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

関根聡
ランゲージ・クラフト研究所
ニューヨーク大学

小林暁雄　**坂地泰紀**
豊橋技術科学大学

竹中孝真
楽天技術研究所

1.　リアルな文書分類問題

サイバーモールのような、様々な店舗が参加するタイプのショッピングサイトには数多くの店舗から様々な商品が出品されている。商品のタイトルや写真を含めたページの作成はすべて店舗にゆだねられ、売上を上げるための工夫が自由にされている。そのため、店舗をまたいだ同一商品の検索は非常に難しく、ユーザーの利便性が損なわれている。例えば、ページのタイトルの欄には商品名だけではなく、人目を引く表現が挿入されていたり（「送料無料」やポイントの表現など）、様々な修飾語や関連情報が付いていたりし、商品名を同定することさえも容易ではない場合が多い（図 1）。これらの問題を解決するために、電化製品や DVD など商品カタログが既に存在する場合には、店舗に商品カタログ ID を入力してもらい、正確な商品の情報を得ようと試みているが、現実的にはすべての店舗が商品カタログ ID を入力しているわけではなく、間違えも含まれる可能性がある。もちろん、酒類や衣料品などカタログを作ること自体が困難な種類の商品も存在する。

そこで、本研究では店舗が独自に作成した商品のページを元に、商品の同一性、類似性を推定する手法について報告する。また、カタログが存在する場合には、カタログとの紐つけも試みてみた。同時に、ショッピングサイトにとって有用な情報を作成することができたので、それらについても報告する。

ここで扱っている問題は、リアルな文書分類問題として捉えられる。つまり、ショッピングサイトにある商品ページが文書群であり、その中での同一性、類似性を推測するという問題である。研究を進めるとリアルであるからこその様々な興味深い問題に直面した。また、本問題は WePS 等（WePS　HP)で行なわれている人名の曖昧性解消の問題において、複数のページが同一の人物を指しているかどうかを認識する問題を、ショッピングサイトの商品を対象に行なったものとも言える。

図１．同一商品を扱う２つの異なるページ　（上段はページタイトル）

2. 手法

同一性、類似性を推定する手法は大きく 2 つのステージに分けられる（図２）。最初のステージでは、商品ページから推定に有用な情報の抽出をおこなう。ここでは、１）タイトル中にある商品関連用語の抽出、２）ページ中にある属性･属性値の抽出､の２つのタスクを試みた。次のステージではそれらの情報を利用した同一性、類似性の推定を行なう。ここでは、３）タイトルの情報を利用したクラスタリングと、４）カタログ情報との紐付けのためにカタログにある商品情報とタイトルの比較方法について試みた。また同時に、１で抽出した属性・属性値の情報を用いて、５）属性名の同義性の推定の実験と、６）各カテゴリに含まれてしまっている、違った種類の商品（ノイズ）の同定の実験も行なった。

これらの内、1,4 の詳細については、本論文誌の（小林ら 10）に、2,3,5 については本論文誌の（坂地ら 10）に詳細を述べている。

商品情報については、楽天株式会社様から白ワイン、ゴルフドライバー、男性用靴についての情報を頂き、それらを対象に実験を行なった。

2.1. タイトル中にある商品関連用語の抽出

商品ページのタイトルは、図１にあるように、シンプルに商品名だけのものもあるが、大抵は右の例のように、商品に関する様々な形容がついていたり､「送料無料」「今だけポイント３倍」のように、売上アップを図るための様々な表現が付加されていたりする。これらの表現は、商品の同一性を求めるためにはノイズとなる。ノイズを除去するため、表現の IDF、同一店舗のタイトルを比較することなども行なったが、上手くノイズを同定できなかった。そこで、この目的のために WEB 検索を利用する方法を考案した。例えば、一般語である「黒猫」で WEB 検索しても、検索結果の上位にはショッピングサイトのワインの商品ページは現れない。また「商品無料」でもワインのページが現れることはまずない。しかしながら、「ツェラー・シュワルツ・カッツ・プリカッツ」という語で検索した場合には、この商品を扱うショッピングサイトの商品ページが検索結果のかなり上位に出現する。このヒューリスティックを用いて、タイトル中にある商品関連用語の抽出を行なった。実際には、タイトルを形態素解析し、ノイズと明らかに分かる表現を除き、ヒューイスティック

