カテゴリに強く関連する語の発見と商品データクリーニングへの適用

村上 浩司　　関根 聡

楽天株式会社 楽天技術研究所
{koji.murakami,satoshi.b.sekine}@mail.rakuten.com

1 はじめに

インターネットの発達と消費者の購買感覚の変化に伴い，企業対消費者間の電子商取引（EC, Electric commerce）は今では必要不可欠な購買方法である．消費者のニーズに応えるため，ECサイトでは扱う商品の種類や数は日々増加している．多くのECサイトにとって商品数の増強は不可欠であるが，サイバーモールのように様々な店舗が商品を販売するサイトでは商品管理が行き届かなくなる問題が生じる．商品登録の方法によりサイバーモール型ECサイトは大きく2つに分けられる．

1. ECサイト側が商品カタログ番号を提供，店舗がそれを利用して出品（アマゾン[1]など）
2. ECサイト側は商品カタログ番号を非提供，店舗が自由に商品を販売（Yahoo!オークション[2]，楽天[3]など）

前者の場合は商品の管理が容易であり，ECサイトが提供するカテゴリに不適切な商品が混在することが少ない，ユーザによる商品の検索が高精度で行えるという利点がある．しかしながら，すべての異なる製品に対して重複なく予め唯一のカタログ番号を規定する必要がある．カタログ番号の策定にはJANコード（Japan Article Code）などの第三者機関が発行する商品コードを参照することが多いが，それらのコードのカバレージの問題や，店舗が独自に販売する商品に対するカタログ番号の策定基準など，考慮しなければならない点も少なくない．これに対して後者の場合，ECサイトは商品情報の管理コストを低く抑えることができるが，ユーザの商品検索に対して網羅的な商品情報を提供できないという問題がある．こうした後者のECサイトでは一般的に，取り扱う商品のカテゴリ階層を提供することで個々の店舗が販売する商品を緩やかに管理しているが，あくまでカテゴリへの商品の提供は店舗の判断によるため，カテゴリ中に不適切な商品が頻繁に登録されてしまう．

ECサイトにとって，提供するカテゴリ内への不適切な商品の混入はサービスの品質という面で大きな問題であり，現在のカテゴリ以外に最適カテゴリが存在する商品（以下，ノイズ商品と呼ぶ）を同定し，最適なカテゴリに修正するデータクリーニング課題が重要である．

本稿では楽天市場の約7,000万件の商品データを対象に，精度90%以上でノイズ商品を同定することを目標として，カテゴリに強く依存する語（以下，強制語と呼ぶ）の効果と強制語の自動発見手法について報告する．2節でタスクの定義と関連研究，3節と4節は強制語の性質とその発見手法，および商品データクリーニングへの適用の実験結果と考察を行い，5節でまとめを述べる．

2 商品の最適カテゴリへの再配置課題

商品のカテゴリ変更は，予め候補となるカテゴリ群が与えられるなら商品の現在所属するカテゴリが何であれ，いわゆる商品のカテゴリ分類問題として設定することができる．新聞記事やWebページなどを対象とした文書分類は数多く研究されている．ECサイトなどで扱う商品を対象とした分類も研究されており，言語情報に着目した商品情報の分類には例えば [4, 1] などがある．新聞記事などの文書分類と異なり，対象を商品情報とした分類では以下の問題を考慮する必要がある．

- カテゴリ間の特徴語の重複問題が深刻
- 商品の種類が多岐に渡る場合，出現する語がスパース
- 商品を提供する店舗が利用するの語彙の相違
- 店舗が提供する商品に関する情報が不足

言語情報以外に画像の情報も考慮した商品分類 [3] も行われている．商品分類の際に有用となる，関連語の抽出 [5]，属性や属性値の抽出 [6, 7] なども行われている．こうした情報を用いた商品分類 [2] の研究も行われている．

3 カテゴリに強く関連する語：強制語

3.1 強制語

関連研究 [4, 1] では分類器を用意し，カテゴリ情報をタグ付けしたデータセットを作成し評価を行う．しかしながら楽天市場商品のカテゴリ階層には，37,355の末端カテゴリがある．最終的に全商品の分類を考えると，分類器の学習のために全カテゴリのタグ付き商品データを

