論 文

Technical Papers

Webからの属性情報記述ページの発見

Finding Specification Pages from the Web

吉永 直樹
Naoki Yoshinaga
日本学術振興会
Japan Society for the Promotion Science
n-yoshi at jaist.ac.jp, http://www.jaist.ac.jp/~n-yoshi/

鳥澤 健太郎
Kentaro Torisawa
北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
torisawa at jaist.ac.jp, http://www-tori.jaist.ac.jp:8000/

keywords: specification finding, Web search, attribute acquisition

Summary

This paper presents a method of finding a specification page on the Web for a given object (*e.g.*, "Ch. d'Yquem") and its class label (*e.g.*, "wine"). A specification page for an object is a Web page which gives concise attribute-value information about the object (*e.g.*, "county"-"Sauternes") in well formatted structures. A simple unsupervised method using layout and symbolic decoration cues was applied to a large number of the Web pages to acquire candidate attributes for each class (*e.g.*, "county" for a class "wine"). We then filter out irrelevant words from the putative attributes through an author-aware scoring function that we called *site frequency*. We used the acquired attributes to select a representative specification page for a given object from the Web pages retrieved by a normal search engine. Experimental results revealed that our system greatly outperformed the normal search engine in terms of this specification retrieval.

1. はじめに

本稿では，対象物名（例: Ch. d'Yquem）とそのクラス名（上位語，例: ワイン）を入力とし，対象物の**属性**の情報を，表や箇条書きなどの視覚的に認知し易い形で記述したページ（以下，**属性情報記述ページ**）を発見する手法を提案する．ここで属性とは，人が知りたい対象物の側面（例えばワインであれば,「造られた土地」や,「原料の葡萄の品種」）のことであり，文書中では具体的な**属性語**（例:「生産地」,「葡萄品種」）によって参照される．属性情報記述ページは，図 1 左の例のように可読性に優れる上に情報の密度が高く，対象物に関する詳細な情報を効率的に得ることができる．

tf-idf [Salton 88] や PageRank [Page 98] などの汎用的なランキング尺度に基づく検索エンジンでは，必ずしも属性情報記述ページが検索結果の上位にくるわけではない．例えば，ワインの名前をクエリとして検索したとき検索結果の上位にくるページは，図 1 右のようなワインの批評家によって書かれた冗長なコメントや歴史を綴ったページである場合も多く，そこから属性の情報を入手するには読解に時間をかけなければならない．また，仮に，検索エンジンにより属性-属性値の情報を含むページが多数得られたとしても，汎用的なランキング尺度ではページ中の属性-属性値の数は通常考慮されないため，得

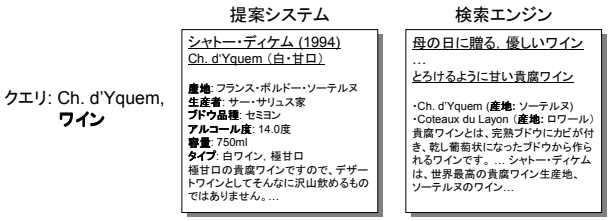

図 1 クエリ「Ch. d'Yquem, ワイン」に対する属性情報記述ページと汎用的な検索エンジンの出力

られたページから一ページだけを読んでも知りたい属性の情報が全て得られるとは限らないし，またページ中での属性情報の書かれ方も様々である．この結果，対象物について属性に関する情報を得たいユーザは複数のページを横断的に読み，散在する断片的な属性情報を統合する必要がある．これは，検索に不慣れな初心者の Web ユーザ，また携帯電話など表示領域の制限されたモバイル端末を用いる Web ユーザにとっては深刻な問題となる．その一方で，我々の提案するシステムでは，対象物名とそのクラス名が入力されると，多くの人が関心を持つという意味で**典型的**な対象物の属性の情報を数多く視覚的に分かりやすい形で含むページを出力する．そのためユーザは提案システムを用いることで，自身の知りたい情報に瞬時に到達できることが期待できるようになる．

これまで，ユーザが対象物に関して詳しい情報を得た

い場合，対象物名に加えて知りたい属性を指し示す属性語（映画であれば,「出演」など）をクエリに足して検索することが一般的であった．しかしながら，同じ属性を指す様々な言いまわし（例: 出演，配役，キャスト）があるため，この検索方法では検索漏れが生じる．また，知りたい情報が Web でどのような属性語により参照されるか思いつかないということもよくある．我々の狙いは，与えられた対象物に対して属性語を陽に指定することなく，最も多くの典型的な属性語を含む属性情報記述ページを発見することで，対象物に関する様々な属性-属性値をユーザへと提供することである．これによりユーザの労力は大幅に削減できる．

我々のシステムは，対象物の属性情報記述ページを，各クラスに対する典型的な属性語の集合（**クラスの属性知識ベース**）に基づき発見する．クラスの属性知識ベースは，前もって，属性情報記述ページにおける属性語の現れ方を考慮した，単純で一般的な教師なし学習により構築しておく．この構築の際には，属性語の候補から属性語として不適切な単語を取り除き，より典型的な属性語を得るために，Web ページ製作者の存在を考慮した**サイト頻度**を用いる．実際にユーザから入力が与えられると，システムはまず，汎用的な検索エンジンを用いて対象物を記述した Web ページを絞り込み，続いてクラスの属性知識を用いて各ページをスコア付けし，典型的な属性の情報の多さという観点から最良のページを発見する．

実験では,（HTML タグを含む）0.7TB の日本語 Web 文書を収集し，そこからクラスの属性知識ベースを構築した．次に，10 のクラスに対して，3 人の被験者に各クラスについて値を知りたい属性を挙げてもらい，各クラスに属する 10 の対象物について，我々のシステムが出力する Web ページに知りたい属性-属性値の情報が，視覚的に分かりやすい形で何対含まれているかを調べてもらった．実験結果から，我々のシステムが出力した Web ページには，Google の検索結果のトップページと比べ，より多くの属性-属性値関係が含まれていることが分かり，提案手法の有効性が示された．

2. 関 連 研 究

2·1 属性知識の自動獲得

本節では，既存の属性知識の自動獲得手法について説明する．属性知識を獲得するためには，これまで 2 種類の手がかりが用いられてきた．まず，自然文を知識源として，対象物とその属性語が共起する構文パターン（例: “the * of the x is”）を用いて属性語を獲得する手法が提案されている [Almuhareb 04, 高橋 05, Tokunaga 05]．しかしながら，属性情報記述ページで用いられる属性語の中には，いわゆる「平書き」の自然文中には出現し難いものが多くあり，自然文から獲得される属性語では不十分である．この点については実験結果を通して検証する．

