商品説明文からの検索語に対する関連語抽出

久保田 敦　佐々木 稔　新納 浩幸
茨城大学工学部情報工学科

1　はじめに

現在のオンラインショッピングのサイトでは登録されている商品を数段階のジャンル分けによって登録することにより、ユーザが商品を検索しやすくしている。しかし、商品には他にも検索する際に役立つ単語が存在する場合がある。それらは商品説明文などに含まれるが、商品と関係のない単語や検索語と関連しない単語とが混在しているため抽出することは難しい。

そこで、本研究では検索結果をさらに絞り込むのに役立つ関連語を抽出する手法を提案する。この手法はまず、ユーザの検索結果 Web ページ内から商品の説明文などに含まれる出現単語を取り出す。次に、取り出した単語の中で頻出する単語に対しジャンル毎の出現頻度や検索語との共起頻度の特徴を調べる。そして、ある特徴を持つ単語を検索語に対する関連語と判断する。

本稿はこの手法に基づき抽出した結果を人手で抽出した結果と比べて評価し、その精度や問題点を考察した結果を報告する。

2　研究背景と関連研究

第 1 節でも示した通りオンラインショッピングのサイトでは複数の商品をサイト独自のジャンルに分割することでユーザの利便性を高めている。しかし、商品によってはそのジャンル以外のジャンル分けに使える単語やその特徴を表す単語など、ユーザが商品検索に使える単語が存在する場合がある。そのような関連語はショッピングサイトでは商品タイトルまたは商品説明文の中に存在している場合が多い。この関連語抽出においてはより商品の特徴を表し、検索に役立つ単語を求めることが重要な問題となる。

このような関連語を抽出する研究はこれまでにいくつか行われている。河野らはブログの記事を商品カテゴリにマッピングさせるためにカテゴリとその特徴語の関連付けを行っている [1]。これは商品カテゴリの特徴語をブログ記事から抽出し、特徴語が出現したブログ記事を商品カテゴリにマッピングする手法である。また、前澤らは商品名寄せシステムの精度向上のために商品タイトル中のフレーズの重要度を判定する手法を提案している [2]。彼らは宣言表現など商品判別に寄与しないフレーズをノイズ、判別に有効なフレーズをシード、ノイズまたはシードのどちらとも言い切れないフレーズを中立フレーズという 3 つのラベルを付けている。しかし、同じ単語でも検索語によって関連語になるかどうかが変わることが目的の本研究ではその手法によって一概に分類することは出来ない。

本研究では前澤らの手法と同様にシードやノイズの判別を検索語に対して行い関連語を抽出する手法を提案する。その際、抽出する範囲を商品タイトルだけでなく商品説明文まで拡げ、より多くの候補から抽出を行う。また、前澤らの研究で言う中立フレーズをなくし、単純にシードであるものを抽出することにする。なぜなら本研究で検索語に対する関連語は、その基準を検索語に対して純粋に関係するものに限定したためである。そのため、中立フレーズのように関連語かもしれないという分類はせずノイズかシードの 2 つに分類できる。

3　提案手法

3.1　概要

検索語に対する関連語は検索語との関連性があると同時に商品検索に役立つ単語でなくてはならない。よって、本研究ではユーザの検索結果 Web ページ内の URL を入力とし数段階の処理によって関連語を出力する。