[1] http://amazon.co.jp
[2] http://auctions.yahoo.co.jp
[3] http://rakuten.co.jp

一定数以上準備するには人手が掛かる，カテゴリ階層は一定期間ごとに更新されるためその都度データを準備することは難しいという理由から現実的でない．

我々は，どのような語がカテゴリ特徴語として最適であるのかに着目して商品データを分析したところ，商品の最適カテゴリを自動的に決定できる語が商品情報内に記述されている場合があることがわかった．商品タイトル例を示す．

(1) ユニオンツール ：超鋼エンドミル スクエア Φ２．５×刃長６．２５

この文に対して 4.2 節で述べる形態素解析を行い，日本語名詞連続に着目すると “ユニオンツール/超鋼エンドミル/スクエア/刃長” の 5 種類が得られる．このとき，“エンドミル” は Wikipedia[4]によると “切削加工に用いる工具（切削工具）であるフライスの一種．” とあり，最適カテゴリが楽天カテゴリ階層内での “研削・研磨” であることを一意に決定することができる．我々は唯一つの最適カテゴリを自動的に決定できる，そのカテゴリに強く関連した語を “強制語” として定義する．“エンドミル” を例に挙げると，楽天商品データ全体でタイトルにこの語を含む商品数は 163,233，これらの商品を 70 店舗が提供し，318 カテゴリに分散して存在している．もしこれらのすべてが “エンドミル” 商品であった場合，正しいカテゴリ “研削・研磨” 以外のカテゴリに属する 145,248 商品がノイズとして同定できる．

3.2 強制語の自動発見手法

商品情報から多くの強制語を抽出して辞書を構築することで，商品データのクリーニングに利用することが可能なる．しかしながら高品質の強制語辞書を構築するには，商品情報から強制語の候補を用意して，それらの語が含まれる商品に対し最適カテゴリを人手で付与する必要があり，これには多くのコストを要する．

そこで，強制語辞書の自動構築について検討する．商品情報内に存在する，固有名詞の商品名は強制語となりうるので，固有表現抽出（NER）が有効な手段として考えられるが，固有表現の中で最も抽出が難しいのが製品名であり，日々増加する製品名を高い精度で捉え，辞書を構築することは実現的ではない．また “エンドミル” のようなモノの種類や総称なども強制語となり得ることから，NER タスクだけでは不十分である．

我々はデータを分析した結果，少数の大規模店舗が大量の商品を不適切なカテゴリに割り当てていることが分かった．そこで商品の数ではなく，商品情報中にその語を利用する店舗数（以下，MF(w_i):Merchant Frequency）とカテゴリ内 MF（以下，CMF(w_i,C_n):Category-Merchant Frequency）に着目した．我々は（1）少数の店舗がカテゴリ内でノイズ商品を提供している（2）あるカテゴリ内で多くの店舗が提供する商品に含まれる語はそのカテゴリと強く関連する，という仮説の元，強制語の自動獲得を試みる．

あるカテゴリ内で語 w_i の CMF が高いならば，多くの店舗がそのカテゴリに類似商品を提供していることがわかる．しかし，同じ語を異なるカテゴリの商品説明で用いることもあるため，ある語 w_i を含む商品を提供する全店舗数 MF に対し，カテゴリ C_n で語 w_i を含む商品を提供する店舗数 CMF の集約度 $d(w_i, C_n) = CMF(w_i, C_n)/MF(w_i)$ を計算する．これにより，ある語 w_i に対して唯一の高集約度となるカテゴリ C_n は，語 w_i と強く関連すると考えられることから語 w_i を強制語として獲得する．反対に低集約度の場合はカテゴリ C_n において語 w_i は，関連性が小さくノイズである可能性が高いと考えられる．

また，複数カテゴリで店舗の集約度が高い場合がある．例えば “ドライバー” を強制語候補と考えた場合，語の曖昧性により “ゴルフドライバー” と “スクリュードライバー” の 2 つのカテゴリで高集約度となると考えられる．このような場合は，候補に対して唯一のカテゴリを決定できないため強制語として獲得しない．

