商品に関する被参照性と類似性に基づく レビュー文書からの競合商品同定手法

石橋 直己† 乾 孝司‡ 山本 幹雄‡
†筑波大学 情報学群 情報科学類
‡筑波大学 システム情報系 情報工学域
{ishibashi@mibel.,inui@,myama@}cs.tsukuba.ac.jp

1 はじめに

消費者が商品の購入を検討する際、その商品と競合関係にある他の商品を比較したいことがよくある。また、商品を生産・販売している企業にとっても競合する他社の商品を発見することは企業の競争力を保つ上で重要である [1]。

本研究では、上記のような背景のもと、Web 上に多量に存在するレビュー文書を活用し、ある商品に対する競合商品を自動同定する手法を提案する。なお、競合商品は、スマートフォン（iPhone と Xperia）や携帯型ゲーム機（PSP とニンテンドー DS）など、同じ商品カテゴリに属する商品間の関係として捉えられる事が多いが、例えば電子書籍を読むツールとして電子書籍リーダーとタブレット端末が競合することもあり、その定義はさまざまである。本研究では「消費者が商品を購入する際によく比較検討する同カテゴリの商品」を競合商品とみなす。

本研究では、競合商品は以下の特徴を持つという仮説を立て、同定手法の検討を進めた。

> 対称性：ある商品 Y が商品 X に対する競合商品とみなせる場合、商品 X は商品 Y に対する競合商品となりやすい。
>
> 類似性：商品 X と商品 Y が競合関係にある場合、それら商品は同様の特徴を有しやすい。

レビュー文書の処理を前提とした場合、上記の対称性仮説が成り立つなら、商品 X のレビュー文書 R_X とその競合商品 Y のレビュー文書 R_Y の間では、互いの商品名が相互参照されやすいと言えるだろう。また、類似性仮説が成り立つなら、R_X と R_Y には同様の記述がされやすく、R_X と R_Y 間の類似度が高くなりやすいと言える。本研究では、これらの特徴を考慮した手法を提案し、評価実験を通してその有効性を検証する。

2 対称性に基づく手法

まず、文書の特徴を表す一般的指標として Term Frequency(TF) があるように、直感的には、商品 X のレビュー文書 R_X に現れる商品名の回数（参照回数）を求めることで、商品 X に対する各商品の競合商品らしさを見積もることができると考えられる。この考えに基づく競合商品同定のアルゴリズム TF を Algorithm 1 に示す。

Algorithm 1 TF

INPUT　X：対象商品,
　　$\mathcal{R}_X$：X のレビュー文書集合,
　　$\mathcal{Y} = \{Y_i | Y_i \neq X\}$：競合候補商品集合,
　　$\mathcal{D}_{Y_i}$：Y_i の異表記集合
OUTPUT　$\mathcal{X}_{rank}$：競合商品のランキング

```
for Y_i ∈ 𝒴 do
    tf[Y_i] = 0
    for y_ij ∈ 𝒟_{Y_i} do
        tf[Y_i] += count(ℛ_X, y_ij)
    end for
end for
𝒳_rank = 𝒴 の要素を tf[Y_i] の降順に整列
```

レビュー文書は一般に多数のユーザから投稿される。また商品の正式名称には比較的長い名称も多いため、略称表記がされやすい。これらを考慮したカウントを実施するため Algorithm 1 では事前に用意された異表記集合を用いている。また、R_X 中に X 自身が記述されることがあるが、X 自身は競合候補の商品集合 $\mathcal{Y}$ に含めないようにする。

	X_1	X_2	X_3
X_1	—	3	5
X_2	1	—	2
X_3	6	1	—

図 1: 商品名の出現カウント結果

	X_1	X_2	X_3
X_1	—	$3+1=4$	$5+6=11$
X_2	$1+3=4$	—	$2+1=3$
X_3	$6+5=11$	$1+2=3$	—

図 2: 商品名の出現カウント結果（補正後）

上記アルゴリズムは参照回数のみを利用している。そこで、これに被参照回数の情報を反映させることを考える。考慮したい全ての商品について、Algorithm 1 の 6 行目までを実行して出現回数をまとめると、図 1 のような表ができる。表の i 行 j 列要素は商品 X_i のレビュー文書に現れた商品 X_j の名称表記の出現回数を表す。また、j 行 i 列要素は商品 X_j のレビュー文書に現れた商品 X_i の名称表記の出現回数を表す。ここで、表の ij 要素に ji 要素の値を加算する補正操作をおこない、図 2 のような表を新たに作成する。そして、図 1 の替わりに図 2 のカウント情報に従って、各商品について Algorithm 1 の 7 行目を実行する。この補正操作によって、相互参照関係にある商品ほど強い補正効果が得られ、その結果として、競合商品として同定されやすくなる。