また，特に Web を知識源として用いる際には，表形式や箇条書きなどの**レイアウト**が，自然文より直接的に属性-属性値を記述する手段として用いられるため注目されている [Chen 00, Yoshida 01]．Chen らは，単一の表形式を入力とし，表形式中のセル間の類似度を用いて属性-属性値の関係を認識する手法を提案している [Chen 00]．また Yoshida らは，表形式の集合を入力とし，各表形式中の属性-属性値の関係を EM 法に基づく教師なし学習により認識し，その後，各表形式中に含まれる属性語に基づき，同一のクラスに関して記述されている表形式（属性-属性値の集合）を獲得する手法を提案している [Yoshida 01]．表形式を手がかりにクラスの属性知識を獲得する際の問題点は，表形式中に必ずしも属性知識獲得対象のクラス名が含まれるとは限らない点である．従って具体的なクラスに関連づけられた属性知識を獲得するには，表形式から得られた属性-属性値に対する適切なクラス名を別途推定するか，具体的なクラスに関する表形式を人手で選別して入力として与える必要がある．

我々の提案する属性知識の獲得手法は，具体的なクラス名を入力とし，まず検索エンジンを用いてクラスの属性語を含み易いページを属性語獲得の知識源として収集する．次に，得られたページに含まれる表形式や箇条書きなどのレイアウトに注目し，特定の HTML タグや文字修飾に基づくパターンにより属性知識を獲得する．知識源のページを具体的なクラス名に対し前もって収集することで，既存手法で問題となっていた，クラスと属性知識の対応は自然にとることが可能である．また，知識源として収集したページ中には入力のクラスと無関係の内容を記述したレイアウトが含まれることがあるため，そのようなレイアウトからは誤った属性知識が獲得される恐れがある．そのため，我々は大量のページを処理して得られる統計値を用いて，パターンにより獲得された属性語候補をフィルタリングすることで，典型性と信頼性の高い属性知識を獲得する．

2·2 属性情報記述ページの発見

本節では，属性情報記述ページの発見に関する関連研究について紹介する．これまで，入力のクラスに属する任意の対象物の属性情報記述ページを収集する手法，および，入力の属性値（数値）に近い属性値を持つ対象物を検索する手法が提案されている．これらの手法は特定の対象物に対する属性情報記述ページを発見する手法ではないが，本研究と関連すると思われるのでここで述べる．

§1 入力のクラスの属性情報記述ページの発見

Yoshida, Nakagawa は，入力のクラスに属する任意の対象物の属性情報記述ページを収集する手法を提案している [Yoshida 05]．彼らの手法では，表形式から得られた属性-属性値の知識を元に，Web ページが入力のクラスに関連する確率を推定する．彼らは実験で，属性語および属性値が属性情報記述ページを発見するのに有用で

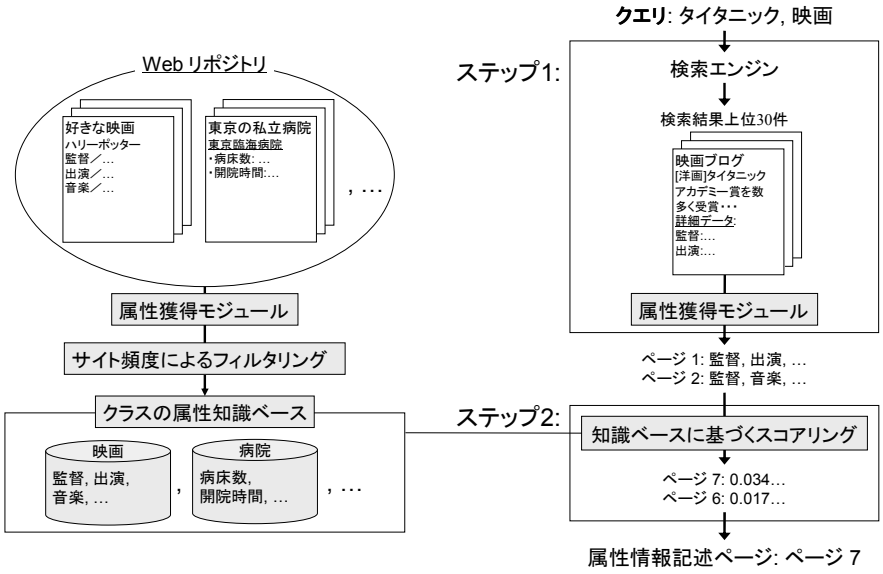

図 2 属性情報記述ページの発見: システム概観

あることを示したが，彼らの手法が前提とするクラスの属性-属性値は，前節で述べた手法 [Yoshida 01] で獲得されたものであり，入力のクラスに対して用いるべき属性-属性値の知識との対応を人手で取る必要があることから，一般ドメインへの拡張は困難であると予想される．

また Shimada, Endo は，クラスを商品に特化して，商品のスペック表を教師あり学習に基づき発見する手法を提案している [Shimada 05]．実験では 3 種類の商品に関して分類器を作成し，それぞれ高い精度でスペック表が得られることを示した．しかしながら，教師あり学習に基づく手法は人手で作成した正解付き学習データを必要とするため，一般ドメインに応用することは困難である．

我々は，検索エンジンを用いて収集したノイズを含む Web ページを，低コストの教師なし学習で大量に処理し，これに統計値を組み合わせて獲得される質の高い属性知識を用いるため，既存研究の本質的な問題であった，一般ドメインへの応用の問題を解決している．

§2 入力の数値に近い属性値を持つ対象物の検索

Agrawal, Ramakrishnan [Agrawal 02] は，属性値の多くが数値で表現されるクラス（例: 自動車，PC のパーツ）に焦点を絞り，対象物の属性値（数値）を入力として，その数値に近い属性値を持つ対象物の属性情報記述ページを検索する手法を提案している．彼らの狙いは，属性値をクエリとして検索することで，属性語の多様性の問題を回避することである．

彼らの手法は，多くの属性値が数値となるクラスにしか適用できない．また，彼らの設定するタスクは，例えば「排気量が 800cc 程度の自動車の詳細なデータが知りたい」といった，属性情報を得たい対象物の幾つかの属性値（数値）に関して大体の見当がついているユーザを想定している．そのため，本研究で対象に含む，対象物に関して属性値などの具体的な知識を持たないユーザに対しては，属性情報記述ページを提示することができない．

