# 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

小林 暁雄 (豊橋技術科学大学　kobayashi@smlab.tutkie.tut.ac.jp)
坂地 泰紀 (豊橋技術科学大学　sakaji@smlab.tutkie.tut.ac.jp)
関根 聡 (ニューヨーク大学　sekine@cs.nyu.edu)
竹中 孝真 (楽天技術研究所　takamasa.takenaka@mail.rakuten.co.jp)

## 1　はじめに

サイバーモールのような，様々な商店が参加するタイプのショッピングサイトでは，各商品のページを，ショッピングサイトに出店している各業者が独自に作成している．このため，消費者は数多くの商品ページの中から欲しい商品を検索するが，業者間でページの表記方法などの統一は行われていない場合が多く，消費者は同一商品と思われる様々なページを閲覧して，購入するサイトを決定する必要がある．特に，各ショップは，他店よりも消費者にアピールするため，商品ページのタイトルに【送料無料】などの情報を付与したり，商品説明をメーカーサイトの商品ページを画像として保存したものを掲載するなどしている．これにより，消費者は，単純にクエリ検索するだけでは，欲しい商品のページ全てを発見することは難しく，ページタイトルも様々な情報で埋め尽くされており，ページタイトルのみから商品を判断することも困難である．このため，商品ページの中から，消費者の望む一品を検出する研究 [1] なども行われている．

このような，商店ごとに異なる表記がなされている商品ページについて，同一商品のページをまとめ上げることができれば，サイバーモールの利便性が向上し，利用者の増加を見込むことができる．そこで，我々はこのような商品ページのまとめ上げの手法を考案する．その手法の概要を図 1 に示す．

本論文では，クラスタリングを行うための技術として，図 1 内の 1. 商品ページタイトルからの商品関連用語の抽出手法と，4. 人手で作成された商品カタログデータへの商品ページの紐付け手法について解説する．手法全体については，[4] において解説する．

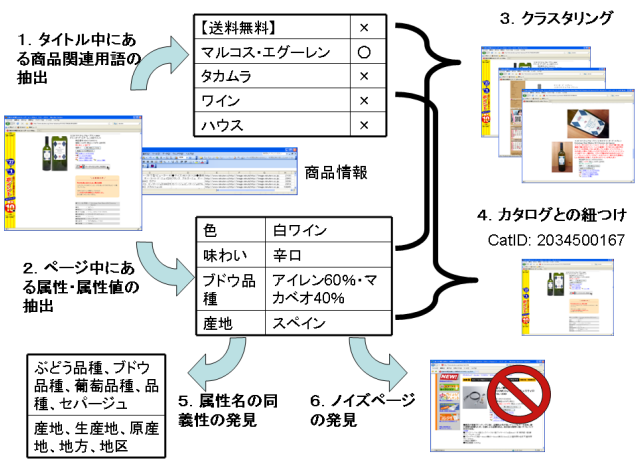


図 1　手法概要

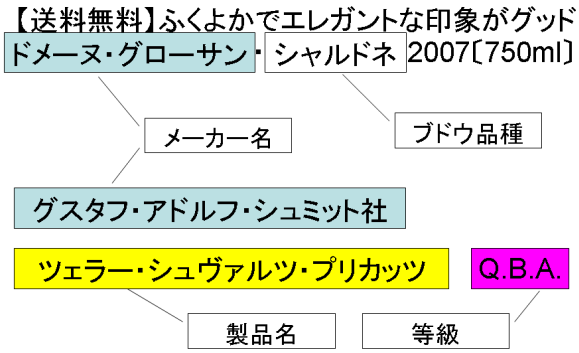


図 2　商品ページタイトルの例

## 2　商品関連用語の抽出手法

ショッピングサイトの商品ページタイトル（以下，タイトルと略記）は，ほぼ全てにおいて商品名が含まれており，同一商品のクラスタリングのための重要な情報源となると考えられる．しかし，ショッピングサイトに出品している商品を取り扱う業者の多くは，他業者よりも商品ページの注目を集めるために，タイトルに様々な情報を付与しており，その箇所がノイズとなり，タイトル中の商品名の判断を自動的に行うことは困難である．図 2 に，タイトルの例を示す．