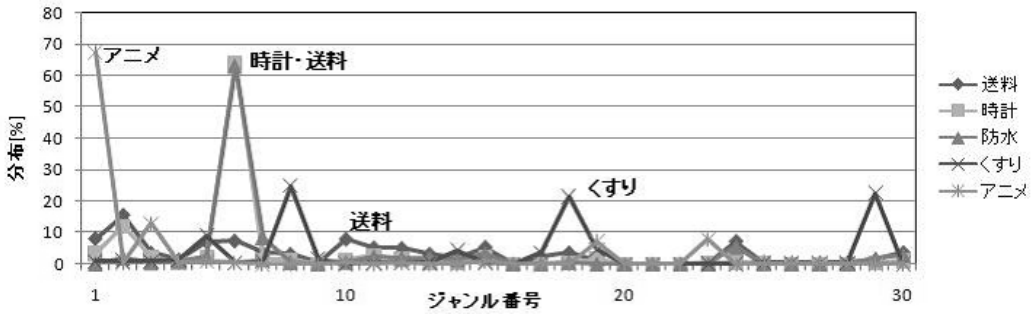

図 1　ヒット商品の分布

3.2　説明文からの単語抽出

本研究では商品の説明文から検索語に対する関連語の候補を見つけ出す。まずショッピングサイトの API を用いて検索語に対する検索結果を取得する。

次にその結果に含まれる商品タイトル、商品説明文を Mecab [1]で形態素解析する。その際以下のような条件の下で出現単語を取り出す。

1. 取り出す単語の品詞は名詞・形容詞・動詞・接頭詞である。
2. 名詞が連続する場合は一つの単語とする。
3. 接頭詞のみでは登録されない。

出現単語毎に出現回数の統計をとり回数の多い上位 2500 件を関連語の候補とする。

3.3　ジャンル毎の出現頻度による判定

本研究では単語の検索語との関連性の強さを求めるためにショッピングサイトでの単語の検索結果を利用する。検索結果 web ページにはサイト独自のジャンル名と、ジャンルそれぞれに対するヒット商品数及び全ヒット商品数が含まれる。図 1 にその情報をグラフで表した例を示す。

検索語と関連語候補とのグラフを比較し、差が少ないものほど検索語との関連性が強いとする。具体的には以下の式でその関連性を計算する。

$$D(P,Q) = \sum_i P(x_i)|P(x_i) - Q(x_i)| \qquad (1)$$

[1]http://mecab.sourceforge.net/

ここで x_i は検索結果ページの各ジャンルに割り当てた番号、P と Q はそれぞれのジャンルにおけるヒット商品数が検索結果の全商品数の何%かを表す分布である。P は検索語の分布であり、Q は比較する関連語候補の分布である。

このとき、$D(P,Q)$ が 1 以上の単語を関連語候補から除外する。

3.4　検索語との共起頻度による判定

関連語は検索語と関連性があり、かつ商品検索に寄与する単語でなければならない。そのため本研究では、関連語候補を検索語と同時に検索に用いた場合のヒット商品数を調べ、その数が検索語によるヒット商品数の 0.2 %以上のものを関連語とした。この判定では検索語と関連語が同時に出現した頻度を数える。これによりさらに関連性を調べ、また検索に役立つ単語かどうか判別する。

4　評価

我々は第 3 節の手法に基づき関連語の抽出を行い、その結果を適合率・再現率・F 値を用いて精度評価する。そこで、提案手法の実行結果と、本研究でその精度を評価する際の基準例を以下に示す。

4.1　実行結果

第 3 節の手法により楽天市場ショッピングサイトで抽出された関連語の例を表 1 に示す。表 1 に示したのは関連語と判断したものを上位から順に取り出したもの

である。本稿において対象となる検索語は便宜上 1 単語と設定する。表 1 を見るとブランド名「GUCCI」に対してはその商品であるバッグ、サングラス] など具体的な商品の種類が取り出すことができ、商品の種類の一つである「時計」や「ギター」に対しては { 腕、電波 }{ エレキ、ソロ } など検索語の前後に付属する単語でより詳細な種類が取り出すことができる事がわかる。

表 1　抽出された関連語

検索語	関連語
時計	腕/電波/機械式/収納ケース/懐中/レディース腕/自動巻き腕/作り出す/ベルト/屋半差得紋/レディースブランド/ケース/機械式腕/高級腕
ギター	エレキ/アコースティック/エレクトリックアコースティック/弦/ベース/クラシック/中古/サウンド/曲/中古エレキ/ソロ/入門セット/
ヘッドホン	出力/端子/インナーイヤー/出力端子/カナルタイプ/密閉型インナーイヤー/ジャック/ポータブル/延長/延長コード/audio/ノイズキャンセリング/audio-technica/
GUCCI	グッチ/ブランド/グッチネクタイ/付属品/トートバッグ/ショルダー/バッグ/サングラス/ショルダーバッグ/ssima

4.2　評価ガイドライン

本研究で抽出された関連語を評価するにはその定義を明確にしなければならない。そのため今回は関連語を純粋に検索語と関連のあるものに限定し、関連商品などの単語はノイズであるとした。そのため、同じ単語であっても検索語によって変わる場合も存在する。検索語に対する関連語を抽出する際シード (適当) やノイズ (不適当) となる単語の例を表 2 に示す。

表 2　評価基準例

検索語	シード	ノイズ
時計	腕/目覚まし/電波/懐中/デジタル/CASIO/	収納ケース/バンド/付属品/大人気/メーカー/
ギター	エレキ/ベース/中古/入門セット/	アンプ/ピック/エフェクター/
ヘッドホン	インナーイヤー/密閉型/ポータブル/ソニー/	端子/延長コード/インピーダンス/
GUCCI	ネクタイ/サングラス/財布/時計/	グッチ/カラー/ロゴ/人気/