3. 提案システム

対象物の属性情報記述ページを発見するには，対象物の属性知識が重要な手がかりとなると考えられるが，あらゆる対象物について属性知識を獲得することは，その膨大さ，増進性から言って現実的ではない．そこで本研究では，対象物に比べ，文書中により頻繁に出現するクラス（上位語）の単位で属性知識を獲得する（図 2 左）．

対象物の属性には，クラスから継承する属性と対象物特有の属性があると考えられるが，このうち対象物特有の属性は，例えばデジタルカメラの新機能の有無（例: 初めて手振れ補正機構を導入したデジタルカメラの対象物に対する属性「手振れ補正機構の詳細」）のように例外的なもので，対象物の全属性に占める割合としては存在したとしてもごく一部と考えられる．そもそも，一つ／種類の対象，より正確には具体物だけが持つ特徴は,「備考」や「特徴」といった特別な属性の値として記述されることが多く，陽に「属性」として扱われることは少ない．逆に言えば,「属性」とは通常，多数の対象に共通の特性を整理するために導入されるものであると考えられ，一つ／種類の具体物だけに適用可能な属性という概念自体が考えづらい（従って，前の例であげた，初めて手振れ補正機構が導入されたデジタルカメラのスペック表に「手振れ補正機構の詳細」という属性が，実際には現れない可能性も高い．むしろ，そのような属性が現れやすいのは多数のデジタルカメラが手振れ補正機構を備えた後であろう）．従って本研究では,「もの」のような過度に抽象的な上位語をクラスとして与えない限り，対象物の属性セットとクラスの属性セットは同一視できるという立場に立って研究を進め，対象物を記述した Web ページから獲得される属性語候補と，クラスの属性語を比較することで，対象物の属性情報記述ページを発見する．

与えられた対象物とそのクラスに対し，我々は対象物の

HTMLソース

```
<HTML><BODY>
<H1>私の趣味</H1>
<H2>映画</H2>
好きな映画のリスト.<BR />
<H3>タイタニック（公開年: 1997)</H3>
監督/ James Cameron<BR />
出演/ Leonardo DiCaprio, Kate Winslet<BR />
詳細:<TABLE BORDER="1">
<TR><TD>ジャンル</TD><TD>ロマンス</TD> </TR>
<TR><TD>上映時間</TD><TD>194 min.</TD></TR>
<TR><TD>...</TD><TD>...</TD></TR>
</TABLE></BODY></HTML>
```

ブラウザ表示

私の趣味

映画

好きな映画のリスト.
タイタニック(公開年: 1997)
監督/ James Cameron
出演/ Leonardo DiCaprio, Kate Winslet
詳細:

ジャンル	Romance
上映時間	194 min.
...	...

図 3 属性語抽出の例

表 1 属性語獲得に用いた HTML タグと文字修飾

HTML タグ: TD, TH, LI, DT, DD, B, STRONG, FONT, SMALL, EM, TT
接頭修飾: ∗, ＊, ●, ○, ■, □, ・, ◆, ◇, ★, ☆, ◎, ・, ○ , ◎
括弧類: 〔-〕, 【-】, 《-》, ［-］, 〈-〉, <->, [-] <->
接尾修飾: ：, :, ／, /, ＝

属性情報記述ページを以下の手順で発見する（図 2 右）.

ステップ 1: 対象物名を含むページからの属性語抽出 入力された対象物名を含む Web ページを検索エンジンによって収集し，各ページから属性語を抽出する.

ステップ 2: 属性情報記述ページの発見 ステップ 1 で収集された各 Web ページから抽出された属性語とそのクラスの属性語を比較することで，最も属性情報の多い属性情報記述ページを発見する.

以下の節で，クラスの属性知識ベースの構築手法と，属性情報記述ページ発見手法について詳しくみていく.

3·1 クラスの属性知識ベースの構築

§1 知識源の Web ページのサンプリング

本研究では，クラスの属性知識を獲得する知識源として，クラスの属性語が多く含まれ易い，クラスをトピックとするページを利用する．具体的には，検索エンジンを用いてクラス名を含む文書を収集し，ページのトピックとなる表現が含まれやすい TITLE, H1 ～ H6, CAPTION, TD[*1], および TH タグでクラス名が囲まれているページを知識の獲得源とし，ページ中でクラス名が最初に現れた位置以降のテキストから属性語候補を獲得する.

§2 Web ページからの属性語候補の抽出

もし前節で収集した各 Web ページがクラスの対象物の属性情報記述ページであれば，ページ中に属性語が HTML タグや文字修飾などによって，視覚的に認知し易い形で記述されているはずである．そこで我々は，属性情報記述ページ中で属性語を強調するのに頻繁に使用される HTML タグ，文字修飾に注目し，特定の HTML タグまたは括弧類で囲まれた文字列，特定の接頭修飾に続く文字列，および特定の接尾修飾を伴う文字列をパターンにより属性語候補として抽出する．表 1 は，本研究で抽出に用いた HTML タグと文字修飾である．この際，表のセルを表現する TD タグに関しては，表中で属性語が記述されるのは主に一行目と一列目のセルであるという観察 [Yoshida 01] から，それらのセルに対応するタグのみに注目する．図 3 は，属性抽出の適用される Web ページの一例であり，図中で太字で強調された文字列がパターンにより獲得される属性語候補である.

表 1 の HTML タグと文字修飾に基づくパターンは，属性語以外にもページの目次や属性値などを抽出することがあるため，我々はパターンにより得られた属性語候補を，ストップワードリストおよび形態素解析[*2]を利用してフィルタリングする．ストップワードリストは，34 のストップワードと 5 つの正規表現（「インターネット」,「リンク」,「ニュース 」,「ページ」,「メール」を部分文字列として含む語にマッチ）から構成されている．また，形態素解析では，属性語候補が固有名詞および数詞以外の名詞だけから構成されるときのみ属性語候補とみなす.

本研究では，クラスの属性語候補を得るために，テキストから知識の抽出パターンを自動獲得するアプローチ [Muslea 99, Kushmerick 00, Sakamoto 01] を取っていない．これは，本研究の知識獲得の対象である属性語が，属性情報記述ページ中では比較的限られた文脈で多数出現するため，少数のパターンでも，大量の文書を処理すれば，大量の属性語候補を獲得できたからである．このステップで誤って獲得された属性語候補については，次節で説明する統計的なフィルタリングにより除去する.