タイトルの例にあるように，「【送料無料】」や「ふくよかでエレガントな印象がグッド！」といった部分は商品名ではないため，ノイズとなっている．一方で，一つ目の例と二つ目の例を比較すると，二つ目の例にはメーカー名，ブランド名などで商品名と思われる部分が構成されていることが分かるが，一つ目の例にはブランド名がない．このように，メーカー名やその他の属性も含めて商品名とするのか否かといった判断は，商品のドメインや生産国などによって異なり，どこからどこまでが商品名であるかを判断することも困難である．また，タイトル中の単語がメーカー名であるか否かといった判定も，商品名を抽出する上で必要となる．

しかし，同一商品をクラスタリングする場合，タイトル中の商品名の同定は必ずしも必要ではない．図 2 の一つ目の例であれば，「ドメーヌ・グローサン・シャルドネ」という語が獲得できれば，このような語を使用してクラスタリングを行うことが

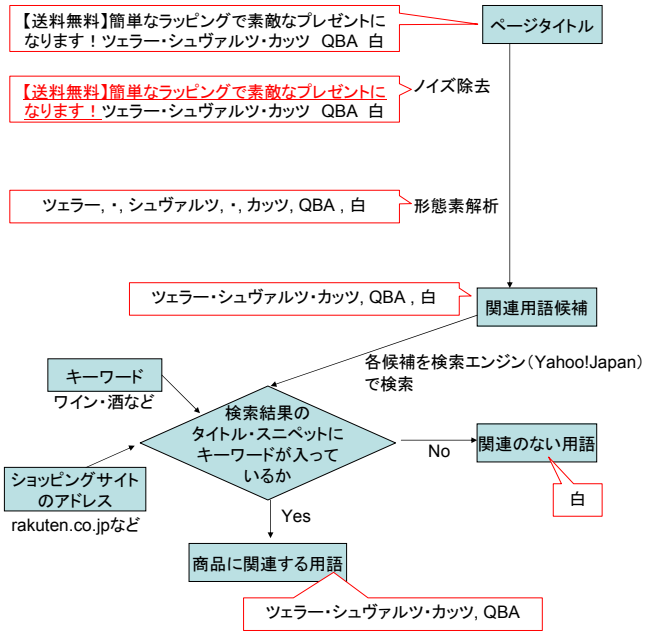


図 3　商品関連用語抽出手法概要

できると考えられる．そこで，我々は，このような語をタイトル中の商品に関連した用語であると判断し，この商品関連用語をタイトル中から抽出する手法を考案した．

### 2.1　商品関連用語の判定

直感として，タイトル中に出現する，商品に関連していると思われる用語を判断する際には，Web 検索エンジンによって検索し，その検索結果を見ることで，その用語が商品に関連した用語であるかどうかを判断できると考えられる．本手法では，この直感から，商品関連用語の候補をタイトルから抽出し，それを Web 検索エンジンの検索クエリとすることで，その検索結果を取得し，その結果中に，商品の属するドメインと同じドメインと考えられるショッピングサイトが出現しているか否かで商品関連用語か否かを判断する．本手法の概要を図 3 に示す．

図 3 に示すように，まず括弧や記号などでくくられている文字列をノイズとして除去する．これは，「【送料無料】」等の商品に関連の低い用語は，括弧でくくるなどして強調されやすいと考えたためである．また，ページタイトルを構文解析した際に，動詞句を主辞とする文節と，その文節に掛かっている文節は，商品を修飾する文になっていると判断し，これもノイズとして除去することにした．

ノイズ除去された，タイトル中の残りの箇所を形態素解析して単語に分割したものについて，名詞連続や中黒を間に挟んだ名詞列は一つの複合名詞であると判断し，これを結合する．こうして取得された用語を商品関連用語の候補として，それぞれ Web 検索エンジンによって検索し，その検索結果上位に出現しているショッピングサイトのタイトル・スニペットに，ドメインに関係するキーワードが含まれているならばその用語を商品関連用語とする．検索結果中にショッピングサイトがない，結果にショッピングサイトがあっても，タイトルやスニペットにキーワードが含まれていないならば商品と関連のない用語と判断する．

表 1　商品関連用語抽出実験設定

| | |
|---|---|
| 形態素解析器 | MeCab |
| Web 検索 | Yahoo!検索 API を用いて Yahoo!Japan にて検索 |
| キーワード | 白ワイン:ワイン，酒 |
| | ゴルフドライバー:ゴルフ，ドライバー，クラブ |
| | 男性用シューズ:シューズ，靴，男性用，メンズ，Men ’ s，MEN ’ S |

