商品カテゴリの階層構造を用いた商品分類

中島 道幸、古宮 嘉那子（茨城大学工学部情報工学科）

Product Classification Using Hierarchical Structure of Categories

Michiyuki Nakajima (Department of Computer and Information Sciences,Ibaraki University)
Kanako Komiya (Department of Computer and Information Sciences,Ibaraki University)

要旨

商品のレビュー文書から競合商品を同定する研究や商品ページの属性や属性値を用いた同一商品のクラスタリング手法の研究等、近年、同一商品の同定に関する様々な研究が行われてきている。本稿では、同一商品の同定に関する研究の足掛かりとして商品カテゴリの階層構造を用いた商品分類を行った結果を報告する。実験には、約 60 万件の楽天市場の商品データを使用した。分類器 svm を使用し、五分割交差検定でそれぞれの階層毎のカテゴリの正解率を求めた。消費者が分類することが目的なので、素性を作成する際には、商品ページから消費者が得られる情報のみを選択した。また、求めた正解率から階層毎、階層全体の重みつき平均を求め、ベースラインとの比較を行った。

1. はじめに

近年、Web 上のサービスを利用して商品を購入する“インターネットショッピング”が普及してきた。ショッピングサイトには様々な企業が出店するサイバーモールのようなタイプのものがある。このようなサイトの商品ページは出店している企業が独自に作成している場合がある。そのため、消費者は自分の求める商品を探すことが困難となっている。商品のタイトルや説明文、写真など商品ページのすべてが店舗にゆだねられている。店舗側は売り上げを上げるために商品タイトルの一部に「送料無料」や「ポイント 2 倍」などの修飾語や関連情報を付けている。このため、消費者は単純にクエリ検索を行うだけでは、望んでいる商品のページにたどり着くことができない。さらに、同一商品であるが、商品タイトルや商品説明文が異なっているものや、異なる商品であるが、用いられている商品画像が同一のものが存在する。このような現状から同一商品の同定をする手法が必要であると考え、ショッピングサイトの商品カテゴリに着目した。商品カテゴリに階層があることを利用して、階層的に分類を行った。本稿では、階層を利用していない場合との比較を行う。

2. 関連研究

カテゴリに関しての研究としては、Web 上の商品情報を利用した商品ページのカテゴリ分類という研究を佐藤らが行っていた(佐藤ら(2010))。彼らは商品ページを自動的にカテゴリ分類する手法を提案している。また（古宮ら(2013)）は既存の手法である Naïve Bayes と Complement Naïve Bayes と提案手法である Negation Naïve Bayes を比較している。分類精度が平均 67.3%とベースラインを上回る結果となり、提案手法が商品ページに対して有効であることがわかった。

分類に関する研究としては、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という研究を豊橋技術科学大学の坂地らが行っていた(坂地ら(2010))。商品ページから属性・属性値を抽出し、属性のまとめ上げを行う。また、二つの商品ページを比較し、類似度スコアをつけることで、商品ページのクラスタリングを行う。

本研究では、カテゴリの階層構造を用いて、商品の分類を行っていく点で、これらの研究とは異なる。

3. 階層構造

商品には、膨大な数の商品の中から消費者の求める商品を探せるように、それぞれジャンルが付けられている。この商品ジャンルは大まかなカテゴリから細かなカテゴリまで分けられている。大まかなカテゴリの例として、インテリアを挙げてみる。インテリアには、時計やテーブル、カーテン、椅子等がある。また、テーブルと一口に言っても、ダイニングテーブル、カウンターテーブル、コーヒーテーブル等に細かい分類をすることができる。図 1 に例を示す。このように、商品ジャンルは大きいカテゴリから小さいカテゴリへと、階層構造で構成されている。消費者が欲しい商品が見つからない場合やお買い得な商品を探したいときに、大きいカテゴリから小さいカテゴリへとジャンルで絞り込んでいくことができる。

