# 商品ページからの属性・属性値抽出と同一商品クラスタリング手法

坂地 泰紀 (豊橋技術科学大学)
小林 暁雄 (豊橋技術科学大学)
関根 聡 (ランゲージ・クラフト研究所/ニューヨーク大学)
竹中孝真 (楽天技術研究所)

## 1　はじめに

Web 上には商品を扱ったページが多数あり，消費者はその商品ページを閲覧して商品の購入を行っている．商品ページには，同一の商品を扱ったものも多数存在しているが，同一の商品を扱っている他のページの検索は消費者自身が行っている場合がある．そこで，本研究では同一商品を探す手間を省くために同一商品を扱っているページのクラスタリングを自動的に行う手法の提案を行う．本研究では，通信販売ショップである「楽天」のデータを対象に研究を行う．

また，商品ページには商品の属性・属性値が含まれていることがある．例えば，ワインに関する商品ページであれば，属性として「品種」，属性値として「シャルドネ」などが出現する．これらの属性・属性値は商品ページのクラスタリングに有効と考えられるため，商品ページからの属性・属性値の抽出も行う．

我々が提案する商品ページのまとめ上げ手法の概要を図 1 に示す．全体については [6] に紹介してある．本論文では，図 1 内の 2. ページ中にある属性・属性値の抽出と，3. クラスタリング，5. 属性名の同義性の発見について説明を行う．

## 2　属性・属性値の抽出

「楽天」の商品ページには，図 2 のような形で属性・属性値が出現する．我々が扱う HTML ソースでみると，「【品種：ヴェルメンティーノ 90%、モスカート・ビアンコ 10%】<br>」という部分の，「品種」が属性となり，「ヴェルメンティーノ 90%、モスカート・ビアンコ 10%」が属性値となる．同じく，「アルコール度数」が属性となり，「12.5%」が属性値となる．これら属性・

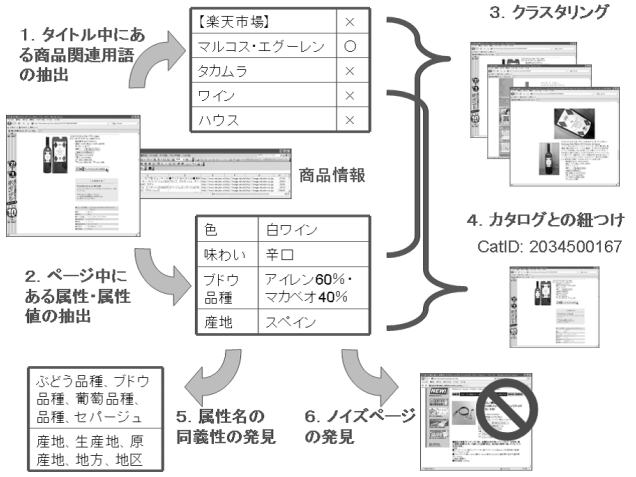


図 1　手法概要


図 2　Web ページ上の属性・属性値

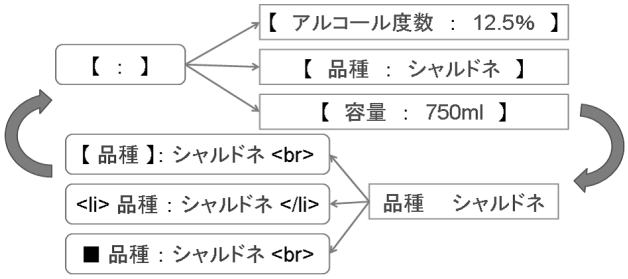


図 3　ブートストラップの概要

属性値の前後に出現する「【 : 】」をパタンとすれば，上記の属性・属性値を抽出することができる．

本研究では，パタンを用いて属性・属性値を抽出する．そこで，パタンを自動獲得する手法の開発を目指す．我々は，解決策として，基礎的技術であるブートストラップ手法を用いて，パタンと属性・属性値を抽出する手法とした．その概要を図 3 に示す．