表 3　商品関連用語抽出実験結果

| ドメイン | 結果項目 | 結果 |
|---|---|---|
| 白ワイン | 抽出ページ数 | 22749/23614(96.3%) |
| | 精度 | 99.4% |
| | 再現率 | 94.2% |
| ゴルフドライバー | 抽出ページ数 | 25338/25837(98.1%) |
| | 精度 | 98.1% |
| | 再現率 | 97.2% |
| 男性用シューズ | 抽出ページ数 | 117826/134122(87.8%) |

### 2.2　商品関連用語抽出実験

本手法により，実際に商品関連用語をショッピングサイトから抽出する実験を行った．実験対象は，楽天市場の「白ワイン」，「ゴルフドライバー」，「男性用シューズ」のドメインの各商品ページである．それ以外の実験設定を表 1 に示す．

### 2.3　実験結果

実験結果の例を表 2 に示す．また，各ドメインにおいて，商品関連用語を抽出できたページの割合と，人手でタイトル中のメーカー名とブランド名にタグをつけたデータに対し，タグ内に含まれる商品関連用語を一つでも抽出できていれば正解として，精度と再現率を求めた結果（男性用シューズについては，人手による正解データが無いので割愛）を表 3 に示す．

結果から，男性用シューズドメインは少し商品関連用語が抽出できたページの割合が低いが，全体として概ね大半のタイトルから商品関連用語を抽出することができた．これにより，クラスタリングの際にも，商品ページの取りこぼしを少なくすることができる．

### 2.4　エラー解析

実験結果中にいくつか本手法における商品関連用語の抽出ミスが発生していた．その詳細は以下の通りである．

**形態素解析間違い**

固有名詞の形態素解析結果に間違いが発生したことにより，商品関連用語の抽出が失敗した（例:はこだてわいん→はこ, だ, て, わい, ん）．

**ノイズ除去による商品関連用語の損失**

構文解析を利用した，動詞句を主辞とする文節と，そこに掛かる文節の除去において，商品関連用語自体がそのような文節に掛かっている場合や，構文解析の際に固有名詞を解析ミスし，動詞句と誤認識された場合に，商品関連用語自体がノイズとして取り除かれてしまうケースがあった（例:クレマン　ド　ロワール　（モンムソー）J.M.MONMOUSSEAU CREMANT DE LOIREANA のファーストクラスに採用された実績のある!! →「ある」が動詞句であり，商品関連用語「クレマン ド ロワール」も除去されてしまった）．

表 2　商品関連用語抽出実験結果例

| ドメイン | 項目 | 出力結果の例 |
|---|---|---|
| 白ワイン | タイトル | フレッシュ＆フルーティー!白ワインのヌーヴォー 2・JJ モルチェ　ミュスカデ・ヌーヴォー |
| | 関連用語 | ワイン, ヌーヴォー, モルチェ, ミュスカデ・ヌーヴォー |
| ゴルフドライバー | タイトル | SRIXON ZR-800SV-3016J T-65 シャフト |
| | 関連用語 | SRIXON,ZR,3016J,65 シャフト,800SV |
| 男性用シューズ | タイトル | havaianas Mens ハワイアナス メンズビーチサンダル Camoflada OliveGreen |
| | 関連用語 | メンズビーチサンダル,Camoflada,OliveGreen |

#### 2.4.1　商品関連用語の候補が商品に関連の低い用語ばかりである場合

商品関連用語の候補を Web 検索した際に，ショッピングサイトが出現しないような一般語ばかりである場合や，商品のドメインとあまり関連が深くない用語ばかりであったために，商品関連用語を少数しか取得できないケースがあった（例:アルファロメオ　ワイン →「アルファ」,「ロメオ」がワインドメインと関係のない検索結果しか出現できなかったため，商品関連用語が「ワイン」しか抽出できなかった）.

### 2.5　まとめ

形態素解析方法やノイズ除去手法について見直す必要があるが，概ね商品関連用語を正しくかつ殆どの商品ページから抽出することができた．今後はエラーを減少する手法について検討する必要がある．また，[2, 3] などの手法を参考に，商品ページから属性・属性値を抽出し，同一ページのまとめ上げへの応用方法の検討や，[5] の手法を参考に，Web 検索部分の改良などを検討する必要がある．

## 3　商品ページのカタログデータへの紐付け手法

ショッピングサイトのページを，人手で構築されたカタログデータに対応付ける手法について解説する．カタログデータは表 4 のようなデータであり，製品名や価格，発売日といった製品情報に関する属性や，ショッピングサイトにおけるジャンル ID などの，ショッピングサイト内での管理用の属性などが付与されたデータとなっている．このようなカタログデータに商品ページを自動で対応付けることで，商品ページそのもの同士を対応付けるよりも，ショッピングサイト内で同一商品の管理がしやすくなると共に，様々な属性情報を商品ページに付与することができるので，カタログデータを編集することで，ショッピングサイト内での商品検索や商品の推薦など，様々な応用に役立てることができる．本手法の概要を図 4 に示す．