3.3 商品情報の重複

ある語にとって低集約度のカテゴリでも，その語を含む商品がノイズ商品であるとは限らない．例えば強制語が “天狗舞” の場合，高集約度カテゴリは “日本酒” となるが，“天狗舞プリント T シャツ” が低集約度カテゴリ “ファッション” に属している場合，強制語に関係なくこの商品は “ファッション” カテゴリが最適である．

もし低集約度のカテゴリ内の商品がノイズ商品で，高集約度のカテゴリに属するべき商品であるなら，その商品の語と高集約度カテゴリの語は類似していると思われる．そこで次式のように低集約度のカテゴリ内の商品 p_j の語集合 $W(p_j)$ と高集約度カテゴリの語集合 $W(c_n)$ の重複を Simpson 係数により計算し，閾値以上ならば高集約度カテゴリに属すべきノイズ商品，そうでなければ低集約度カテゴリに属すべき非ノイズ商品と判断する．

$$\frac{|W(p_j) \cap W(c_n)|}{|W(p_j)|} \rightarrow \begin{cases} Noise(> Threshold) \\ No\_Noise(Otherwise) \end{cases} \quad (1)$$

4 評価実験

獲得した強制語によるノイズ商品同定の性能を実験により検証する．カテゴリ特徴語から構成されるカテゴリベクトルを用いた手法と比較実験を行い，考察を行う．

4.1 カテゴリベクトルを用いる手法

強制語を用いる手法に対して，カテゴリベクトルを用いる手法をベースライン手法とする．我々は，各カテゴ

[4]http://ja.wikipedia.org

リ毎に n 次の特徴語ベクトルを予め用意し，分類対象の商品の語ベクトルと比較して類似度が最も高いカテゴリを商品の最適カテゴリと決定する手法を用いる．カテゴリの特徴語は TFIDF の考え方を拡張し，出現する語の頻度情報と語の出現するカテゴリの情報を用いる．TF に対応する RDF(Relative Document Frequency)，および IDF に対応する ICF(Inverse Category Frequency) との内積とする．ある n 番目のカテゴリ C_n に属する商品情報内に出現する語 w_i を含む商品の数を $df(w_i, C_n)$ とし，語 w_i が出現するカテゴリの頻度を $cf(w_i)$ で表すと，カテゴリ内の語頻度 RDF とカテゴリ頻度の逆数 ICF は以下のように定義できる．

$$RDFIDF(w_i, C_n) = RDF \cdot IDF \quad (2)$$

$$RDF(w_i, C_n) = \frac{df(w_i, C_n)}{D_{C_n}} \quad (3)$$

$$ICF(w_i, C_n) = log\frac{N}{cf(w_i)} \quad (4)$$

このとき D_{C_n} はカテゴリ C_n に属する商品数，N はカテゴリ総数を示す．この RDFICF 値の高い語を，カテゴリ特徴語とする．

4.2 実験設定

ここでは，ベースライン手法，低集約度のカテゴリ内の商品と高集約度のカテゴリ内の語の重複を考慮しない強制語による手法および重複を考慮する強制語による手法の 3 種類の比較実験を行う．

実験データは楽天市場商品情報（カテゴリ数 37,355，商品数 70,196,962）を利用した．名詞抽出には形態素解析器 MeCab[5]を IPA 辞書準拠で利用した．解析結果で名詞が連続する場合，平仮名，片仮名，漢字のみを含む名詞連続を利用する．

ベースライン手法でのカテゴリベクトル生成には，全商品のタイトルと商品説明から得られる単名詞と名詞連続を用いた．RDFICF 値の上位 400 個をカテゴリ特徴語とした．

強制語の抽出は以下の手順で行った．商品タイトルのみを対象とした．まず MeCab による形態素解析と，連続する名詞のまとめあげにより 20,907,697 の名詞連続が得られた．これらに対して，（1）店舗の集約度（CMF/MF）の閾値は 0.8（2）同様に単語頻度の集約度も（CTF/TF）も考慮し閾値は 0.7 （3）重複判定（式 1）の閾値は 0.8（4）接尾辞を含まない（5）店舗数（MF）が 10 以上（6）商品カテゴリ階層内の “その他” のカテゴリのデータは対象外（7）数値，アルファベットのみから構成される語は対象外，の 7 つの条件で候補をフィルタリングした．その結果，名詞連続全体の 0.6%となる 126,271 の強制語が得られた．強制語の高集約度カテゴリ数は 8,469，強制語がタイトルに含まれる商品数は 13,240,117 であった．