本手法では，属性とパタンにスコア付けをし，各反復においてスコアの上位 $N$ 個の属性，もしくは，パタンを抽出する．

### 2.1　属性の抽出

パタンを用いて属性・属性値を抽出する部分について説明する．本手法では，パタンにより獲得した属性候補が適切であれば，属性値候補も適切であるという仮定に基づいて抽出を行う．以下の二つの特徴を用いて，適切な属性を選択する．

1. 属性候補が出現した店舗の異なり数
2. カテゴリごとの属性候補の出現確率

1. 属性候補が出現した店舗の異なり数で属性候補にスコア付けし，スコアの高い順に並び替える．その後，2. カテゴリごとの属性候補の出現確率を用いて，不適切な属性候補をフィルタリングする．

**属性候補が出現した店舗の異なり数**

我々は，多種の店舗ページに出現した属性候補は属性として適切であるという仮定に基づいて，属性候補が出現した店舗の異なり数をスコアに用いた．例えば，ワインのページにおいて適切な属性である「品種」という属性候補は多種の店舗ページに出現する．それに対して，いずれかのパタンにマッチした「神の雫」という不適切な属性候補は 1 店舗の商品ページからしか獲得されなかったため，スコアが低くなり，属性として抽出されなくなる．

**カテゴリごとの属性候補の出現確率**

他のカテゴリにおいて同様に出現する属性候補は，属性として不適切であるという仮定に基づいて，フィルタリングを行う．店舗の異なり数で順位を付けた場合，例えば，「送料無料」のような属性候補は多数のページに出現するため，どうしてもスコアが高くなってしまう．そこで，カテゴリごとの属性候補の出現確率を用いる．例えば，「品種」という属性候補はワインカテゴリのページにはよく出現するが，ドライバや靴などのカテゴリには出現しない．

### 2.2 パタンの抽出

属性・属性値を用いてパタンを抽出する部分について説明する．本手法では，様々な属性・属性値と共起するパタン候補は適切であるという仮定に基づいて，エントロピー $H(t)$ を用いてパタンのスコア付けを行う．実際には，以下の式 (1) を用いて計算を行う．

$$H(t) = -\sum_{i \in I} P_t(i) \log_2 P_t(i), \quad P_t(i) = \frac{f(i,t)}{N_t} \qquad (1)$$

ただし，$T$ はパタン候補集合，$I$ は属性・属性値の組の集合とする．$P_t(i)$ はパタン候補 $t$ と属性・属性値の対 $i$ が共起する確率，$f(i,t)$ はパタン候補 $t$ と属性・属性値の対 $i$ の共起数，$N_t$ はパタン候補 $t$ の獲得数とする．

## 3 属性のまとめ上げ

属性の中には同じ概念を持つものが存在している．例えば，「品種」「ぶどう品種」「ブドウ品種」「セパージュ」「葡萄品種」は全て同じ概念の属性である．我々は，同一の属性値を持つ属性は同じ概念であるという仮定に基づいて，まとめ上げを行った．例えば，A 属性と B 属性は同一の属性値を持っていたら，同じ概念であると判定する．しかしながら，この仮定だけでは，うまくまとめ上げを行うことができなかった．例えば，ゴルフドライバで「360g」という属性値は，属性「総重量」の属性値にも，属性「ヘッド重量」の属性値にも含まれていたため，「総重量」と「ヘッド重量」が同じ概念であると誤って判定されてしまった．さらに，例えば，ワインカテゴリにおいて属性としては不適切な「白」という語が属性として抽出されていた場合，それが悪影響を及ぼした．この例だと，ページ中に「【白：辛口】」と出現したため，誤って属性として「白」，属性値として「辛口」を獲得しまっていた．そのため，「辛口」を属性値として持つ「タイプ」が「白」と同じ概念であると誤って判定されてしまった．