4.3　精度評価

第 4.2 節のガイドラインに基づき人手で抽出した関連語を正解データとして、本システムで抽出された関連語を比較することにより求めた適合率・再現率・F 値を表 3 に示す。なお正解データは楽天市場ショッピングサイトでの検索語の検索結果から、ランダムな 50 商品の商品タイトルと商品説明文から人手で抽出したデータである。

表 3　精度評価

関連語	適合率	再現率	F 値
時計	0.79	0.48	0.59
ギター	0.42	0.29	0.34
ヘッドホン	0.56	0.67	0.61
GUCCI	0.63	0.61	0.62

5　考察

5.1　実行結果に関する考察

表 1 の結果には検索に使用できる関連語が 4.1 節で示したように検索語に関連する単語が人目で見ても多く含まれていることがわかる。また、「ギター」に対するエレキ、アコースティック、「GUCCI」に対するサン

グラス、バッグなど表 2 のシードのような検索語に関連し、その結果を絞りこむのに役に立つであろう単語が含まれている。しかし、表 3 からもわかるように人手で抽出したものと比較すると、他にも関連語として適当な単語は多く存在していることがわかる。本節ではその原因について考察したものをまとめる。表 1 で抽出された単語には、それに関係する商品や関連性はあるが、検索には役に立ちにくい単語などが多く含まれている。また、正解データと比較したとき抽出できなかった単語の特徴を以下に示す。

- メーカー名

  表 3 の検索語がギターや時計などの場合検索語に役立つ単語にメーカー名があげられる。メーカーの数は非常に多く有名なメーカー以外は抽出されにくかった。

- 関連性が非常に高い語

  検索語との関連性は高いものの、検索にはあまり役に立たない語は関連語として取り出されやすい。例としてギターに対する「曲」「音楽」「エフェクター」などがあげられる。

5.2 問題点

本節では、特に精度が低かった例や求める結果と大きく異なった例についてその原因を考察したものをまとめる。

- 検索語が検索結果ページのカテゴリ名と似ている場合

  この場合第 3.3 節による判定で検索語のグラフの形は「送料」などのサイト内で頻繁に使用される単語と同じになる。そのためノイズが抽出されやすくなってしまう。

- 検索語がすでに特定の商品名を示す場合

  検索語との関連性は高いものの、ほとんどが商品名の型番や検索語を言い換えた表現などの検索には役に立たない語になってしまう。

- 複数の品詞の単語を抽出したい場合

  小説や映画のタイトルなどは分割されたしまうため、それ単体ではあまり意味のある語にはならず、関連語として抽出されない。そのため関連語の全体数が少なくなってしまう。

6 まとめと今後の課題

本稿では商品説明文から検索語に対する関連語を抽出する手法を提案した。まず、商品説明文からその候補となる単語を抽出し、その候補をジャンル毎の出現頻度と検索語との共起回数の情報を用いて評価し適当な関連語を絞り込んでいった。

その結果、本手法により関連語として適当な語が抽出できることがわかった。抽出された関連語には検索語と頻繁に共起する単語や有名なメーカー、検索結果を分類できる単語があり十分に検索に役立つものであった。しかし、総合的にみるとその品質は十分とはいえず抽出できなかった単語やノイズが多く、検索システムとしては余り良い精度ではなかった。また、単語によってはほとんどがノイズであり、どのような検索結果 Web ページに対しても関連語を満足のいく精度で抽出することは難しかった。

今後の課題は、抽出できなかった場合やノイズを含んでしまう場合の原因を第 5 節のように突き止め対処し、判定条件の改良や追加をして関連語の抽出精度を上げたいと考えている。また、他のショッピングサイトにも対応させて検証を重ねたり、関連性のみが高い単語などを今回扱わなかった中立フレーズに分類し、その有効利用方法を検討するなど将来この手法をより検索に役立てるシステムへと成長させることを検討していきたい。

参考文献

[1] 河野洋史, 柴田知秀, 黒橋禎夫. ブログ記事の商品カテゴリへの自動マッピング. 言語処理学会第 14 年次大会,pp.733-736,2008

[2] 前澤敏之, 山下達雄, 萩原由岐恵. ”商品カテゴリ ”および ”取扱店舗 ”の統計情報を用いた商品タイトルに含まれるフレーズの重要度判定. 言語処理学会第 14 年次大会,pp.1081-1084,2008