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

表 1: ベースライン手法のノイズ・非ノイズ商品同定結果

		システム出力 ノイズ	非ノイズ	合計
データ	ノイズ	141	27	168
	非ノイズ	7	4	11
合計		148	31	179

表 2: 強制語候補のノイズ・非ノイズ商品同定結果

		システム出力 ノイズ	非ノイズ	合計
データ	ノイズ	168/**162**	0/**6**	168/**168**
	非ノイズ	11/**5**	0/**6**	11/**11**
合計		179/**167**	0/**12**	179

すべての強制語の評価はできないため 32 語をランダムに抽出し，それらをタイトルに含む商品に対し現在のカテゴリ内でノイズ商品であるかのタグ付けを行い評価セットとした．このタグ付けは，商品カテゴリ階層の第 1 階層に対してのみ行った．楽天市場のカテゴリでは “食品”，“スポーツ・アウトドア”，“家電・AV・カメラ” などに対応する．これは第 1 階層で商品が不適切なカテゴリに属する場合は明らかにノイズ商品であり，修正の優先度が高いためである．32 語をタイトルに含む商品数は 179 であり，のべ 88 カテゴリ（うち，低集約度カテゴリ数 56）である．タグ付けの結果 179 商品のうち 168 商品がノイズ，11 商品が非ノイズであった．

4.3 実験結果

ベースライン手法の結果を表 1 に，強制語によるノイズ商品同定の結果を表 2 にそれぞれ示す．表 2 中では，商品情報の重複を考慮しない場合が左，考慮する場合が右（太字）である．ベースライン手法はノイズ同定の精度が 0.95(=141/148) であったが，全体の正解率は 0.81(=145/179) と強制語による手法に比べて低い．

強制語を用いた手法では，重複の考慮に関係なくどちらも正解率は 0.94(=168/179) である．重複を考慮しない場合，ノイズ商品同定は精度 0.94(=168/179) であるが，非ノイズの商品に対しては何も対応できない．一方，重複の考慮により非ノイズ商品同定の精度は 0.50(=6/12) となり，ノイズ商品同定の精度も 0.97(=162/167) に向上した．これにより低集約度カテゴリに属する各商品と高集約度カテゴリとの間で語の重複を考慮することは，ノイズ商品同定の精度向上に貢献することが分かった．

4.4 考察

手法の違いによる誤り傾向の違いについて分析したところ，主な原因はベースライン手法ではカテゴリ間での特徴語の重複に大きく影響を受けていることであった．例えば “美容・コスメ・香水”，“ダイエット・健康” カテゴリ内の商品はカテゴリ階層内の第一階層であっても，

単語	RDFCIF	TF	CF	CTF	CDF	MF	CMF
サラタッピング	0.788	349	7	160	160	6	1
ドライ	0.439	56150	3238	4285	333	313	57
ネジ	0.423	69581	1394	224	224	1659	1
フラワー	0.381	141284	5989	455	363	9876	64
ステンレス	0.308	515192	4211	246	246	5185	1
コーススレッド	0.248	1393	19	57	57	38	1
資材	0.208	30289	338	275	114	77	3
アレンジ	0.114	34830	2086	69	69	2775	6

表 3: RDFCIF によるカテゴリ特徴語

比較的類似した商品が取り扱われている．強制語となった “華麗終” という商品は店舗によりどちらにも出品されていた．強制語による手法では語としての “華麗終” の存在に対して最適カテゴリを決めるのに対して，ベクトルベースの手法では他のカテゴリ特徴語の影響により，最適カテゴリを誤選択する場合がある．

このカテゴリ間の特徴語の重複は，語の重複を考慮した強制語による手法の場合にも影響する．表 2 中の語の重複を考慮した結果において，False Negative となった 5 商品に含まれる強制語は全て “焼かつお” であった．高集約度となったカテゴリは “キャットフード”，低集約度カテゴリは “かつお節” である．“かつお節” カテゴリ内で商品は非ノイズであったが，タイトル中の語がキャットフードカテゴリ中の商品のタイトル内の語と重複が多かったことが原因である．

