# ブログ記事の商品カテゴリへの自動マッピング

河野 洋志　柴田 知秀　黒橋 禎夫
京都大学大学院情報学研究科
{kouno, shibata, kuro}@nlp.kuee.kyoto-u.ac.jp

## 1　はじめに

近年、ブログやソーシャルネットワーキングシステムなどの CGM(Consumer Generated Media) が注目を浴び、ネット上の「クチコミ」が消費者の購買行動に大きな影響を与えている。CGM の普及に伴い、消費者の興味・関心に即した広告を提示するコンテンツ連動型広告の市場がますます大きくなってきている。現在運用されているコンテンツ連動型広告としては、Google アドセンス[1]やマイクロアド[2]などがある。これらのシステムではまず広告主がキーワードを設定しておき、システムがコンテンツを解析し、それに基づき最適な広告を掲載している。

本研究では、人手で付与されたキーワードを利用せずに、商品カテゴリの特徴語を自動学習し、ブログ記事を商品カテゴリにマッピングする手法を提案する。商品カテゴリとしては JICFS カテゴリを採用し、ブログとしては携帯ブログサイトを用いる。まず、各商品カテゴリにおける特徴語を大規模 Web テキストから自動学習する。そして、ブログテキストに特有の口語調テキストを頑健に形態素解析し、キーワードを抽出することにより、最適な商品カテゴリにマッピングする。

## 2　カテゴリ特徴語の自動学習

ブログ記事に商品カテゴリを付与するためには、各カテゴリの特徴語を用意する必要がある。例えば、「マスカラ」というカテゴリに対しては「アイライナー」、「アイブロウ」、「ビューラー」、「まつ毛」などといった特徴語が必要となる。各商品カテゴリの特徴語は、カテゴリの多さや新規カテゴリの出現の問題から人手で整備するには大変コストがかかり、自動獲得を行なうことが必須となる。そこで本研究ではカテゴリの特徴語を Web テキストから自動学習する。

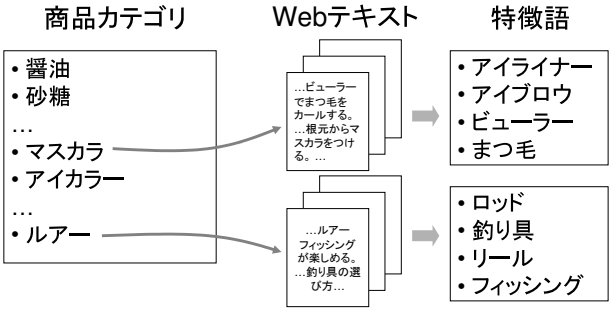


図 1: カテゴリ特徴語の自動学習

表 1: JICFS カテゴリの例

| JICFS カテゴリ (細分類) | 製品例 |
|---|---|
| 蚊取り線香 | 金鳥の渦巻 |
| 蚊取りマット・リキッド | ベープリキッド |
| マスカラ | アイラッシュトニック |
| テニスバッグ・ケース | ヨネックスラケットバッグ |

### 2.1　JICFS カテゴリ

本研究では商品カテゴリとして JICFS カテゴリ[3] を用いる。JICFS とは商品情報を一元的に管理するためのデータベースシステムである。JICFS 分類は大分類、中分類、小分類、細分類の 4 レベルで構成されており、例えば、細分類「醤油」は、(食品) - (加工食品) - (調味料) - (醤油) という分類になっている。本研究では細分類を商品カテゴリとして採用する。合計 2,161 カテゴリあり、表 1 に JICFS カテゴリの例と各カテゴリの製品例を示す。

### 2.2　特徴語の自動学習

本研究では、各カテゴリの特徴語を Web テキストから自動学習する。その概要を図 1 に示す。まず、JICFS カテゴリ名をクエリとして、検索エンジンから Web テキストを取得する。商用の検索エンジン API では十分なテキスト量を得ることができないので、我々が

[1] http://www.google.com/adsense/?hl=ja
[2] http://www.microad.jp/

[3] JAN Item Code File Service, ジクフスと読む。
http://www.dsri.jp/company/jicfsifdb/top.htm