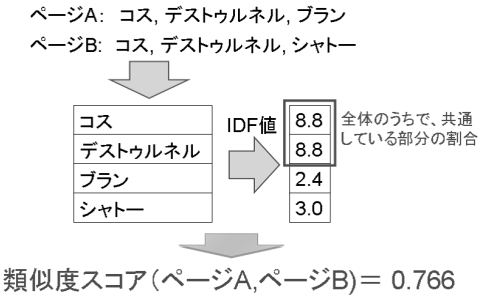

図 4 類似度スコア

そこで，同じ概念らしさのスコア付けを行い，属性のまとめ上げの実験を行った．すでに，大前ら [5] は，ジャッカード係数を用いた属性のまとめ上げについて提案している．本研究では，ジャッカード係数を含めた様々なまとめ上げ手法を比較実験することにより，最適なまとめ上げ手法を検討する．具体的には，以下の 4 つの手法を比較した．

each A 属性の属性値の中で B 属性が持っている属性値と共通なものの割合と，B 属性の属性値の中で A 属性値が持っている属性値と共通なものの割合を掛け合わせ，それをスコアとする．

ent $each$ の割合を元にエントロピーを計算し，掛け合わせ，それをスコアとする．

jac 大前ら [5] と同様に，ジャッカード係数を計算し，それをスコアとする．

abs A 属性と B 属性の属性値中で共通なものの種類数をスコアとする．

## 4 商品ページクラスタリング

本研究では，二つの商品ページ組に対して類似度スコアを付与することにより，商品ページのクラスタリングを行う．ページタイトルには，商品の名前やメーカー名が記述されていることが多い．タイトルには，謳い文句などの商品名やメーカー名以外の情報も記述されていることがあり，これらを正確に抽出することは難しい．そこで，商品名やメーカー名などの商品ページクラスタリングに役立つ用語を商品関連用語と定義し，これを抽出する．商品関連用語の抽出に関しては小林ら [4] の手法を用いる．小林らの手法は，まずページタイトルにフィルタリングを行い商品関連用語の候補を獲得する．その後，商品関連用語の候補から商品関連用語を抽出している．

商品ページ $a$ と $b$ の組の類似度スコア $S_{ab}$ を以下の式 (2) により計算する．

$$S_{ab} = \frac{\sum_{i \in I_{ab}} IDF_i}{\sum_{u \in U_{ab}} IDF_u}, \quad IDF_l = \log \frac{N}{N_l} \qquad (2)$$

ただし，$U_{ab}$ は商品ページ $a$ と商品ページ $b$ の商品関連用語の集合，$I_{ab}$ は商品ページ $a$ と商品ページ $b$ において共通の商品関連用語の集合とする．$IDF_l$ は商品関連用語 $l$ の IDF 値であり，全商品ページ数 $N$ と商品関連用語 $l$ が出現するページ数 $N_l$ により求める．図 4 に，スコア付けの概要を示す．