3 類似性に基づく手法

類似性仮説が成り立つなら、ある 2 つの商品の各レビュー文書には、それらに関する属性や使用感といった共通の事柄が多く記述されるはずである。そこで、レビュー文書間の類似度を求め、その類似度によってある商品に対する各商品の競合商品らしさを見積もることを考える。この考えに基づく競合商品同定のアルゴリズム Sim を Algorithm 2 に示す。

アルゴリズム中の関数 $cos\_sim()$ ではコサイン類似度を計算している。各引数に与えられたレビュー文書集合は、その文書集合中に現れる名詞についてその頻度を重みとして生成されたベクトルに変換された後、実際に類似度が計算される。

先の Algorithm 1 は競合商品を求めたい対象商品のレビュー文書のみが必要であるが、Algorithm 2 は対象商品および競合商品候補の両者のレビュー文書が必要である。

Algorithm 2 Sim

INPUT X：対象商品,
$\mathcal{R}_X$：X のレビュー文書集合,
$\mathcal{Y} = \{Y_i | Y_i \neq X\}$：競合候補商品集合,
$\mathcal{R}_{Y_i}$：Y_i のレビュー文書集合

OUTPUT $\mathcal{X}_{rank}$：競合商品のランキング

1: **for** $Y_i \in \mathcal{Y}$ **do**
2: $sim[Y_i] = cos\_sim(\mathcal{R}_X, \mathcal{R}_{Y_i})$
3: **end for**
4: $\mathcal{X}_{rank} = \mathcal{Y}$ の要素を $sim[Y_i]$ の降順に整列

4 統合手法

ここまでに述べた 2 つのアルゴリズムの結果を統合する。具体的には、式 (1) で得られるスコアを競合商品候補ごとに求め、このスコアに従って再ランキングを実施する。式中の $Rank(\cdot, Y_i)$ は各手法で得られた商品 Y_i の順位であり、定数 α（$0 \leq \alpha \leq 1$）は組合せの重みである。

$$comb[Y_i] = \frac{\alpha}{Rank(TF, Y_i)} + \frac{1-\alpha}{Rank(Sim, Y_i)} \quad (1)$$

5 評価実験

5.1 実験条件

代表的な価格比較サービスである価格.com[1]に投稿されたレビュー文書のうち、自動車とデジタルカメラ（デジカメ）のデータを利用した。自動車カテゴリの売れ筋ランキング上位の商品のうち、正解となる競合商品が自動的に取得できる商品を選別し、465 種の自動車に対するレビュー文書を準備した。これらの車種に対し、各商品ページに存在する「ライバル車種比較[2]」欄に挙げられている車種を正解競合商品とみなした。これにより、平均 4.9 件の正解競合商品を得た。同様にデジカメカテゴリでは 263 種のデータを準備した。ただし、デジカメカテゴリには「ライバル車種比較」に相当する欄がないため、代わりに「この商品を見ている人はこんな商品も見ています」欄に提示される商

[1] http://kakaku.com/

[2] 現在、この欄はなくなっており、新たに「この商品を見ている人はこんな商品も見ています」欄が登場している（2012/1/23 に確認）。

品を正解競合商品とみなし、平均 16 件の正解競合商品を得た。なお、上記データは、自動車カテゴリは 2011 年 7 月 23 日〜2011 年 7 月 25 日にかけて取得し、デジカメカテゴリは 2011 年 10 月 20 日〜2011 年 10 月 23 日にかけて取得した。

また、Algorithm 1 において、正確な商品名のカウントを実現するために、

トヨタ ランドクルーザー
→ ランドクルーザー, ランクル

CANON PowerShot SX40 HS
→ SX40, SX40HS

のような正式名称に対する異表記リストを全対象商品に対して人手で作成して用いた。今後、任意の商品を対象として提案手法を適用する場合には異表記リストの自動生成が必要不可欠であり、文書内の商品名の記述を自動認識する手法（例えば文献 [2]）の精度向上が望まれる。

評価尺度には以下に示す 3 つの指標を用いる。これら指標は、下側ほどランキングの順位変動に対し敏感に反応するような指標となっている。

正解率 ランク 1 位として出力された商品が正解競合商品に含まれていれば正解とみなし、全対象商品において正解の割合を求めた値。

Precision@k ランク上位 k 件の出力のうち、それらが正解競合商品に含まれる割合を適合率とし、それを全対象商品で平均した値。本実験では $k = 5$ とし、以降この指標を Precision@5 と記す。

MAP@k 1 位から k 位に対する平均適合率を求め、それを全対象商品で平均した値。本実験では $k = 5$ とし、以降この指標を MAP@5 と記す。Algorithm 1 は、対象商品のレビュー文書において言及されない競合商品候補は実質的に結果から排除される。本実験では、この影響を抑えて評価するために、情報検索分野などで標準的に使用されている MAP 指標とは異なる定義を採用しているので注意されたい。