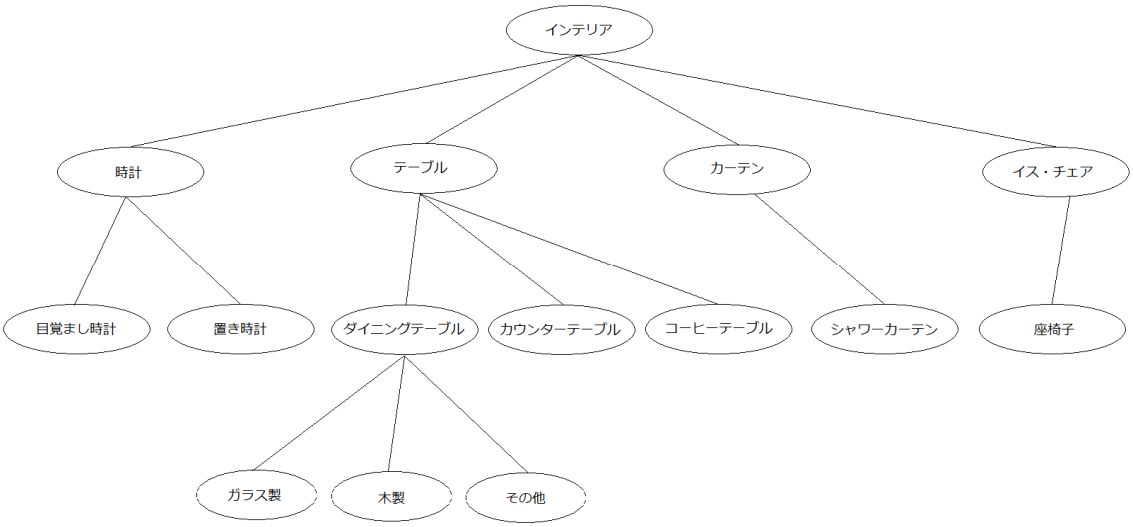

図 1：階層構造の例

本研究では、この階層構造を用いて、商品のカテゴリを機械学習による手法で絞り込んでいく手法をとる。

4. 実験データ

4.1. 実験に使用したデータ

本研究では、約 60 万件の楽天市場の商品データを使用した。商品データは 2014 年 4 月 1 日公開のものである。楽天市場の商品データは 11 個の情報で構成されている。その要素を表 1 に示す。基本的には表 1 のようなフォーマットで商品データは構成されている。実際の商品データの例を図 2 に示す。

商品コードは「店舗コード：商品 ID」と示される。販売方法別説明文とは商品説明文に入らない場合に使用される説明文である。空白となる場合もある。商品 URL はユニーク部分のみが示されている。「http://item.rakuten.co.jp/[店舗コード]/[商品 URL]/」で商品ページの URL となる。ジャンル ID は、その商品カテゴリに割り当てられた番号である。

新約「巨人の星」花形 1› guruguru2:11452005› 400›村上 よしゆき 画梶原 一騎 他原作週刊少年マガジンKC本[コミック]詳しい納期他、ご注文時は ご利用案内・返品のページをご確認ください出版社名講談社出版年月2006年11月サイズISBNコード9784063637533コミック ≫ 少年（中高生・一般） [講談社週刊マガジンKC] › 商品説明新約「巨人の星」花形 1シンヤク キヨジン ノ ホシ ハナガタ 1 シユウカン シヨウネン マガジン コミックス KC ケ-シ- 42265-53※ページ内の情報は告知なく変更になることがあります。あらかじめご了承ください登録日2013/04/08› 9784063637533› http://image.rakuten.co.jp/guruguru2/cabinet/b/7/533/9784063637533.jpg› 0› 0.00› guruguru2› 101941↓

図 2：実際の商品データの例

表 1：商品データフォーマット

順番	データ内容
1	商品名
2	商品コード
3	商品価格
4	商品説明文
5	販売方法別説明文
6	商品 URL
7	商品画像 URL
8	レビュー件数
9	レビュー平均
10	店舗コード
11	ジャンル ID

