Webからの属性情報記述ページの発見

吉永 直樹 †‡
† 日本学術振興会
n-yoshi AT jaist.ac.jp

鳥澤 健太郎 ‡
‡ 北陸先端科学技術大学院大学
torisawa AT jaist.ac.jp

1 はじめに

本稿では，与えられた対象物とそのクラスから，対象物を記述した代表的な（最も多くの情報を含む）web 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，属性情報記述ページとは，対象物の包括的な**属性**，すなわち人が知りたい対象物の側面の情報を，表やリスト形式など，視覚的に認知し易い形で記述したページのことである（図 1 左）．これまで，ユーザが対象物について詳しい情報を Web から見つけるには，汎用的な検索エンジンの返す対象物に関する複数のページ（図 1 右）を横断的に見て，自然文で書かれた部分的な知識を統合する作業が必要であった．ここで，可読性の高い属性情報記述ページを提示することができれば，このようなユーザの要求を瞬時に満たすことが可能である．

我々のシステムは，対象物（例: Ch. Mouton Rothchild）の属性情報記述ページを，そのクラス（例: wine）の属性知識ベースに基づき発見する．我々はまず，クラス属性の知識ベースを，属性情報記述ページにおける属性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する．この学習の際には，属性候補から属性として不適切な単語を取り除くために，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，通常の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スコア付けし，最良のページを発見する．

我々は，0.7 TB（HTML タグを含む）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，3 人の被験者に各クラスについて知りたい属性をあげてもらい，各クラス 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，我々の出力した web ページに，被験者の知りたい属性のうち平均で 69% が含まれていることが分かった．

2 関連研究

2.1 web からの属性知識の獲得

これまで，2 種類の異なる手がかりを用いた属性知識獲得の研究が行われてきた．最初の手がかりは，対象物とその属性が共起する構文パターン（例: “the * of the x is,”）であり，これとある種の統計量を組み合わせることで，自然文から属性知識を獲得することが試みられている [1, 7, 4]．しかしながら，属性情報記述ページで用いられる属性の中には，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では不十分である．

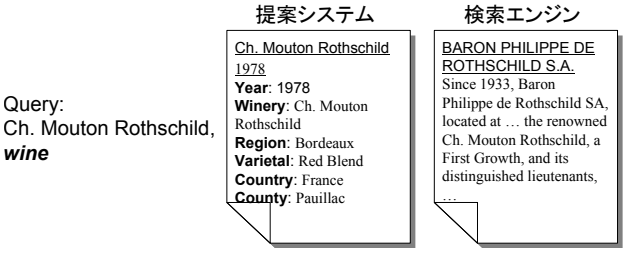

図 1: “Ch. Mouton Rothschild, wine” に対する属性情報記述ページと通常の検索エンジンの出力

一方，Web を知識源として用いる際には，**レイアウト情報**が，より直接的に属性-属性値を記述する方法として研究されてきた．具体的には，表中のセル間の類似度 [2] や，EM に基づく教師なし学習 [6] が HTML で記述された表から属性を発見するのに用いられてきた．これらの手法は，表認識に計算コストがかかる上に，知識源の表形式を人手で注意深く用意する必要があった．

我々の属性獲得手法は，表形式に限らず様々な形式から属性知識を抽出可能であり，計算コストも少ないことから，大量のページを処理することで，包括的で信頼性の高い属性知識を獲得することが出来る．またこの手法は，対象物を記述したページから属性語を認識する際にも現実的な時間で動作する．

2.2 属性情報記述ページの発見

島田ら [3] はクラスを商品に特化して，商品の仕様表を教師あり学習に基づき発見する手法を提案した．しかしながら，教師あり学習を用いる場合，一般ドメインの対象物に応用することは困難である．

吉田ら [5] は，与えられたクラスの対象物を記述した属性情報記述ページを収集する手法を提案した．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 ページが与えられたクラスに関係する確率を計算する．彼らは実験で，属性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あることが示したが，そもそも彼らの用いている知識ベースは，獲得の際の計算コストが高いこと，知識源が注意深く選別された表形式に限定されていることなどから一般ドメインに拡張することは困難である．

我々は，検索エンジンによって収集されたノイズを含む 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回避している．

表 1: 属性獲得に用いた 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:	:, :, ／, /, =

```
<H3>タイタニック (公開年:1997)</H3> 監督/ James Cameron
<BR />詳細: <TABLE><TR><TD>出演</TD><TD>Leonardo
DiCaprio, Kate Winslet</TD></TR><TR><TD>上映時間</TD>
194 分</TD></TR></TABLE>
```