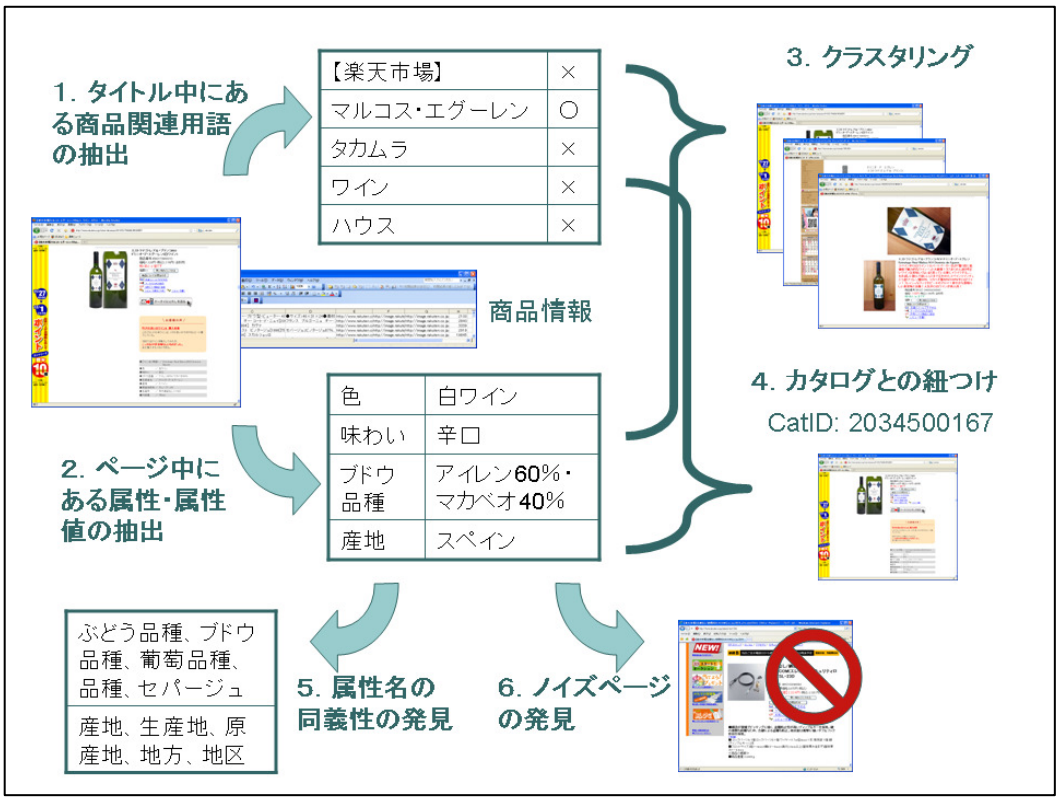

図２．実験手法の概要

で単語切りを修正した表現を Yahoo API を利用して検索した。検索結果の 30 位以内に「ワイン」「酒」などカテゴリに関連する単語がスニペットに含まれるショッピングサイトがあった場合には、その表現は商品関連用語と判断した。評価結果は人手による判断に対し精度 99～98%、再現率 97～94%であった。用語抽出が全くできなかった商品ページの割合は 2～12%で、主な原因は形態素解析誤りであった。

2.2. ページ中にある属性・属性値の抽出

例えば、ワインには色、産地、アルコール度数のような属性があり、商品ページにはそのような属性が書かれていることが多い。このような属性を高精度で抽出することができれば、商品の同一性、類似性判断に役立つと考えられる。商品ページ中の属性の表現を調べた所、それらは以下のような決まったパターンで書かれていることが多いことが分かった。

【品種：シャルドネ】
容量：750ml

全てのパターンを、人手で書き出すことは難しいため、少数のパターンを種にブートストラッピングを行ない、パターンの発見と属性・属性値の抽出を行なった。属性のスコアリングには、その属性候補が現れる業者数を利用したが、そのままだと「送料無料」のように属性でないものも含まれてしまう。しかし、このような表現は複数のカテゴリで出現するため、それぞれのカテゴリで特徴的に出現する表現を属性名として抽出した。また、なるべく多くの属性・属性値と共起するパターン候補はより適切なパターンであると考えられ、このヒューリスティックを利用してパターンのスコアリングを行なった。様々なパラメータセッティングで実験を行なった結果、ワインでは 41 の属性、ドライバーでは 35 の属性を抽出することができた。

2.3. クラスタリング

同一商品を扱う商品ページのクラスタリングを行なった。ここでは、2.1 において抽出した商品関連用語を利用した。商品関連用語はその商品を特定する情報であり、クラスタリングに有効であると考えられる。しかし、抽出した用語の中には「シャトー」「ブラン」といったそのカテゴリにおける一般用語とも言えるような単語が含まれていた。そこで、各用語に対してそのカテゴリ内の IDF 値で重みを付けた。そして、2 つのページの類似性は、2 つのページ間で重複している用語の IDF で重みをつけた割合で求めた。評価データとして、同一商品の組み、そうでない組みが半数ずつくらいある商品ページの組みを用意した。そして、それらをスコア順にソートした結果、同一商品である組みが上位に現れるかどうかで評価した。比較実験として、関連用語抽出を行なわずタイトルにある全ての単語を利用した実験も行なった。その結果、特に上位の結果において抽出した関連用語よりも良い結果が得られた。分析の結果、これはタイトルが完全一致した場合には同一商品であるという事実によるものであった。そこで、最初は抽出した用語の類似度で順位付けした後、同点の場合には、全ての単語での類似度で再順位づけするという方法を試みたところ、その方法が最適であった。ランダムや、共通する表現の中の最大 IDF 値を用いたシンプルなベースラインに比較し、本手法は有意な結果を示した。