開発している検索エンジン基盤 TSUBAKI[4]の API を用いて、最大 1000 件のテキストを取得する。

取得した Web テキストに対して、形態素解析器 JUMAN[1] で形態素解析を行ない、品詞が名詞 (ただし細分類が時相名詞のもの、ひらがな一文字、カタカナ一文字を除く) または未定義語である語を抽出する。

ここで、抽出した語は JUMAN が出力する代表表記で扱う。これにより、例えば、「喉」、「のど」、「ノド」を同一の代表表記「喉」で扱うことができ、表記の揺れを解消することができる。

次に、各カテゴリにおいて、カテゴリと語の共起度をスコアとして計算し、スコアの高いものを特徴語として採用する。具体的には、カテゴリ $C$ において、カテゴリ $C$ と語 $w$ の自己相互情報量 (Pointwise Mutual Information, PMI) を計算する。$PMI$ は以下の式で計算される。

$$PMI(C,w) = \log \frac{P(C,w)}{P(C) \cdot P(w)} = \log \frac{N \cdot C(C,w)}{C(C) \cdot C(w)} \quad (1)$$

ここで、$P(X)$ は X の生起する確率、$C(X)$ は検索エンジン TSUBAKI での X のヒット件数を表し、$N = 100,000,000$ である。そして、相互情報量が閾値 ($th = 4$) 以上となる語をカテゴリ $C$ の特徴語とする。

ここで、すべてのカテゴリ $C$ において、すべての語 $w$ に対してヒットカウント $C(C,w)$ を得ることは非常に計算コストがかかるので、以下の尺度で上位 $L$ 件の語 $w$ についてのみヒットカウントを取得し、相互情報量の計算を行なう。

$$LDF \cdot IGDF(w) = LDF(w) \cdot \log(\frac{N}{C(w)}) \quad (2)$$

ここで、第一項はカテゴリ $C$ での語 $w$ の文書頻度、第二項は Web 全体での語 $w$ の IDF である。また、現在のところ、$L = 50$ としている。

### 2.3 高頻度語の削除

相互情報量を計算することにより一般的な語は特徴語となりにくいが、一般的な語が特徴語となっている場合があるので、高頻度な語は特徴語とならないようにする。具体的には、ヒットカウントが 2,000,000 以上の語を高頻度語として捨てる (表 2)。

### 2.4 自動学習された特徴語の例

自動学習されたカテゴリ特徴語の例を表 3 に示す。カテゴリ特徴語とみなすための相互情報量の閾値はすべてのカテゴリで同一のため、カテゴリごとに特徴語の数が異なっている。また、表 4 に、逆に語がどのカテゴリの特徴語となっているかの例を示す。これを利用して、次節でブログの商品カテゴリへのマッピングを行なう。

[4]http://tsubaki.ixnlp.nii.ac.jp/index.cgi

表 2: 単語のヒットカウント

| rank | 単語 | ヒットカウント |
|---|---|---|
| 1 | 出来る | 38607944 |
| 2 | 成る | 37088235 |
| 3 | サイト | 36084598 |
| 4 | 言う | 29109087 |
| 5 | 視 | 25870115 |
| ... | ... | ... |
| 1180 | 明 | 2005042 |
| 1181 | 一杯 | 2003305 |
| 1182 | ブルー | 2001276 |
| 1183 | 工場 | 1999643 |
| 1184 | 熊本 | 1999642 |
| 1185 | 即 | 1999404 |
| ... | ... | ... |
| 6114 | モンクレール | 38565 |
| 6115 | 木管 | 38561 |
| 6116 | アイライナー | 38476 |
| 6117 | 遊泳 | 38420 |
| 6118 | 不詳 | 38400 |
| 6119 | ヒスタミン | 38394 |
| ... | ... | ... |