図 2: 属性抽出の例

3 手法

本節では，与えられた対象物とそのクラスから，属性情報記述ページを発見する手法について述べる．

属性情報記述ページを発見するには，対象物の属性が有用であるが，あらゆる対象物について属性知識を獲得するのは，その膨大さ，増進性から言って実際的ではない．そこで本研究では，対象物より頻繁に文書中に出現するクラスの属性知識を獲得する．対象物はクラスの属性を継承するため，対象物を記述した web ページから獲得される属性候補と，クラスの属性との一致度により，対象物の属性情報記述ページを発見できる．

与えられた対象物とそのクラスに対し，我々は対象物の属性情報記述ページを以下の手順で発見する．

STEP 1: ページからの属性抽出 与えられた対象物を含む web ページを検索エンジンによって収集し，各ページから属性を抽出する．

STEP 2: 属性知識に基づく属性情報記述ページの発見 STEP 1 で収集された各 web ページから抽出された属性とクラス属性を比較することで，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の各ステップ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

3.1.1 知識源の web ページのサンプリング

我々はまず，属性知識獲得の知識源として，検索エンジンを用いてクラス単語を含む文書を収集する．この際，クラスと無関係の文書をなるべく取り除くために，クラス単語がページのトピックとなり易い，TITLE, H1∼H6, CAPTION, TD, および TH に囲まれているページを知識の獲得源とし，ページ中でクラス単語が最初に現れた位置以降から，属性候補を獲得する．

3.1.2 Web ページからの属性候補の抽出

次に，HTML タグと文字装飾に基づくパターンを用いて属性候補を抽出する．表 1 は，本研究で用いた HTML タグと文字装飾である，この際，表のセルを表現する TD タグに関しては，表中で属性が記述されるのは主に一列目と一行目のセルであるという観察から，それらのセルに対応するタグのみに注目する．図 2 は，属性抽出の一例である．

このようにして得られる属性候補を，形態素解析およびストップワードによってフィルタリングして得られるものをページから得られる属性候補とする．

3.1.3 属性候補のサイト頻度に基づくフィルタリング

前節の手法により得られる属性候補は，知識源の web ページ集合中にクラス単語と無関係のページが含まれること，及び抽出に用いた HTML タグが単なる文字の強調にも用いられるものであることからノイズを含む．そこで統計値により，得られた属性候補をフィルタリングする．

我々は，予備実験で獲得された属性候補を，既存手法 [7, 4] で用いられている，*df* や *idf*, および相互情報量を用いたフィルタリングした結果を人手で分析した．その結果，*idf* や，相互情報量は，対象のクラスに特化した属性のみを重要視する傾向があることが分かった．実際には，クラスの属性には，クラスに特化した属性（例: 映画の字幕，配給）と，複数のクラスに現れる属性（映画，ドラマ，演劇などの出演）があり，必ずしもクラスに特化した属性を重要視することは望ましくない．一方，*df* [4] については，Web 製作者が，同一のテンプレートから同種のページを大量に生成している場合，それらのページに含まれる属性候補の *df* は結果として非常に大きくなる．もし，不適切な属性候補がそのようなページに現れていた場合，*df* ではそのような属性候補を取り除くことが非常に難しくなる，

本稿で提案する**サイト頻度**は，*df* の問題点を解決するものであり，一般に *df*, *df-idf* や相互情報量といった既存の統計値より有効に働く．属性候補 x のサイト頻度 $sf(x)$ は，以下のように定義される．

$$sf(x) = \text{属性候補 } x \text{ を抽出した web サイトの数} \quad (1)$$

サイト頻度を獲得するためには，Web ページを Web サイトを同値類としてまとめる必要がある．ここで言う Web サイトとは，同一 Web 製作者により記述された Web ページ群のことであり，本研究では Web ページの URL のパスを辿って `/^(?:index|default|main)\..+/`にマッチするファイル名のファイルを最初に含むディレクトリまでのパスを Web サイトとして定義した．サイト頻度は，属性 x をクラスの属性として用いた**Web 製作者の数**を属性の信頼度として表現したものであると言える．

3.2 STEP 1: 対象物を含むページからの属性抽出

我々はまず，対象物の属性情報記述ページの候補として，対象物を含むページを検索エンジンを用いて収集する．対象物 x を含む各ページから抽出された属性が，x のクラスの属性と類似していれば，そのページは属性情報記述ページであると言える．そこで，各ページからまず，3.1.2 節で述べた方法を用いて属性を抽出する．この際，単語の出現位置の制約は用いない．