§3 属性語候補のサイト頻度に基づくフィルタリング

本節では，前節で獲得した属性語候補から誤った属性語候補を取り除き，属性情報記述ページの発見に用いる典型的な属性語を得る手法について述べる.

前節では，クラスをトピックとして記述したページを知識源としてクラスの属性語候補の獲得を行った．しかしながら，収集されたページにクラスと無関係な内容が含まれうるため，得られる属性語候補には誤った属性語候補が含まれる．例えば，複数のクラスをトピックとして扱ったページからは，他のクラスに関する属性語も属性獲得の対象としているクラスの属性語候補として抽出されてしまう．また，抽出に用いたパターンは，属性語の強調以外に単なる文字列の強調にも用いられるため，前

∗1 ただし一行目と一列目のセルに対応するタグのみを考慮した.

∗2 MeCab: http://mecab.sourceforge.jp/

節で述べた手法の結果には本質的に誤った属性語候補が含まれる．そこで，前節で Web ページから獲得されたクラスの属性語候補から，典型的な属性語候補がより重要視されるような統計値を用いて属性語候補をフィルタリングすることで，誤って獲得された属性語候補を取り除く．

本研究では，基本的に多数の Web ページ制作者が共通して記述する属性語は，ユーザが知りたい典型的な属性語であるという仮説に基づき，まず予備実験で獲得された属性語候補を，既存手法 [高橋 05, Tokunaga 05] で用いられている相互情報量，*idf*，および上の仮定と類似した発想に基づく *df* などの統計値を用いてランキングし，典型的な属性語が得られるかどうか結果を分析した．その結果，*idf* や相互情報量は，対象のクラスに特有の属性語を重要視する傾向があることが分かった．クラスの属性語には，クラスに特有の属性語（例: 映画の「字幕」）と複数のクラスに共通する属性語（映画，ドラマ，演劇などの「出演」）があるが，クラスに特有の属性語が複数のクラスに共通する属性語よりも必ずしも典型的であるとは限らないため，クラスに特有の属性語を重要視することは望ましくない．一方で属性語候補の *df* は，Web ページ製作者が同一テンプレートから大量の Web ページを生成している場合，テンプレートに含まれる属性語候補の *df* が非常に大きくなるという問題が生じた．そのため，不適切な属性語候補がテンプレート中に現れていた場合（例: amazon.com の ASIN (Amazon Standard Identification Number)），*df* ではそのような属性語候補を取り除くことが非常に難しくなる．

本稿では，前段落で述べた仮説をより直接的に用いるために，*df* の問題点を解決する**サイト頻度**を提案し，クラス特有の属性語を重要視する相互情報量や *idf* といった統計値と組み合わせることなく，この統計値単独で属性語のフィルタリングを行う．サイト頻度は，我々のテストの範囲内では相互情報量，*idf* や *df* といった既存の統計値より有効に働いた．属性語候補 x のサイト頻度 $sf(x)$ は，以下のように定義される．

$$sf(x) = \text{属性語候補 } x \text{ を抽出した } \textbf{Web サイト}\text{の数} \quad (1)$$

サイト頻度を獲得するためには，各 Web ページの属する Web サイトを認識する必要がある．ここで言う Web サイトとは，同一 Web ページ製作者が作成した Web ページ群のことである．Web ページの属する Web サイトを正確に推定することは難しいが，本研究では Web ページの URL（例: `http://ex.org/foo/bar.html`）のパスを末尾から逆に辿り (`http://ex.org/foo/`→`http://ex.org/`)，Web サイトのトップページのファイル名となり易い，正規表現`/^(?:index|default|main)\..+/`にマッチするファイル名のファイルを含む最下層のディレクトリまでのパスを求め，そのパスを Web サイトと一対一に対応するものと仮定した．ただし，そのようなディレクトリが存在しなかった場合は，サーバー名（例: `http://ex.org/`）を単に Web サイトとして定義した．

サイト頻度は，属性語候補 x をクラスの属性語として用いた **Web ページ製作者の数**を，属性語候補の属性らしさとして表現したものであると言える．前に述べた仮説，すなわち基本的に多数の Web ページ制作者が共通して記述する属性語は，ユーザが知りたい典型的な属性語であるという仮説が正しければ，このサイト頻度により，前節までの手法で集めた大量の属性語候補から，典型的な属性語候補を落とすことなく，誤って獲得されたノイズの属性語候補を除くことができると期待される．この点については，実験結果を通して確認する．

なお，ここまでで述べたクラスの属性知識ベースの構築は，1 クラス辺り数分程度で行えることを確認しており，数千のクラスに対する大規模な属性知識ベースも数日程度で構築することが可能である．

3·2 属性情報記述ページの発見

本節では，前節で獲得されたクラスの典型的な属性語を用いて，そのクラスに属する対象物の最良の属性情報記述ページを発見する手法について述べる．

§1 ステップ 1: 対象物を含むページからの属性語抽出

ステップ 1 ではまず，対象物の属性情報記述ページの候補として，対象物名を含むページを検索エンジンを用いて収集する．次に，各ページについて，ページ全体から 3·1·2 節で述べた方法を用いて属性語を抽出する．その後，以下のステップ 2 で各ページから抽出された属性語とクラスの属性語を比較し，代表的な属性情報記述ページを発見する．

§2 ステップ 2: クラスの属性知識に基づく属性情報記述ページの発見

1 章で属性情報記述ページの定義を述べたが，実際に属性情報記述ページの「良さ」を計るには様々な基準が存在する．本研究の目的は，ユーザが知りたい典型的な属性の情報をできるだけ多く含むページを発見することであるため，ここではページに含まれる典型的な属性語の数に従ってページをランキングし，典型的な属性語の情報が数多く記述されているページを発見する．

具体的には，入力の対象物 x とそのクラス c に対し，ページ p の属性情報記述ページとしての良さを表すスコアを，ページ p から獲得した属性語の集合 $\mathcal{A}_p$ とクラス c の属性語の集合 $\mathcal{A}_c$ に基づき，以下のように計算した:

$$score(p,c,x) = \frac{\#(\mathcal{A}_p \cap \mathcal{A}_c) \times ratio(\mathcal{A}_p, \mathcal{A}_c)}{ave(\mathcal{A}_p, p) \times text\_size(x,p)}. \quad (2)$$