4.2. ジャンル ID

ジャンル ID は商品ジャンルに割り当てられた番号である。その商品ジャンルに当てはまる商品には、その商品ジャンルの番号であるジャンル ID がつけられる。また、その商品ジャンルには親ジャンル ID というものが割り当てられており、階層構造となっている。つまり、親ジャンル ID を辿っていくと、1 階層にある 34 種類のジャンルに辿り着く。この 34 種類のジャンルは、楽天市場のトップページから検索できる最上層のカテゴリである。階層構造の例で挙げたダイニングテーブルならば、ジャンル ID が「111346」となり、親ジャンル ID は「215476」となる。図 3 に楽天市場のトップページにあるジャンルの一部を例として示す。

図 3：1 階層のジャンルの例

5. 実験

5.1. 実験内容

次の二つの実験を行った。(1)をベースラインとし、カテゴリの階層構造を用いた実験を(2)として、(1)と(2)の重みつき平均の比較を行う。

(1)60 万件のデータを 50 分割し、svm で五分割交差検定を行う。正解ラベルは、その商品のジャンル ID(最下層)とする。

(2)階層毎に分類する手法。60 万件のデータをまず、第 1 階層カテゴリに分類し、分類されたカテゴリ中の商品をそのカテゴリの下の第 2 階層カテゴリに分類するということを最下層まで繰り返す。正解ラベルはその階層のジャンル ID とする。そして、階層毎に五分割交差検定で正解率を求めた。重みつき平均は階層毎に求め、それらを掛けることで階層全体の重みつき平均とする。

5.2. 実験設定

(1)において、60 万件のデータを 50 分割にしたのは PC のスペックの都合である。メモリが 8MB のマシンで動く最低限の分割数が 50 分割であった。

正解率を求める際は、svm のツールとして libsvm を使用する。Option に関してはカーネルのタイプを linear(線形)で行った。これは以前、カーネルタイプの比較を行った実験の結果から、本実験では線形カーネルが適切であると判断した。

(2)において、分類されたカテゴリ中の商品をそのカテゴリの下の階層に分類するとあるが、商品によっては最下層のカテゴリではなく第 2 階層から第 4 階層のカテゴリが正解のものある。そのため、2 階層まではすべてのデータが用いられるが、3、4、5 階層となっていくにつれてデータ数は減っていくということである。

素性として扱う情報については 5.1 で前述した中から商品名、商品価格、商品説明文、販売方法別説明文、商品 URL、商品画像 URL、レビュー件数、レビュー平均に絞る。これは、本研究の背景として、一般の消費者が商品分類を行うことを想定しているため、消費者が商品ページから取得できる情報に限定する必要があるからである。商品説明文に関しては、mecab で形態素解析したものを素性として使用する。また、4.1 節で説明した商品データのフォーマットにしたがってない商品データについては、素性データには含めていない。

(2)についての重みつき平均の計算方法を説明する。はじめに、それぞれの商品データの件数と svm から得られた正解率を掛け、正解数を求める。正解数を計算する際に、それぞれの階層まででおわっているものについては、それ以降の正解率を 100%として計算する。例えば、3 階層まででおわっているものについては、4、5 階層では、正解率を 100%にする。本来は最下層である 5 階層まで細かく分類したいわけだが、細かいカテゴリに属さないため、途中でおわっているものについては、それ以降の階層では、100%分類できると仮定する。次に、求めた正解数を階層毎に足し合わせる。そして、正解数の合計を用いた商品データの全件数で割ることで、階層毎の重みつき平均を求めることができる。最後に、すべての階層の重みつき平均を掛け合わせることで、階層構造全体の重みつき平均を求める。