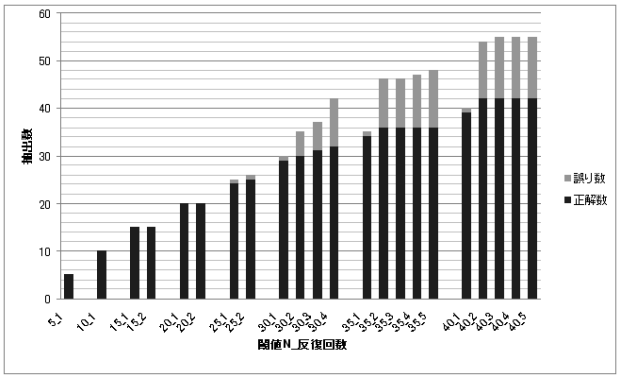


図 5　ワインの属性抽出結果

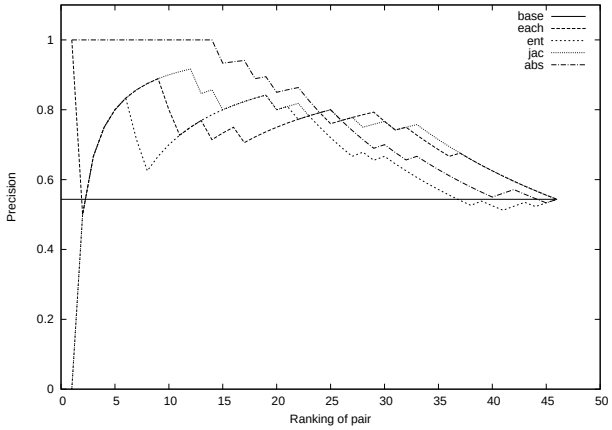


図 6　ワインの属性のまとめ上げ結果

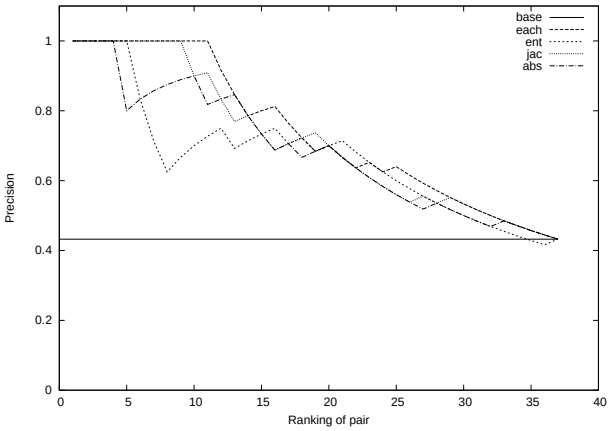


図 7　ドライバの属性のまとめ上げ結果

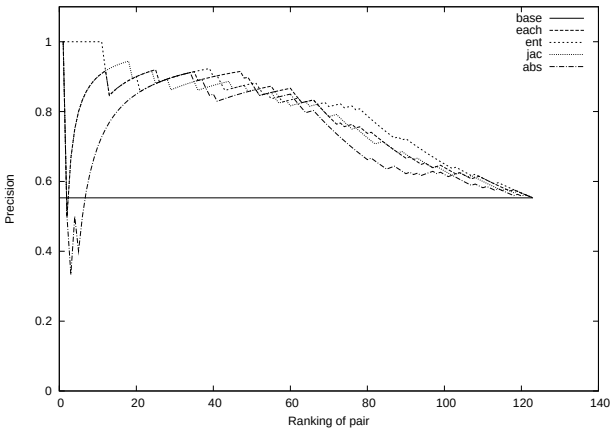


図 8　靴の属性のまとめ上げ結果

## 5　評価実験

本手法を評価するために，実験を行った．属性・属性値抽出，属性のまとめ上げ，商品ページクラスタリング，それぞれ以下の節で示す．

### 5.1　属性・属性値抽出

属性・属性値抽出の実験には，白ワインカテゴリは 20,667 商品ページ，ゴルフドライバは 23,824 商品ページ，男性用靴は 119,928 商品ページを実験に用いた．図 5 に白ワインの属性の抽出結果を示す．ゴルフドライバと男性用靴の結果は関しては，白ワインと同様の結果になったため，割愛する．

図 5 より，閾値が低いときは精度が高く，閾値が高くなるとノイズが増え，精度が低くなっていることが分かる．結果を分析した結果，属性のフィルタリングは効果的であったが，パタンのフィルタリングについては効果的とはいえず，今後の課題となる．

### 5.2　属性のまとめ上げ

属性のまとめ上げ実験には，属性・属性値抽出で獲得した属性・属性値を用いた．白ワインは閾値 40 反復回数 1 で抽出した属性の組 46 組 (正解組数：26 組) を用いた．ゴルフドライバは閾値 35 反復回数 1 で抽出した属性の組 37 組 (正解組数：1 組) を用いた．男性用靴は閾値 35 反復回数 1 で抽出した属性の組 123 組 (正解組数：68 組) を用いた．図 6，図 7，図 8 に属性のまとめ上げ結果を示す．