2.4. カタログとの紐つけ

今回の実験対象のひとつのゴルフドライバーのように、商品カテゴリによってはカタログが存在する場合がある。そこで、カタログと商品との紐付けの実験を行なった。カタログにある商品名は「MP CRAFT R1」のように商品名とそのマイナーバージョンなどが載っているが、商品ページの説明やタイトルとそのまま完全にマッチする場合は少ない。例えば、商品タイトルには「MP クラフト R-1」や「CRAFT T-1 ドライバー　クラフト R1」などのように記載されている。英語とカタカナの言い換えだけではなく、略語、バージョンの省略、ニックネームなど様々な問題があるが、このような言い換えの問題を正面から解くことは非常に難しい。そこで我々はカタログにある商品名と商品ページの説明文とタイトルの部分マッチを行うことで、商品のカタログへの紐付けを実現した。問題はマッチングにおいてどの部分を優先するかという部分である。我々は、メーカー毎の IDF 値を単語毎に求め、それが高い物を優先することにした。そのようにすると、ミズノのドライバーでは MP よりも CRAFT、CRAFT よりも R1 を優先することになる。実際のマッチングは、最初に全ての単語をマッチさせ、マッチしない場合には、順番に優先度の低い物を除いていった単語群で行なう。最初に単語群がすべてマッチした時の、対応する商品カタログ ID（複数ある場合もある）をその商品ページのカタログ ID とする。その結果、精度は 82%、再現率は 79%であった。エラーの理由としては、

表記揺れの問題や優先度計算の不完全さ、比較のための他の製品名が説明文やタイトルにあった場合があった。

2.5. 属性名の同義性の発見

2.2 で述べたように属性・属性値を抽出したが、属性名の中には、「生産地」「産地」「地方」など、カテゴリ依存の同義の属性名が存在する。これらの同義性を発見できれば、属性名の一般化が行なえ、同一性・類似性の認定に便利である。直感的には、共通の属性値を持つ属性は同じ概念であると考えられるが、実験結果に含まれるノイズの影響などで、単に重複度を計算しても共通な属性名が上手く発見できなかった。そこで、4 つの手法を試みた。それらは重複度合いを掛け合わせたもの、互いのエントロピーを計算し掛け合わせたもの、ジャカード係数、共通属性値の種類数の 4 つである。今回対象にした 3 つのカテゴリで実験した所、カテゴリによって最適な計算手法が異なっていた。またワインの場合を除いては、ほとんどの計算手法の結果が同であった。今後の分析が必要である。

2.6. ノイズページの発見

様々な商店が参加するタイプのショッピングサイトは店舗が自由に商品を出品し、出品するカテゴリの判断も店舗に任されているため、違ったカテゴリに出品してしまう場合がある。例えば、白ワインのカテゴリには、「ワインクーラー」「コンピューターのセキュリティロック」「北海道産小豆」「ドックフード」の商品が出品されている。これらのノイズを除去することはユーザー、店舗、ショッピングサイトの全てにとって有用である。このようなノイズを発見するために、例えば、シンプルに「ワイン」のカテゴリには「ワイン」という単語が入っていないものはノイズであるという判断は有効そうであるが、実際には「ワインクーラー」などワイン関連の商品が発見できないという欠点がある。そこで、2.1 で抽出した属性・属性値を利用することを考案した。つまり、ワインの属性値がまったく現れていない商品はワインではないであろうという仮説である。実際には、2.1 で抽出した属性値の内、頻度が 15 以上の属性値が商品のアブストラクトの中に最大 1 回しか現れていないページをノイズページと判断した所、再現率 100%、空白ページなどを除いた場合の適合率が 48%という結果となった。

3. まとめ

リアルな応用の場において、言葉を扱う有用な技術を開発した。リアルな世界には、評価を目的に設計された人工的な課題設定と違い、思いもしない問題が存在する。しかし、人間は柔軟にその問題を解決し、いとも簡単に自分に有用な情報のみを認識しているように思える。この仕組みを自動的にコンピューターに行なわせることができれば、リアルな世界で役に立つ様々なシステムを開発することにつながるであろう。

本論文では、ショッピングサイトの商品を対象に、その同一性・類似性を推定する手法を開発した。商品タイトルにある商品関連用語の抽出、属性・属性値が表現されるパターンの発見、クラスタリングや紐つけにおける重要な情報の認識、属性の同義性、ノイズページの同定を行なった。

リアルな問題は非常に興味深い。世の中には言語に関する技術が役立つと思われるようなリアルな問題が溢れており、様々な問題解決の機会があるように思える。今後も、そのような問題に取り組んでいきたいと考えている。

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。

参考文献

小林、坂地、関根、竹中「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 回言語処理学会年次大会(2010)

坂地、小林、関根、竹中「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」第 15 回言語処理学会年次大会(2010)

WePS homepage: http://nlp.uned.es/weps/