5.3. 実験結果

表 2 に実験結果を示す。括弧内の数値は途中までで階層がおわっているジャンルを 100%で計算せずに、値として加えない場合の結果である。

表 2：実験結果

正解ラベル	重みつき平均
最下層	31.24%
1 階層	85.80%
2 階層	89.96%
3 階層	84.22%(83.48%)
4 階層	85.07%(79.95%)
5 階層	93.20%(75.77%)
階層全体	51.54%

6.　考察

5 章で行った実験の結果を考察する。まず、(2)の実験における階層毎の結果と階層全体の結果がベースラインである(1)の実験における最下層の結果を上回る結果を得られたため、本研究で提案した商品カテゴリの階層構造を用いた商品分類システムは妥当であるといえる。

(1)における実験結果は 3 割程度の結果であった。(1)は最下層のラベルということで、2 階層や 3 階層等、途中でおわるものから 5 階層にまで亘る広いカテゴリで分類したため、あまりポイントが高くならなかったのではないかと考えられる。

一方、階層毎に分類した結果では、すべて 8 割を上回った。5 階層の結果が 9 割を超えているが､途中までで階層がおわっているジャンルを加えない場合の結果は 7 割程度である。これは、途中までで階層がおわっているジャンルを正解率 100%で加えた結果が大きく関係していると考えられる。また、階層が下になるにつれて途中までの階層に当たるデータが増えてくることで、5 階層で用いるデータが減ってくる。そのため、ジャンル毎に正解率を求めている過程から、五分割交差検定での正解率が 0%になるところも増えてくる。このような理由から括弧内の結果が少し低くなっていると考えられる。

階層全体の実験結果は、5 割を超え、ベースラインを超える結果となったが、それぞれの階層のエラーの累積が全体の正解率を押し下げる結果となっている。特に階層が下った際の正解率の低下が全体の正解率の低下の原因と見て取れる。

今後の課題としては、4 階層、5 階層等の下の階層の分類精度の向上である。考えられる方法としては、末端の訓練事例数を増やすことである。今回は 60 万件で実験を行ったが、マシンのスペックがよければ、データ数を増やすことができる。また、商品データを分割する必要もない。

本研究は、商品カテゴリに関しての分類であるので、商品そのものの分類や同定ではない。なので、今後は階層構造を用いて、単一商品の分類や同定をすることを目指したい。

7. まとめ

本稿では、商品カテゴリの階層構造を用いた商品分類を行った結果を報告した。実験では、正解ラベルを階層毎に設定したものと、最下層に設定したもので重み付き平均の比較を行った。結果は提案した階層構造を用いたシステムの方が 20 ポイント高くなった。今後の課題としては、下の階層の分類精度あげることである。そのためには、訓練事例数を増やすこと等でシステムの向上を目指したい。また、将来的にはこのシステムを用いて、同一商品の同定を可能にしたい。

謝　辞

データを提供していただいた、楽天株式会社と国立情報学研究所に御礼申し上げます。また、この研究は、文部科学省科学研究費補助金[若手 B（No：24700138）]の助成により行われました。ここに、謹んで御礼申し上げます。

文　献

坂地泰紀、小林暁雄、関根聡、竹中孝真(2010)「商品ページから属性・属性値抽出と同一商品クラスタリング手法」言語処理学会第 16 回年次大会発表論文集、pp.371-374.(http://www.anlp.jp/proceedings/annual_meeting/2010/pdf_dir/PA1-27.pdf よりダウンロード可能)

佐藤直人、藤本浩司、小谷善行(2010)「ウェブ上の商品情報を利用した商品のカテゴリ分類」人工知能学会代第 87 回知識ベースシステム研究会、pp.7-10.

古宮嘉那子、伊藤裕佑、佐藤直人、小谷善行(2013)「文書分類のための Negation Naive Bayes」自然言語処理 Vol. 20、 No. 2、pp.161-182.
(https://www.jstage.jst.go.jp/article/jnlp/20/2/20_161/_pdf よりダウンロード可能)