ここで，分子の $\#(\mathcal{A}_p \cap \mathcal{A}_c)$ は，良い属性情報記述ページはクラスの属性語を多く含むという傾向を反映した項であり，$\mathcal{A}_p$ と $\mathcal{A}_c$ に共通する属性語の数として計算される．また， $ratio(\mathcal{A}_p, \mathcal{A}_c)$ は，対象物名が複数のクラスのインスタンスを指しうる場合（例: 映画 ↔DVD）に，入力のクラスに属する対象物のページを発見するための

表 2 Web から獲得された各クラス上位の属性語（最大 29 語）と被験者により提示された属性語

クラス	提案手法により獲得された属性語	被験者の提示した属性語
デジタルカメラ（例: Caplio RZ1, DSC-T7）	電源 レンズ ホワイトバランス 重量 撮像素子 *オート *注意 記録媒体 価格 セルフタイマー メーカー サイズ メモリ 測光方式 動画 シャッタースピード ファインダー フォーカス 露出補正 *ＣＰＵ 有効画素数 商品名 撮影モード 記録メディア 付属品 外形寸法 静止画 絞り デジタルズーム	(**メーカー**[2] **会社名**) (**画素**[2] **画素数**) (**価格 値段**) **重量 記録サイズ** 充電時間
競走馬（例: ウイングアロー, クロフネ）	父 コメント 馬名 母 名前 名 毛色 *牡 順位 *競馬場 詳細 *牝 性別 着順 *問い合わせ *メッセージ レース * 応募方法 *受付時間 *担当 タイトル 備考 *黒鹿毛 *業務内容 *勤務地 品種 *不明 夢 価格	(**戦績**[2] **競争成績**) 生産者[2] 生年月日[2] (収得賞金 賞金) **性別 血統** 馬主
野球選手（例: 青木宣親, 福留孝介）	高校 *ニコニコ *昨日 *秘密 *後編 ドラフト *出典 *祝 *トラックバック時刻 *今日 *累計	(球団[2] 所属球団) 生年月日[2] ポジション 性別 **出身校** 背番号 打率 身長 出身
俳優（例: 妻夫木聡, オダギリジョー）	生年月日 *監督 タイトル 出演 映画 名前 作品名 *場所 ビデオ コメント *脚本 役名 特集 *ＮＯ 趣味 血液型 舞台 名 *助演男優賞 *演出 *ブルース 特技 *配給 氏名 *コメント数 *音楽 *音楽賞 *順位 *原題	(出身地[2] 出身) (**出演作品**[2] **出演番組**) **生年月日**[2] (所属 所属事務所) 年齢 **名前**
病院（例: 東大病院, 聖路加国際病院）	電話番号 所在地 *午前 住所 診療科目 *午後 診療時間 内容 休診日 備考 受付時間 内科 院長 電話 小児科 場所 整形外科 問い合わせ先 名 名称 *日時 連絡先 外科 *講師 病床数 医師 土曜日 泌尿器科 皮膚科	**電話番号**[3] **(住所**[2] **場所) (診療科目**[2] **診療科) (診療時間**[2] **診療受付日時)**
株式会社（例: ツムラ, カルビー）	資本金 所在地 従業員数 電話番号 住所 代表者 事業内容 設立 本社 *価格 代表者名 ＦＡＸ番号 電話 本社所在地 会社名 営業時間 会社概要 備考 売上高 代表取締役 取引銀行 代表取締役社長 設立年月日 連絡先 担当 交通 問い合わせ 業務内容 問い合わせ先	**(住所**[2] **所在地) 資本金**[2] **電話番号**[2] 株価 **売上高 社員数 設立 社長**
ワイン（例: オーパス・ワン, シャトー・マルゴー）	価格 容量 *住所 コメント 作り方 サイズ *営業時間 *白 *赤 *赤ワイン 商品番号 生産者 *電話 *塩 商品名 タイプ *日本酒 場所 産地 備考 *白ワイン ぶどう品種 材料 *電話番号 名前 *焼酎 アルコール分 料理 原産国	**(生産地**[2] **産地) 種類**[2] **価格**[2] 生産年[2] **飲み口 容量 ワイナリー**
博物館（例: 国立科学博物館, 印刷博物館）	休館日 開館時間 場所 住所 所在地 入館料 電話 内容 問い合わせ 料金 問い合わせ先 入場料 備考 駐車場 日時 交通 無料 大人 営業時間 *期間 アクセス *主催 時間 *対象 問合せ 交通案内 *参加費 *期 電話番号	(**場所**[2] **住所**) (**入館料**[2] **料金**) (**入館時間 利用時間 開館時間**) **展示物 定休日 電話番号**
文庫（例: ノルウェイの森, TSUGUMI）	著者 出版社 価格 備考 書名 著 *下 著者名 イラスト タイトル 発売日 発行 内容 解説 原作 サイズ 分類 あとがき 作家名 *住所 定価 作品名 名 コメント *入力 税込価格 行 *日時 概要	**価格**[3] **(著作 著者 作者) 発売日**[2] **書名 出版社** ジャンル ISBN
遊園地（例: はなやしき, としまえん）	駐車場 コメント 住所 観覧車 *無料 料金 場所 内容 テーマパーク *おすすめ書籍 営業時間 入場料 *なし *学校 ガイド 所在地 メリーゴーランド フリーパス 電車 地図 トイレ 交通 *定員 *不明 子供 電話 休園日 *プール *水族館	**(場所**[2] **住所) (入園料**[2] **入場料) (入場時間 開園時間 営業時間) 定休日 電話番号 アトラクション**

項であり，$\mathcal{A}_p$ に含まれる属性語のうち $\mathcal{A}_c$ に含まれる割合（すなわち，$\frac{\#(\mathcal{A}_p \cap \mathcal{A}_c)}{\#(\mathcal{A}_p)}$ ）として計算される．また，分母の $ave(\mathcal{A}_p, p)$ は，複数の対象物を含むカタログページよりも，対象物のみについて記述したページを選ぶために用いた項であり，ページ p 中における全属性語 $a \in \mathcal{A}_p$ の出現回数（ただし，表 1 の HTML タグと文字修飾に基づくパターンで抽出されたもののみを考慮し，他の文脈で使われているものは考慮しない）の平均として計算される．最後に $text\_size(x, p)$ は，対象物をトピックとして記述するページでは，属性情報のレイアウトに対象物名を含む短い表題が付くことが多いという事実を反映した項である．具体的にこの項は，ページ中で最初に対象物名を含む任意の HTML タグで囲まれた文字列（例えば，対象物「タイタニック」に対して図 3 のページを得た場合，H3 タグで囲まれた文字列「タイタニック（公開年：1997）」）の長さとして計算される．このようにして計算されるスコア $score(p, c, x)$ が最大のページ p を，クラス c に属する対象物 x の最良の属性情報記述ページとして出力する．