## 3 商品カテゴリへの自動マッピング

### 3.1 アルゴリズム

前節で得られた各カテゴリの特徴語を利用し、ブログを商品カテゴリにマッピングする。以下にアルゴリズムを示す。

1. ブログ記事に対して、JUMAN による形態素解析を行ない、カテゴリ特徴語の学習時と同様に、品詞が名詞 (ただし細分類が時相名詞のもの、ひらがな一文字、カタカナ一文字を除く) または未定義語の語を抽出する。そして抽出した語を代表表記で扱い、語の頻度を計数する。
2. 抽出した語のうち、いずれかのカテゴリで特徴語となっているものをキーワードとする。ただし、以下の 3.2 節で述べる、ひらがなの形態素解析誤りと疑わしい語はキーワードとしない。
3. カテゴリ $C$ のスコア $Score(C)$ を以下の式で計算する。

$$Score(C) = \sum_{w} PMI(C,w) \cdot tf(w) \quad (3)$$

表 3: 自動学習された特徴語

| カテゴリ | 特徴語 |
|---|---|
| 鼻炎用剤 | ヒスタミン:5.687, 鎮痛:5.410, 気管支:5.075, ···, 花粉:4.010 (23 語) |
| 栄養ドリンク | リポビタン:4.953, 滋養:4.874, タウリン:4.803, ···, 栄養:4.016 (13 語) |
| アイライナー | アイライナー:7.882, プードゥルサテン:7.863, フェイスカラパウダー:7.742, ···, まつ毛:4.711 (36 語) |
| マッサージ椅子 | フットマッサージャー:4.968, マッサージチェア:4.824, 指圧:4.256, マッサージ:4.184, リクライニング:4.008 (5 語) |
| ... | ... |

表 4: 語が属するカテゴリ

| 特徴語 | カテゴリ |
|---|---|
| 風邪 | トローチ剤:4.169, 解熱鎮痛剤:4.033 (2 カテゴリ) |
| アイライナー | アイライナー:7.882, アイブロウ:7.366, ···, フェイス用化粧道具:4.162 (10 カテゴリ) |
| 包丁 | 土瓶・鉄瓶:5.133, フライパン類:4.647, たわし・スポンジ:4.567 (3 カテゴリ) |
| サンドイッチ | 食パン:4.286 (1 カテゴリ) |
| ... | ... |

ここで、$tf(w)$ はブログ記事中でのキーワード $w$ の頻度を表す。

4. カテゴリのスコアの高い順にソートし、スコア上位のカテゴリにブログをマッピングする。

### 3.2 形態素解析誤りへの対処

ブログテキストには口語調の表現が多いので、新聞テキストに比べて形態素解析誤りが多く見られる。以下の例のように、特にひらがな表記の形態素解析誤りが目立つ。

(1) 言われてたん だなぁ

この場合、キーワード「だな」がカテゴリ「たな一般」の特徴語となっており、カテゴリ「たな一般」にスコアが与えられてしまう。

ブログ、特に本研究で扱う携帯ブログに特徴的な口語調のテキストを正確に形態素解析を行なうのは大変困難である。そこで、本研究では、形態素解析誤りの典型的なパターンである「あやしい」ひらがな語をキーワードとしないようにする。「あやしい」とは、以下のようなものである。

- 前後いずれかの形態素にひらがな・カタカナ一文字の未定義語がある場合

  (2) 絵をペ そり と。

  「そり」の前の形態素「ペ」が未定義語となるため、「そり」を「あやしい」形態素だとみなし、「そり」をキーワードとしないようにする。

- 前後いずれかの形態素が小さなひらがなである場合

表 5: 精度

| Precision | Recall | F |
|---|---|---|
| 81.0%(64/79) | 80.0%(64/80) | 0.805 |

(3) 言われてたん だなぁ

「だな」の後ろの形態素「ぁ」が小さなひらがなであるため、「だな」をキーワードとしないようにする。

## 4 実験と考察

### 4.1 実験

まず、TSUBAKI API を用いて各 JICFS カテゴリの特徴語を自動学習した。ここで、JICFS カテゴリのうち、「その他」という文字列を含むものを除いた 1,771 カテゴリを用いた[5]。1,771 カテゴリのうち、一つ以上の特徴語が学習されたカテゴリが 1,464 カテゴリ、カテゴリあたりの特徴語の平均数は 9.5 語であった。