図 6，図 7，図 8 の横軸はスコアが高い順に並べた属性組で，縦軸はそのときの精度を表す．例えば，横軸が 10 のときの *abs* は精度 1 である．これは，*abs* におけるスコアの高い順の上から 10 組の属性組は，全て同じ概念の属性組であったことを意味する．それに対して，*ent* では，10 のとき精度が 0.7 であり，これは *ent* におけるスコアの高い順の上から 10 組の属性組のうち 7 組が同じ概念の属性組であったことを意味する．

図 6，図 7，図 8 より，ワインに関しては *abs* が優れ，ドライバに関しては *each*，靴に関しては *ent* が優れていた．また，ワインの場合を除いて，各手法間に大きな差がなかったため，今後の分析が必要である．

### 5.3　商品ページクラスタリング

商品ページクラスタリングの実験には，白ワインに関する 20,667 商品ページから 282 組の商品ページ組をサンプリングし，それを用いた．商品ページクラスタリングの評価実験では，提案手法を *selected_keywords* とする．また，商品関連用語ではなく，商品関連用語候補を用いた手法を *all_keywords* として実験を行った．二つの商品ページ間に出現する商品関連用語の IDF 値の中で最大のものをスコアとする手法 *max_idf* と，ランダムに抽出する手法 *base* をベースラインとした．理想的に最も良い結果を *best* とした．図 9 に商品ページクラスタリングの結果を示す．

図 9 の横軸は，スコアが高い順に並べた商品ページ組で，縦軸がそのときの累積正解組数を表す．例えば，横軸 50 のとき (スコアの高い順に上から 50 組を獲得することを意味する)，

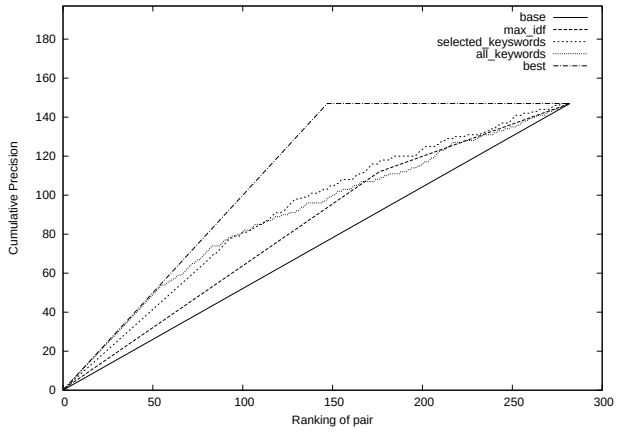


図 9　ワインページのクラスタリング結果

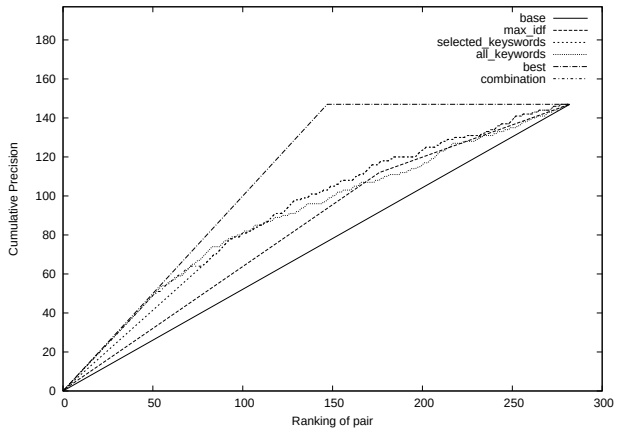


図 10　combination の結果

*all_keywords* は 50 組全てが正解で，*selected_keywords* は約 40 組が正解，*max_idf* は約 30 組が正解，*base* は約 22 組が正解ということを示す．つまり，*best* に近づくほど良い手法であるということである．