図 4 に示すように，本手法では，カタログデータ中の製品名について，それを構成している単語がそれぞれ商品ページのタイトルと説明文に含まれているか否かで商品ページとカタログデータの対応付けを行う．商品データ・カタログデータはあらかじめメーカー毎に分類して同じメーカーで対応付けた上で，カタログデータ中の各製品と商品ページとの対応付けを行う．カタログデータの製品名に含まれる単語は，それぞれ同一メーカーの各製品名を 1 ドキュメントとして IDF 値を求め，その高いものを含む製品名から商品ページとの対応付けを行う．製品名中の単語全てが商品ページのタイトル・説明文に含まれていない場合，先ほど求めた優先度が低い語を取り除いた部分単語列との対応を調査する．このようにして，優先度順に全ての単語が対応する部分単語列が得られるまで，全ての部分単語列と商品ページの対応を判定する．一致する部分単語列が，他の製品名にも出現する場合，その部分単語列を含む全ての製品のカタログデータと商品ページを対応付ける．

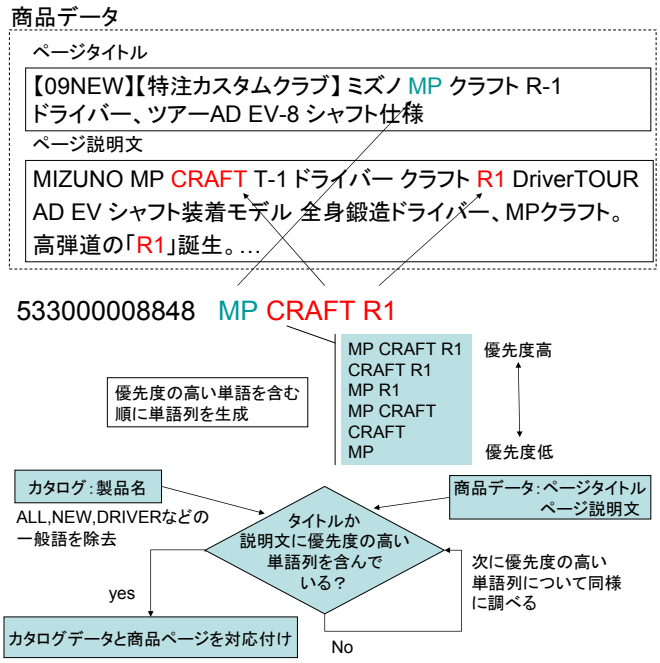


図 4　商品ページのカタログデータへの紐付け手法概要

### 3.1　商品ページのカタログデータへの紐付け手法実験

本手法により，商品ページとカタログデータとの対応付け実験を行った．この際，データには，商品ページに「ゴルフドライバー」のページのうち，人手で正解が付与されたデータ 1,513 ページと，ゴルフドライバーのカタログデータ（収録ドライバー数 338 件）を使用した．前処理として，一般語と思われる語「All,New,Driver」をカタログデータの製品名からあらかじめ取り除いた上で実験を行った．また，製品名中にアルファベット一文字のみの単語を含む場合，この一文字からなる部分文字列については，ノイズとなることが多いと考えられるため，商品ページとの対応付けの判定を行わないことにした．さらに，製品名の各単語について，「-」,「・」を含む場合，これらの文字の箇所で単語を分割した単語列についても商品ページとの対応付けの判定を行った．

### 3.2　実験結果

実験結果を表 5 に示す．また, 対応付けの例を表 6 に示す．

結果から，多くの場合で正解を含むカタログデータと商品ページを対応付けることができた．

表 4　カタログデータの例

| カタログ ID | ジャンル | メーカー | 製品名 |
|---|---|---|---|
| 533000008848 | ドライバー (201706) ＞ ミズノ (201721) | ミズノ | MP CRAFT R1 |
| 533000008847 | ドライバー (201706) ＞ ミズノ (201721) | ミズノ | MP CRAFT T1 |
| 533000008849 | ドライバー (201706) ＞ ミズノ (201721) | ミズノ | MP CRAFT S1 |
| 533000010405 | ドライバー (201706) ＞ キャロウェイ (201710) | キャロウェイ | FT-iQ |
| 533000010408 | ドライバー (201706) ＞ キャロウェイ (201710) | キャロウェイ | LEGACY AERO |