3.3 STEP 2: クラス属性に基づく属性情報記述ページの発見

属性情報記述ページの候補の Web ページ p について，ページから抽出される属性 $\mathcal{A}_p$ とクラス属性 $\mathcal{A}_C$ とを比較し，属性情報記述ページとしてのスコアを計算する:

$$score(p) = \frac{\#(\mathcal{A}_p \cap \mathcal{A}_C) \times ratio(\mathcal{A}_p, \mathcal{A}_C)}{ave(\mathcal{A}_p, p) \times text\_size(x, p)}$$

表 2: Web から獲得された各クラス上位 15 属性と被験者により提示された属性 (括弧内の数字は，複数の被験者がその属性を提示した場合その人数を表す)

クラス	獲得された属性	被験者の提示した属性
デジタルカメラ	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	価格 値段 メーカー 画素 画素数 重量 記録 サイズ 会社名 充電時間
競走馬	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	戦績 (2) 生年月日 (2) 性別 血統 取得賞金 生産者 (2) 馬主 競争成績 賞金
野球選手	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) 生年月日 (2) ポジション 所属球団 性別 出身校 背番号 打率 身長 出身
俳優	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	名前 生年月日 (2) 出演作品 (2) 出演番組 出身 出身地 (2) 年齢 所属 所属事務所
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 住所 (2) 診療科目 (2) 電話番号 (3) 診療科 診療時間 (2) 診療受付日時
株式会社	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	株価 住所 (2) 所在地 資本金 (2) 電話番号 (2) 売上高 社員数 設立 社長
ワイン	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	種類 (2) 価格 生産年 (2) 産地 生産地 (2) 飲み口 価格 容量 ワイナリー
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	場所 (2) 住所 展示物 料金 入館時間 定休日 入館料 (2) 利用時間 開館時間 電話番号
文庫	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	作者 書名 価格 (3) 著作 出版社 著者 ジャンル 発売日 (2) ISBN
遊園地	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	場所 (2) 住所 入園料 (2) 入場時間 定休日 開園時間 営業時間 入場料 電話番号 アトラクション

ここで，分子の $\#(\mathcal{A}_p \cap \mathcal{A}_C)$ および，$ratio(\mathcal{A}_p, \mathcal{A}_C)$ は，良い属性候補記述ページは，クラス属性の大部分を含むという事実，対象物の属性のほとんどはそのクラスの属性と一致するという事実をそれぞれ表現している．また，分母の $ave(\mathcal{A}_p, p)$ は，ページ中の属性の出現回数が小さいページを選好する目的で追加された項であり，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ている．最後に，$text\_size(x, p)$ は対象物について記述したページを選ぶための項である．対象物を含む最初の HTML タグで囲まれたテキストが短ければ短いほど，ページが対象物について記述したページである可能性が高くなる．このようにして計算されるスコアの最大の候補ページを，属性情報記述ページとして出力する．

4 評価実験

クラス属性を獲得する知識源として，我々はまず GNU wget[1] を用いて 0.7TB（HTML タグ含む）の日本語 web リポジトリ及び全文検索エンジン[2]（以下，*local_search*）を構築した．クラスの属性知識は，*local_search* の出力するクラスを単語として含む全文書からランダムに選んだ 10,000 件を知識源として獲得した．

次に，我々のシステムを評価するテスト用データを構築した．ランダムにクラスと対象物を生成することは困難なため，本研究に関与しない第三者に，彼らが興味があるクラスとその対象物を挙げてもらった．表 2 右がテストに用いたクラスである．対象物については，ページ数の関係で省略する．

4.1 クラス属性の獲得

各クラス C について，節 3.1 の獲得手法により，*local_search* の出力から，ランダムに 10,000 ページを選んで属性候補を獲得し，サイト頻度によるランキング上位 30 属性をクラス属性として用いた．表 2 中央は用いた各クラスについて獲得された属性の上位 15 属性を示したものである．

我々は徳永らにより提案された接尾辞追加質問テスト [4] により，各属性が妥当かどうかを簡単に人手で調べた（* の付加された属性が不適当と判断された）．野球選手以外のクラスについては，ほとんどが妥当な属性であった．野球選手については，属性候補の知識源のページが十分な数得られなかったため，適当な属性が獲得できなかった．

これらの属性を詳しく見ていくと，自然文には現れないような属性が，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，対象物の同じ属性を表現する様々な同義語の属性語に，自然文中には現れない省略されたものもあった（例: 文庫の著者名 ↔ 著）．これは，他の属性と区別する際の曖昧性が無い場合は，属性語がかなり柔軟に省略されることを示唆している．他の例としては，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，（上位 15 属性には含まれていないが）対象物への電車でのアクセスを示す「電車」なども獲得できている．