図 9 より，上位の結果では *all_keywords* の結果が良く，その後，*selected_keywords* が良いという結果になった．これは，タイトルが完全に一致した場合，同一商品である可能性が高いということに起因する．そこで，最初は *selected_keywords* で順位付けした後，同点の場合には *all_keywords* の結果を用いるという手法 *combination* を実装し，評価した．その結果を図 10 に示す．図 10 より，*selected_keywords* と *all_keywords* の良い特徴をとらえ，それを反映させた結果となった．

## 6　関連研究

Tokunaga らは，自然文を対象として，対象物とその属性語が共起するパタンを用いて属性語を抽出する手法を提案している [1]．また，属性語を抽出するために属性情報が記述されたページを発見する手法を吉永らは提案している [7]．しかしながら，これらの手法は属性を抽出するだけで，我々の提案手法のように属性・属性値の対を抽出することができない．

また，Web 上の表や DOM ツリーなどのレイアウトには，直接的に属性・属性値が記述されている場合があり，それらを対象に属性・属性値を抽出する手法がいくつか提案されている [2][3][5]．大前らは，Web 上の表を対象に $\chi^2$検定 を用いた属性・属性値の獲得と，ジャッカード係数を用いた属性のまとめ上げを行う手法の提案をしている [5]．Yoshida らは，表形式の集合を入力とし，EM 法を用いて表中の属性・属性値を自動的に認定し，その後，クラスごとに属性・属性値をまとめ上げる手法を提案している [2]．鶴田らは，あるカテゴリに属する不均一なフォーマットで記述された属性・属性値を同一ジャンルのウェブサイトのトップページ URL の集合，及び，リンクを辿る際のヒント文字列を用いて自動的に抽出する手法を提案している [3]．しかしながら，これらの研究は表や DOM ツリーなどのレイアウト情報だけを対象にしているため，HTML に直接記入されている属性・属性値情報を獲得することができない．

## 7　まとめ

本研究では，同一商品を扱っているページのクラスタリングと，属性・属性値の抽出を自動的に行う手法の提案を行った．

## 謝辞

今回の研究の機会を与えてくださり，貴重なデータを提供頂いた楽天株式会社様には感謝致します．特に，安武様，森様，三條様には共同研究の設定，西岡様，平手様にはディスカッションにて貴重な意見を頂きました．また，本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の支援を受けた．

## 参考文献

[1] Kosuke Tokunaga, Jun'ichi Kazama, and Kentaro Torisawa. Automatic discovery of attribute words from web documents. In *proc. of IJCNLP 2005*, pp. 106–118, 2005.

[2] Minoru Yoshida, Kentaro Torisawa, and Jun'ichi Tsujii. Extracting ontologies from world wide web via html tables. In *proc. of PACLING*, pp. 332–341, 2001.

[3] 鶴田雅信, 関根聡, 増山繁. 企業の公式 web サイトからの基本情報抽出. 人工知能学会第 23 回全国大会 (JSAI 2009), 2009.

[4] 小林暁雄, 坂地泰紀, 関根聡, 竹中孝真. ショッピングサイトの商品ページタイトルからの商品関連用語の抽出と商品カタログへの商品ページの紐付け手法. 第 15 会言語処理学会年次大会, 2010.

[5] 大前信弘, 黄瀬浩一. Web の表を対象とした属性の自動識別. 情報処理学会研究報告, pp. 43–48, 2006.

[6] 関根聡, 小林暁雄, 坂地泰紀, 竹中孝真. ショッピングサイトにおける商品の同一性、類似性の推定手法. 第 15 会言語処理学会年次大会, 2010.

[7] 吉永直樹, 鳥澤健太郎. Web からの属性情報記述ページの発見. 人工知能学会論文誌, Vol. 21, No. 6, pp. 493–501, 2006.