表 5　カタログデータへの商品ページの紐付け実験結果

| | |
|---|---|
| 全データ数 (カタログに載っているドライバー) | 1513 |
| (1) 正解と出力が完全一致 | 286 |
| (2) 出力中に正解を含む | 910 |
| (3) 出力中に正解を含まない | 258 |
| (4) 出力なし | 59 |
| 精度 ((1),(2) を正解として計算) | 82.3% |
| 再現率 ((1),(2) を正解として計算) | 79.0% |

表 6　カタログデータへの商品ページの紐付け結果の例

| | |
|---|---|
| タイトル | MIZUNO レディース JPX-E310-1W 12.5 度【OUTLET-JPX-E310-1W】MIZUNO レディース JPX-E310-1W 12.5 度 |
| 一致単語列 | E310,JPX,MIZUNO |
| 対応カタログ | 532000014113 MIZUNO JPX E310 |
| タイトル | 【送料・手数料無料】テーラーメイド r7 スーパークワッド TP ドライバー (QUAD-TROTECH 65)【SPAP0115P05】 |
| 一致単語列 | TP,r7, クワッド と 460,TP,r7 と SUPERQUAD,TP,r7 |
| 対応カタログ | 533000000101 r7 SUPERQUAD TP |

### 3.3　エラー解析

紐付け実験において不正解となった対応付け結果について，その原因について解説する．

**商品ページにおける商品名の表記ゆれ，カタログデータにおける製品名の表記ゆれ**

商品ページの表記が「ナノ V」であるのに対し，正解のカタログデータでは，「ナノブイ」と表記されているケースや，カタログデータ中で「SasQuatch」と「SQ」のように表記が統一されていないことに起因する紐付けミスが発生していた．

**優先度の高い語が他の商品に含まれている場合**

「FCT」という単語は，製品名では，テーラーメイド社の「XR FCT」という製品にしか出現しない．しかし，商品ページでは，同社の「R9」という商品のシリーズにおいても「FCT」という単語が出現する場合が多く，そのような商品ページは，単語の優先度から，正解である製品名「R9」ではなく，「XR FCT」に対応付けられる対応付けミスが発生していた．

**製品の特別仕様と一般仕様における優先度差による紐付けミス**

単語の優先度のみで対応付けする順を決定しているため，製品名「PRGR GN 502」と「PRGR GN 502 Tour」では，「PRGR」「GN」「502」の 3 単語はどちらの製品にも出現するため，「Tour」と比べて優先度が低く，特別仕様よりも一般仕様の方が優先度が下がってしまい，一般仕様の製品が正しい対応付け先であっても，商品ページ中に「Tour」という語が出現する場合，対応付け先を特別仕様の製品のカタログデータとしてしまう紐付けミスが発生していた．

**その他のエラー**

商品ページにおいて，他の製品を比較対象として紹介している場合に，そちらの製品名の方が優先度が高かったため，そちらのカタログデータに対応付けられてしまうミスが発生していた．また，製品名に含まれる，数値のみが一致したため，異なる製品カタログデータと商品ページが対応付けられてしまうミスが発生していた．

### 3.4　まとめ

多くの商品ページを正しいカタログデータと対応付けることができた．しかしながら，いくつかの紐付けミスは残っているため，今後はこれらのミスを解決していく必要がある．

## 4　まとめ

タイトルからの商品関連用語抽出については，商品関連用語を正しく，かつ殆どの商品ページから抽出することができた．カタログデータと商品ページの紐付けについては，多くの商品ページを正しいカタログデータと対応付けることができた．今後は，どちらもエラーを解決していくことが課題である．

**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いただいた楽天株式会社様に感謝致します．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いただきました．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」による支援をいただきました．

## 参考文献

[1] Dan Shen, Xiaoyuan Wu, and Alvaro Bolivar. Rare item detection in e-commerce site. In *WWW 2009 MADRID!*, 2009.

[2] Kosuke Tokunaga, Jun'ichi Kazama, and Kentaro Torisawa. Automatic discovery of attribute words from web documents. In *IJCNLP 2005*, Vol. 3651, pp. 100–118, 2005.

[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. *The 23rd Annual Converence of the Japanese Society for Artifical Intelligence*, 2009.

[4] 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.

[5] 本間大輝, DanuShka Bollegala, 松尾豊, 石塚満. Web を用いた人物の別名抽出. NLP 若手の会第 2 回シンポジウム, 2007.