4.2 評価基準

我々は次に，ユーザが実際に web を検索する際の状況を考慮して評価基準を設計した．まず，各被験者は，各クラスについて知りたい 4 つの属性を対象物を見ることなく決める（表 2 左）．各被験者は次に，我々のシステムを含む 3 つのシステムが各クラスの 10 の対象物について出力したページに知りたい属性-属性値が何対含まれているか評価してもらった．この際に用いた 3 つのシステムは，は，Google の検索結果の第一位を出力するもの（以下，Google），と Google のランキング上位 30 件を候補ページとする提案手法 (SP)，また *local_search* の対象物に対する出力からランダムに選んだ 10,000 件を候補ページとする提案手法 (SP*) の 3 システムである．

各ページの評価は以下の手順で行った．まず，被験者

[1] 商用検索エンジンの出力を知識源とすると，ランキング手法の不透明性から獲得される属性の分析が困難となるため，用いなかった．

[2] FreyaSX: `http://www.delegate.org/freyasx/`

表 3: 属性情報記述ページの発見: 実験結果

クラス	# 対象物	被験者 1 Google	被験者 1 SP	被験者 1 SP*	被験者 2 Google	被験者 2 SP	被験者 2 SP*	被験者 3 Google	被験者 3 SP	被験者 3 SP*	被験者 1-3 (平均) Google	被験者 1-3 (平均) SP	被験者 1-3 (平均) 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**
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

は各ページが対象物を含むページどうかを調べる．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クラスの対象物を参照していると判断した対象物のみを評価の対象とした．次に，3 人の被験者それぞれ，各ページに属性-属性値の情報が，属性（またはその同義語），属性値共に明示的にテキストとしてページに含まれており，かつ属性と属性値の関係が自然文ではなく，視覚的に即座に分かる形で書かれているかどうかを調べてもらい，各ページを含んでいる属性-属性値対の数をポイントとして評価付けしてもらった．[3]

4.3 属性情報記述ページの発見

表 3 は，3 つのシステムが出力した全てのページに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．表中の各カラムは，3 つの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．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判断した．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（約 69%）の属性-属性値関係を含むページを出力した．また，SP* が Google よりも良いページを出力していることに注目されたい．SP* が，Google の 10% 以下の web ページのから選んだ高々 10,000 ページから属性情報記述ページを選んでおり，さらに 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全く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．ただ，Google のランキングはそれでも有用であることは SP が SP* より良いことから分かる．

各クラスの結果をもう少し詳しく見て見ると，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 が高くランク付けするいわゆる権威ページ（例えば，Yahoo! や Amazon.com など）がたまたま包括的なデータベースを持っていたからである．一方，権威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 の結果はあまり属性情報を得るのに役に立たないことが分かる．これは，SP* がワインや博物館，遊園地について Google のランクを使った 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上位にランク付けしてしまっている．

[3]以上の評価基準は，非常に厳しい評価になっていることに注意されたい．というのも，属性名を省略しても，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，属性名が省略されることが多いためである．

5 まとめ

本稿では，与えられた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，クラスの属性の知識ベースを構築する．この知識ベースを用いて，ユーザにより与えられた対象物とそのクラスから検索時に，与えられた対象物を含むページの中から最も属性-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-属性値関係が最も含まれると判断された．

今後の課題としては，属性に基づく属性情報記述ページのスコア関数の改良，また上位語-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などが挙げられよう．

参考文献

[1] A. Almuhareb and M. Poesio. Attribute-based and value-based clustering: an evaluation. In *Proc. EMNLP*, 2004.

[2] H.-H. Chen, S.-C. Tsai, and J.-H. Tsai. Mining tables from large scale html texts. In *Proc. COLING*, 2000.

[3] K. Shimada and T. Endo. Acquisition of new training data from unlabeled data for product specifications extraction. In *Proc. PACLING*, 2005.

[4] K. Tokunaga, J. Kazama, and K. Torisawa. Automatic discovery of attribute words from web documents. In *Natural Language Processing – IJCNLP 2005*, volume LNAI 3651. Springer-Verlag, 2005.

[5] M. Yoshida and H. Nakagawa. Specification retrieval – how to find attribute-value information on the web. In *Natural Language Processing - IJCNLP 2004*, volume LNAI 3248. Springer-Verlag, 2005.

[6] M. Yoshida, K. Torisawa, and J. Tsujii. A method to integrate tables of the World Wide Web. In *Proc. WDA*, 2001.

[7] 高橋哲朗, 乾健太郎, and 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. In **言語処理学会第 *11* 回年次大会論文集**, 2005.