上記の属性情報抽出ページ発見の手続きは非常に高速に動作する．例えば，検索エンジンで絞った 30 の Web ページから属性語を抽出し，上記の関数で属性情報記述ページを選ぶ場合，ページのダウンロード時間を除けば，通常の検索プロセスに追加してかかる時間は 1 秒未満であり，サーバー側で行う処理としては実用可能な基準を達成していると言える．

4. 評 価 実 験

本章では，我々のシステムを通常の検索エンジンと比較することで，我々のシステムの有効性を示す．

クラスの属性知識を獲得する知識源として，我々はまず GNU wget を用いて 0.7TB（HTML タグ含む約 5,900 万文書）の日本語 Web リポジトリおよび全文検索エンジン[*3]（Local Search Engine, 以下 LSE）を構築した[*4]．

次に，評価用のクラスと対象物のペアを作成した．本システムは，実際にユーザに利用してもらうことを想定しているため，ユーザが実際に関心があるクラスの対象物について評価を行うこととした．具体的には本研究に関与しない第三者に，彼らが興味があるクラスを 10，またそのインスタンス，すなわち対象物を各クラスに対して 10 ずつ挙げてもらった．表 2 左がテストに用いたクラスとその対象物（一部）である．

今回比較するシステムは，Google[*5]の対象物名とクラ

*3 FreyaSX (0.99.10): `http://www.delegate.org/freyasx/`
*4 商用検索エンジンを用いて知識源の文書を得た場合，ランキング手法の不透明性から獲得される属性語の分析が困難となるため，用いなかった．
*5 `http://www.google.co.jp/`

ス名をクエリとして検索した結果のトップページを出力するシステム（以下，GOOGLE），と Google のランキング上位 30 ページを対象に提案手法を用いて属性情報記述ページを発見するシステム（以下，SP），最後に LSE を用いて対象物名とクラス名をクエリとして検索した結果からランダムに選んだ 10,000 ページを対象に提案手法を用いて属性情報記述ページを発見するシステム（以下，SP*）の 3 システムである．GOOGLE と SP*を比較することで，属性情報記述ページを発見するために，我々の提案するスコアがどれだけ有用か把握することができる．また，SP と SP*を比較すれば，Google のランキングがどれだけ属性情報記述ページの候補を制限するのに有用かを知ることができる．実験では，各システムが表 2 に挙げたクラス名およびその対象物に対して出力したページを 3 人の被験者によって評価してもらい，各システムの性能を比較する．以下，クラスの属性知識の獲得実験，3 つのシステムの出力する属性情報記述ページの比較実験の順に説明する．

4·1 クラスの属性知識の獲得

まず，評価に用いた各クラスについて，LSE を用いて得られたクラス名を含む全ページからランダムに選んだ 10,000 ページを[*6]を知識源として，3·1 節の獲得手法により属性語候補を獲得し，サイト頻度によるランキング上位の属性語（最大 29 語）をクラスの属性知識として用いた（表 2 中央）．

我々は接尾辞追加質問テスト [Tokunaga 05] により，各属性語が妥当かどうかを簡単に人手で調べた（∗ の付加された属性語が不適当と判断された）．その結果，野球選手とワイン以外のクラスについては，ほとんどが妥当な属性語であった．ワインについては約半分が不適当な属性語であるという結果が得られたが，これはワインの属性語を獲得する知識源としたページが，他のクラス（例えばワインショップ）のページを多数含んでおり，結果としてそれらのクラスの属性語が混入したためだと考えられる．また，野球選手については， LSE のインデクスが辞書に基づく単語分割により作られており，検索に漏れがあることから，属性語候補の獲得のための知識源のページが十分な数得られなかったことが[*7]，適当な属性語が獲得できない原因であった．

提案手法により獲得された属性語を詳しく見ていくと，同じ属性を表すために，自然文には現れないような属性語が，属性情報記述ページでは多数使われていることが分かる．まず最初のケースとして，属性語がクラスに特化した専門語で記述されている場合があった（例: デジタルカメラの撮像素子）．また，二番目のケースとして，対象物の同じ属性を表現する様々な同義の属性語には，表現が過度に省略されているものもあった（例: 文庫の著者名 ↔ 著）．これは，他の属性語と区別する際の曖昧性が無い場合には，属性語がかなり柔軟に省略されることを示唆している．その他のケースとしては，属性情報記述ページ特有の属性語として，例えば午前中の営業時間を表す「午前」や，電車でのアクセスを示す「電車」，子供料金を表す「子供」なども属性語として獲得されている．これらの属性語は，前段落で述べた接尾辞追加テストでは妥当な属性語とはみなされなかったが，属性情報記述ページを発見するという観点からは，妥当な属性語と考えられる．属性情報記述ページが，そもそも対象物のクラスについて良く知る人間が属性情報を簡潔に記述することを目的として作成することが多いことを考えると，上で分析したように（自然文には属性語として現れにくい）簡潔かつ専門的な表現が属性語として好まれるのは自然なことと思われる．

このように，平書きの文から抽出した属性語と，属性情報記述ページで用いられる属性語は，同じ属性を指すにも関わらず異なる語が使われることが頻繁にある．従って平書きの文から抽出した属性語を用いたのでは，異なる属性語の使われる属性情報記述ページを発見することは困難であると予想される．また，人が属性語を追加して属性情報を検索しようとする場合に最初に思いつくのは，平書きで書かれるような「自然な」属性語であることから，提案手法で得られる属性語により，人が属性語を追加して検索することでは発見が難しい属性情報記述ページを見つけられる可能性もある．これらの結果から，対象物の属性情報記述ページを発見するために，そのクラスの属性情報記述ページを属性語獲得の知識源とした方針は正しかったと言える．

次にクラスの属性知識を評価するために，3 人の被験者に，それぞれ各クラスについて自身が知りたい 4 つの属性を対象物を見ることなく決めてもらった（表 2 右）．表中，属性語の上付きの数字は，属性語が複数の被験者により提示された場合にその人数を表し，括弧でグループ化された属性語は，著者の直感で同義であると分類した属性語である．各同義な属性語のグループに対し，提案手法により獲得した属性語に同義と考えられるものが含まれていた場合，それらのグループは太字で示した．各被験者は，単語としては異なるが，互いに同義で同じ属性を意味する属性語を知りたい属性として提示していた．