次に、ブログの自動マッピングの解析結果を評価した。本研究ではブログとしてロックウェーブ社の携帯ブログサイト aimew(http://aimew.jp/) を用い、50 件のブログ記事に対して最大 3 カテゴリを人手で付与した。そして、システムはスコア上位 3 件のカテゴリを出力し、適合率・再現率・F 値で評価した。結果を表 5 に示す。適合率 81.0%、再現率 80.0%、F 値は 0.805 であった。

[5]カテゴリから「その他」を除いた文字列が別のカテゴリとして存在していることが多いため、「その他」という文字列を含むものを除いた。例えば、「電子ゲーム」と「電子ゲームその他」というカテゴリがあるので、「電子ゲームその他」は除いた。

ノド が痛くてご飯も痛みを我慢して食べてる俺が今カラオケ に来ておりやすアホだよね（笑）自分の体より楽しいこと好きだからさでも…いつも通り声でないからちょっとテンション落ち目悲しいよぉ一応友達連中には歌うまい俺で通ってるのにショック大。

1. カラオケ・歌集・歌謡曲楽譜 (4.726)
2. トローチ剤 (4.554)
3. うがい薬 (4.019)

図 2: 解析例 (下線をひいた語はキーワードを示し、カテゴリの後ろの数字はカテゴリのスコアを示す。)

また、図 2 に解析例を示す。下線をひいた語はキーワードを示す。

### 4.2 考察

誤り例は以下のようにまとめられる。

- 不適当なカテゴリ特徴語

  現在はカテゴリの名前をそのままクエリとして、Web テキストを収集しているが、例えばカテゴリ「スキー防具」のようにカテゴリ名から具体的な商品が連想しにくい場合、収集される Web テキストは高品質でない場合が多い。この問題に対しては、クエリに具体的な製品名を加えて AND 検索を行なうなどして、品質のよい Web テキストを得られるように改良する予定である。

- 形態素解析誤り

  (4) 雷は見てるんは綺麗 やけど 音は嫌い。

  例えば、上記の文でキーワード「やけど」がカテゴリ「アンカ, カイロ」の特徴語となってしまっているので、「あやしい」形態素とみなす条件を加える予定である。

- 固有表現の問題

  「スパイダース」の「ダース」がカテゴリ「ラクロスボール」の特徴語となっている。この問題に対してはブログテキストの固有表現解析を行い、固有表現内の形態素はキーワードとみなさないようにする予定である。

- 多義語の問題

  「…ブランコ に乗る」の「ブランコ」がカテゴリ「遊具」だけでなくカテゴリ「釣用履物」の特徴語となっている (釣りでブランコ仕掛けというものがある)。この問題に対しては、ブログ記事中の多義語の曖昧性を解消することによって対処する予定である。

## 5 関連研究

増沢らは CGM から話題を抽出し、それをもとに関連する広告を配信するシステムを構築している [2]。しかし、このシステムは個別のコンテンツに関連した広告を配信するのではなく、CGM 全般で流行している話題についての広告を配信している。

あるドメインの特徴語を学習する研究として、峠らは意見情報を獲得するためにクエリーに関連するドメイン特徴語を Web 掲示板の文書から抽出している [3]。例えば、車のドメインにおいて、評価対象となる語 (ハンドル、アクセル、シートなど) を特徴語として収集している。クエリーと特徴語の関連度は、メインクエリー、メインクエリーの隣接語、ドメイン特徴語の検索エンジンのヒットカウントを用いて計算している。

## 6 おわりに

本研究では、ブログ記事を解析し、商品カテゴリへ自動マッピングする手法を提案した。まず、Web テキストからカテゴリ特徴語を自動学習し、それに基づき、ブログ記事を頑健に形態素解析し、商品カテゴリへマッピングした。今後の課題としては、不適当なカテゴリ特徴語の改善、形態素解析誤りへの対処、固有表現・多義語の問題への対処があげられる。

## 参考文献

[1] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pp. 22–28, 1994.

[2] 増澤晃, 南野謙一, 渡邊慶和. CGM による話題連動型広告配信システムの開発. 情報処理学会研究報告 情報システムと社会環境研究報告, pp. 19–26, 2007.

[3] 峠泰成, 山本和英. 意見情報獲得のためのクエリー関連のドメイン特徴語抽出. 言語処理学会第 12 回年次大会, pp. 85–88, 3 2006.