また他の大きな要因は，データ中に元々含まれるノイズ商品の情報がカテゴリベクトル作成に大きな影響を与えていることであった．ベクトルベースの手法は,「あるカテゴリに登録されている商品は特定の言語表現が多く出現する」という仮説に基づいている．しかしながら実データにはどれくらいのノイズが予め存在するのかはわからない．そのためカテゴリ内にノイズ商品が多い場合は，表 3 の例のように適切な特徴語を選択できていないカテゴリが多くなる．これらの語は “ドライフラワー” カテゴリで得られたものである．表中の TF は語の頻度，CF はその語を含む商品が属するカテゴリ数，CTF および CDF はそれぞれカテゴリ内の TF，および商品数を表す．明らかにドライフラワーとは関連性が低い語であるネジ，ステンレスなどが高いスコアとなっている．RDFICF 値は基本的にカテゴリ内のノイズ商品数は考慮しないため，カテゴリ内でノイズ商品数が極端に増加した場合にスコアが不適切に計算される．

これに対して強制語による手法では強制語候補を獲得する際に表中の MF や CMF に着目し，多くの店舗の商品の説明に使われる語を強制語選別の基準とすることで，カテゴリと関連性の低い語を除外し，“フラワー” や “アレンジ” といったカテゴリとより高い関連を持つ語を選別できることがわかる．

強制語による手法は，商品を提供する店舗の数 (MF) を利用することで高精度のノイズ商品同定ができることが分かったが，今回獲得した強制語でカバーする商品数は市場全体の 20%の商品に満たない．カバーできなかった商品には MF の低い語が多く含まれている．こうした商品に対しては 1 語でカテゴリが決定できる強制語ではなく，低頻度，低 MF であってもカテゴリに関連した語であれば，複数の語の共起を利用してパターンを作成して利用することでノイズ商品を同定できると考えられる．

5 おわりに

本論文では大量の商品データから高精度でノイズ商品を同定するため，適切なカテゴリが一意に決まる強制語の自動獲得手法の提案し，小規模なテストセットに対する評価実験を行い，約 97%の精度でノイズ商品を同定できることを示した．

評価実験はカテゴリ階層中の第 1 階層のみが対象であったことから，末端のカテゴリまでの評価を行い全体の精度を調べる必要がある．また店舗の集約度の閾値を変更して，より多くのノイズ商品を同定できるように改善する必要がある．

謝辞

ベクトルベースのアプローチは，弊社のインターン期間に豊橋技術科学大学の坂地泰紀氏が研究開発した手法である．ノイズ商品同定タスクのベースラインとして利用させて頂いた．心より深謝する．

参考文献

[1] Rakesh Agrawal and Ramakrishnan Srikant. On integrating catalogs. In *Proc. of the 10th WWW international Conference on World Wide Web*, 2001.

[2] Anitha Kannan, Inmar E. Givoni, Rakesh Agrawal, and Ariel Fuxman. Matching unstructured product offers to structured product specifications. In *Proc. of the 9th ACM SIGKDD international Conference on Knowledge Discovery and Data Minning*, 2011.

[3] Anitha Kannan, Partha Pratim Talukdar, Nikhil Rasiwasia, and Qifa Ke. Improving product classification using images. In *Proc. of International Conference on Data Minning*, 2011.

[4] Sunita Sarawagi, Soumen Chakrabarti, and Shantanu Godbole. Cross-training: Learning probabilistic mappings between topics. In *Proc. of the 9th ACM SIGKDD international Conference on Knowledge Discovery and Data Minning*, 2003.

[5] 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. 言語処理学会第 16 回年次大会, pp. 367–370, 2010.

[6] 坂地泰紀, 小林暁雄, 関根聡, 竹中孝真. 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. 言語処理学会第 16 回年次大会, pp. 371–374, 2010.

[7] 宇佐美祐, 萩原正人, 関根聡. 商品説明文に対する属性値タギング. 言語処理学会第 17 回年次大会, 2012.