実際に，LSE の仕様により，ほとんど適切な属性を獲得できなかった野球選手を除く 9 のクラスについて，被験者が提示した 81 語（異なり語数，以下同様）の属性語のうち，66 語 (約 81%) の属性語については，提案手法により獲得された属性語あるいはその同義語と考えられるものであった．また，各被験者が提示した属性語のうち，デジタルカメラの「メーカー」や「画素数」のよう

∗6 実際には，3·1·1 節で述べたクラス名の出現位置の制約を満たすページのみが知識源として用いられるため，知識源はこれより少ない数のページに制限される．

∗7 FreyaSX の最新版では，文字ベースの漏れのない検索が可能であり，野球選手についても妥当な属性語が得られることを確認した．

表 3 属性情報記述ページの発見: 実験結果

		被験者 1			被験者 2			被験者 3			被験者 1-3 (平均)		
クラス	対象物の数	GOOGLE	SP	SP*	GOOGLE	SP	SP*	GOOGLE	SP	SP*	GOOGLE	SP	SP*
デジタルカメラ	4/10	1.50	**3.25**	1.00	1.00	**3.00**	1.25	2.00	**3.00**	1.75	∗1.50	∗**3.08**	**1.33**
競走馬	6/10	**4.00**	**4.00**	3.33	**4.00**	**4.00**	3.33	**4.00**	**4.00**	3.33	∗**4.00**	∗**4.00**	∗3.33
野球選手	1/10	0	0	0	0	**1.00**	0	1.00	**3.00**	1.00	0.33	**1.33**	0.33
俳優	4/10	**1.00**	0.75	0.75	**1.75**	1.50	1.50	2.00	**2.25**	0.75	∗**1.58**	1.50	1.00
病院	4/10	0.75	**2.50**	1.25	0.25	**3.75**	1.25	0	**3.75**	1.25	0.33	∗**3.33**	1.25
株式会社	0/10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
ワイン	5/10	0.40	0.60	**1.60**	1.20	1.20	**3.20**	3.00	2.80	**3.40**	∗1.53	1.53	∗**2.73**
博物館	6/10	0	2.67	**3.17**	0.67	2.67	**3.00**	0.17	3.67	**3.83**	0.28	∗3.00	∗**3.33**
文庫	7/10	**4.00**	2.71	1.29	2.00	**2.29**	2.14	**4.00**	3.00	2.43	∗**3.33**	2.67	∗1.95
遊園地	5/10	1.00	2.60	**3.00**	1.00	2.60	**3.00**	1.20	3.20	**3.40**	1.07	∗2.80	∗**3.13**
対象物ごとの平均	42/100	1.71	**2.41**	1.98	1.55	**2.60**	2.38	2.17	**3.24**	2.62	1.81	**2.75**	2.33
ベスト 5 (∗)											2.59	**3.26**	2.86

に，三人全員の被験者が提示した属性に関する 39 語の属性語に注目すると，提案手法により獲得された属性語は 37 語 (95%) となり，二人の被験者が提示した属性に関する 15 語の属性語に関する獲得数 8 語 (53%)，また一人の被験者のみが提示した属性に関する 27 語の属性語に関する獲得数 21 語 (78%) より良いことから，3･1･3 節で述べた，多数の Web ページ制作者が共通して記述する属性語はユーザが知りたい典型的な属性語であるという仮説が正しいことも確認できる．

4･2 属性情報記述ページの発見

次に，前節と同一の 3 人の被験者に，GOOGLE, SP, および SP*が各クラスの 10 の対象物について出力した各ページを評価してもらった．具体的には，まず被験者に，各ページが指定された対象物を含むページどうかを調べてもらい，3 人の被験者が，3 つのシステムが出力するどのページも与えられた対象物を参照していると判断した対象物についてのみ，評価した．次に各ページについて，3 人の被験者に，被験者自身が前節で各クラスについて知りたいと挙げた 4 つの属性とその属性値の情報が，属性語（または，被験者の主観でその同義語と考えられる語），属性値ともに明示的にテキストとしてページに含まれており，かつ属性語と属性値の関係が自然文ではなく，表や箇条書きや文字修飾などのレイアウトによって，視覚的に即座に分かる形で書かれているかどうかを被験者自身に調べてもらった．その後，ページが含んでいる属性-属性値対の数で各ページを評価付けしてもらった．例えば，ワインのクラスについて，被験者が「生産地」,「葡萄品種」,「アルコール度数」,「価格」を属性として想定している場合，図 1 左のページでは,「生産地」,「葡萄品種」,「アルコール度数」と同義の属性語およびその属性値を含んでいるため 3，一方で図 1 右のページは「生産地」に関する情報のみ含んでいるため 1 となる．ここで，以上の属性-属性値がページに含まれるかどうかの判断基準は，非常に厳しいものであることに注意されたい．というのも，属性語を省略しても属性値がどの属性の値か明確な場合（例えば，デジタルカメラのメーカー，病院の電話番号，文庫の書名など）は，属性語が省略されることが多いためである．

表 3 は，3 つのシステムが出力した全てのページに全員の被験者が対象物が含まれていると判断した 42 の対象物に対する実験結果である．表中の各セルは，各クラスの対象物について，各システムが出力したページに各被験者が含まれていると判断した属性-属性値対の数の平均値である．SP は，被験者が対象物に対して知りたい 4 つ属性のうち，平均して 2.75 個（約 69%）の属性の情報を含むページを発見することに成功した．さらに，各システムについて，それぞれ被験者 3 人の平均の評価の最も良い 5 つのクラス（表 3 中 ∗）の結果をみると，SP が出力するページには，平均して 3.26 個（約 82%）の属性の情報が含まれていた．ここで注意して欲しいのは，被験者ごとに知りたいと提示した属性は異なるため，それぞれの被験者に対して精度は異なるものの，全ての被験者が，SP の出力したページに最も属性-属性値関係が含まれていると評価したことである．これは，できるだけ多く属性語が含まれているページを出力するという我々の戦略が，ユーザを満足させる上で一般的に有効であったことを意味する．また，SP*が GOOGLE よりも良いページを出力していることにも注目されたい．SP*が，PageRank などの対象物の代表ページを見つけるためのスコアを使ってないことを考えると興味深い結果である．このことからも，我々の属性語に基づくアプローチが，属性-属性値関係を含む属性情報記述ページの発見に非常に有用であることが分かる．

