearchService/V1/imageSearch
42) EDR 電子化辞書．http://www2.nict.go.jp/r/r312/EDR/J_index.html
43) 舟橋卓也，上田高徳，平手勇宇，山名早人：商用検索エンジンの検索結果では取得できないランキング下位部分の収集・解析，電子情報通信学会第 19 回データ工学ワークショップ（DEWS'08）論文集，A2-5 (2008).
44) Davis, J. and Goadrich, M.: The Relationship between Precision-Recall and ROC curves, *Proc. 23rd ACM International Conference on Machine Learning* (*ICML'06*), pp.233–240 (2006).

(平成 20 年 6 月 20 日受付)
(平成 20 年 10 月 11 日採録)

(担当編集委員　今村　誠)



服部　峻（学生会員）

2004 年京都大学工学部情報学科卒業．2006 年同大学大学院情報学研究科社会情報学専攻修士課程修了．同年同博士後期課程入学後，2007 年より日本学術振興会特別研究員 DC2．主にユビキタス社会の情報アクセス技術の研究に従事．電子情報通信学会，日本データベース学会各学生会員．

田中　克己（正会員）

1974 年京都大学工学部情報工学科卒業．1976 年同大学大学院修士課程修了．1979 年神戸大学教養部助手．1986 年同大学工学部助教授．1994 年同大学工学部教授（情報知能工学専攻）．1995 年同大学大学院自然科学研究科情報メディア科学専攻専任教授．2001 年京都大学大学院情報学研究科社会情報学専攻教授，現在に至る．工学博士．主にデータベースとマルチメディア情報システムの研究に従事．人工知能学会，日本ソフトウェア科学会，IEEE Computer Society，ACM 等各会員．