各クラスの結果をもう少し詳しくみてみると，各システムの傾向の違いがよりはっきりと分かる．例えば，GOOGLE は他のシステムより競走馬と文庫のクラスについて良い結果を得ることができたが，これは，Google が高くランキングするいわゆる権威の高いページ（例えば，Yahoo!や Amazon.com など）が網羅的なデータベースを持っていたからである．一方，権威の高いページがそのようなデータベースを持っていない場合は，Google の検索結果の上位に属性情報を多く含むページが含まれない場合もある．これは，SP*がワインや博物館，遊園地

について Google のランキングを使った GOOGLE, SP より良い結果を出したことからも裏づけられる．これらのページについては，GOOGLE は，ショッピングサイトやニュース，blog など，属性-属性値情報を得るにはあまり役に立たないページをトップにランキングしてしまっていた．しかしながら，それ以外のクラスで SP が SP*より良い結果を出したことを考えると，Google の用いる PageRank などのランキング尺度が有効に働く場合もあると考えられ，良い属性情報記述ページを得るためには，我々の提案する属性語に基づく尺度とランキング尺度を同時に考慮することが必要であることが分かる．

最後に，本システムの適用範囲について定性的な考察を行う．本システムの性能は，与えられたクラスの対象物に関する属性情報記述ページがそもそも存在するかという点に依存する．基本的に属性情報記述ページが多く存在するクラスであれば，そのようなページを集めて知識源として用いることで，多くの典型的な属性語を得ることができる．著者の調べた範囲では，今回実験に用いた 10 のクラスに関しては，属性情報記述ページが多く存在しており，(情報検索のニーズのある）多くのドメインでは，対象物の属性情報記述ページが存在する可能性が高いと期待される．

5. ま と め

本稿では，対象物名とそのクラス名を入力とし，対象物の詳細な属性情報を記述したページを Web から発見する手法を提案した．我々のシステムは，予め HTML タグと文字修飾に基づくパターンにより Web から獲得した属性語候補を Web の特性を考慮したサイト頻度と呼ばれる統計値でフィルタリングし，クラスの属性知識ベースを構築する．そして，このクラスの属性知識ベースを用いて，入力の対象物を含むページの中から最も典型的な属性-属性値関係を含むと考えられるページを発見する．3 人の被験者に Google と我々のシステムの出力するページを比較してもらったところ，我々のシステムが出力するページに，被験者の知りたい属性-属性値関係が最も多く含まれると判断された．この結果は，今回の評価基準の厳しさ（属性語と属性値が共にページ中に視覚的に分かり易い形で含まれる場合のみ，属性-属性値関係が含まれるとした）を考慮すると有望な結果である．

今後は，属性情報記述ページのスコア関数の改良，また上位-下位語関係を用いて，入力された対象物のみから属性値を発見する手法の開発を行う予定である．

謝 辞

本研究の一部は，文部科学省研究費補助金（平成 15 年度若手研究 (A) 15680005，平成 15 年度萌芽研究 15650015）ならびに同省科学技術振興調整費（任期付若手研究員支援プログラム）の支援を受けた．記して謝意を表する．

◇ 参 考 文 献 ◇

[Agrawal 02] Agrawal, R. and Srikant, R.: Searching with Numbers, in *Proc. WWW*, pp. 420–431 (2002)

[Almuhareb 04] Almuhareb, A. and Poesio, M.: Attribute-Based and Value-Based Clustering: An Evaluation, in *Proc. EMNLP*, pp. 158–165 (2004)

[Chen 00] Chen, H.-H., Tsai, S.-C., and Tsai, J.-H.: Mining Tables from Large Scale HTML Texts, in *Proc. COLING*, pp. 166–172 (2000)

[Kushmerick 00] Kushmerick, N.: Wrapper induction: Efficiency and expressiveness, *Artificial Intelligence*, Vol. 118, No. 1–2, pp. 15–68 (2000)

[Muslea 99] Muslea, I.: Extraction Patterns for Information Extraction Tasks: A Survey, in *AAAI Workshop on Machine Learning for Information Extraction.*, pp. 7–14 (1999)

[Page 98] Page, L., Brin, S., Motwani, R., and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Database Libraries Working Paper (1998)

[Sakamoto 01] Sakamoto, H., Murakami, Y., Arimura, H., and Arikawa, S.: Extracting Partial Structures from HTML Documents, in *Proc. FLAIRS*, pp. 264–268 (2001)

[Salton 88] Salton, G. and Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, Vol. 24, No. 5, pp. 513–523 (1988)

[Shimada 05] Shimada, K. and Endo, T.: Acquisition of New Training Data from Unlabeled Data for Product Specifications Extraction, in *Proc. PACLING*, pp. 272–277 (2005)

[Tokunaga 05] Tokunaga, K., Kazama, J., and Torisawa, K.: Automatic Discovery of Attribute Words from Web Documents, in *Natural Language Processing – IJCNLP 2005*, Vol. LNAI 3651, pp. 106–118, Springer-Verlag (2005)

[Yoshida 01] Yoshida, M., Torisawa, K., and Tsujii, J.: Extracting Ontologies from World Wide Web via HTML Tables, in *Proc. PACLING*, pp. 332–341 (2001)

[Yoshida 05] Yoshida, M. and Nakagawa, H.: Specification Retrieval – How to Find Attribute-Value Information on the Web, in *Natural Language Processing - IJCNLP 2004*, Vol. LNAI 3248, pp. 338–347, Springer-Verlag (2005)

[高橋 05] 高橋 哲朗, 乾 健太郎, 松本 裕治：言語パタンと統計的共起尺度による属性関係抽出, 言語処理学会第 11 回年次大会論文集, pp. 432–435 (2005)

〔担当委員：佐藤 理史〕

2006 年 4 月 28 日 受理

著 者 紹 介

吉永 直樹

2000 年東京大学理学部情報科学科卒業．2002 年より 2005 年まで日本学術振興会特別研究員 (DC1)．2005 年東京大学院情報理工学系研究科博士課程修了．同年 4 月より日本学術振興会特別研究員 (PD)．計算言語学の研究に従事．博士（情報理工学）．

鳥澤 健太郎(正会員)

1992 年東京大学理学部情報科学科卒業．1995 年同大学大学院理学系研究科情報科学専攻博士課程退学，同年より同大学院理学系研究科情報科学専攻助手．1998 年より 2001 年まで科学技術振興事業団さきがけ研究 21 研究員兼任．2001 年より北陸先端科学技術大学院大学情報科学研究科助教授．計算言語学の研究に従事．博士（理学）．