5.2 実験結果

各手法を単独で適用した場合の結果を表 1 および表 2 に示す。表中の TF が Algorithm 1 の結果、TF（補正）は TF に対して対称性に基づく補正操作を加えた結果、Sim が Algorithm 2 の結果である。また、

表 1: 各手法の結果（自動車）

手法	正解率	Precision@5	MAP@5
TF	0.359	0.186	0.245
TF（補正）	**0.439**	**0.245**	**0.317**
TF（パタン）	0.275	0.108	0.137
TF（パタン, 補正）	0.346	0.156	0.196
Sim	0.314	0.161	0.225

表 2: 各手法の結果（デジタルカメラ）

手法	正解率	Precision@5	MAP@5
TF	0.449	0.270	0.294
TF（補正）	**0.536**	**0.355**	0.392
TF（パタン）	0.354	0.159	0.175
TF（パタン, 補正）	0.464	0.236	0.262
Sim	0.483	0.347	**0.405**

TF（パタン）は、表 3 に示すような比較表現パタンに該当する文脈である場合のみ商品名を数え上げるよう、TF を変更した場合の結果である。これは先行研究 [3,4,5] において、テキストから競合関係にある事物対を特定する手がかりとして比較表現が役立つことが示唆されていることから、本研究でも効果を検証するために比較手法として追加した。

表 3: 比較表現パタンの例 (A：商品名)

A と比べ	A に比べ	A と比較	A より
A には	A にも	A では	A の方

表 1 から、自動車カテゴリでは TF が Sim よりも良好な結果であることがわかる。TF に補正操作を加えると飛躍的な性能の向上が確認でき、各手法を単独で適用した場合では TF（補正）がもっとも良い性能となっていた。TF と TF(補正) の実行結果の例を表 4 に示す。正解商品は太字で表示しており、括弧内の数字はそれぞれのスコアである。TF において比較表現パタンを用いてカウントした場合は逆に性能の低下を招いた。この原因を調べたところ、パタンにマッチしている事例自体は性能向上に寄与していると言えるが、現れている商品名に対するパタンの被覆率が極端に低くなっており、データ全体を通して評価すると、比較表現パタンを用いない場合よりも性能が下がる結果となっていた。

デジカメカテゴリでもほぼ同様の結果であったが、自動車カテゴリに比べて Sim の性能が上がっており、商品カテゴリによって手法の挙動が異なることが確認された。

次に、統合手法の実験結果を表 5 および表 6 に示

表 4: 自動車カテゴリの商品:「日産 スカイライン」に対する実行結果の例

順位	TF	TF(補正)
1 位	**フーガ**(11)	**マーク X**(20)
2 位	プリメーラ (7)	**フーガ**(16)
3 位	**ティアナ**(5)	**ティアナ**(10)
4 位	ダイハツ COO(5)	レクサス IS(9)
5 位	ウィンダム (4)	プリメーラ (8)

表 5: 統合手法の結果（$\alpha = 0.5$，自動車)

手法	正解率	Precision@5	MAP@5
TF	0.381	0.212	0.278
TF（補正）	**0.434**	**0.244**	**0.317**
TF（パタン）	0.325	0.189	0.249
TF（パタン, 補正）	0.383	0.216	0.286

す。組合せの重みは $\alpha = 0.5$ で固定した。自動車カテゴリでは、概ねの設定で性能が向上していることから、ランキング結果が改善されていることがわかるが、どの設定でも TF（補正）の単独適用を上回ることはなかった。一方のデジカメカテゴリでは、各評価指標および各手法において性能の向上が確認できた。今回の結果は、両手法の性能が同程度であれば、両手法は互いに相補的な関係にあり、手法の組合せが競合商品ランキング結果の改善に有効であることを示している。

5.3 組合せ重みの影響

最後に、組合せ重みの影響について調査するために、重み α を変化させながら、性能変化の様子を測定した。なお、どの結果でも同様の傾向を示していたため、ここでは代表してデジカメカテゴリにおける Precision@5 の結果のみ図 3 に示す。ここから、重みの変更に対する性能変化は鈍く、実利用においては適当な重み設定（例えば $\alpha = 0.5$）のもとで手法の組合せを検討すれば十分であると言える。

6 おわりに

評価実験を通して、提案手法がレビュー文書からの競合商品の同定に有効であることを示した。今後の課題として以下のような項目が考えられる。

相互参照の考え方を複数の商品間に拡張することで、PageRank[6] のようなリンク解析の技術を競合商品同定に適用できるかを検討する。

表 6: 統合手法の結果 ($\alpha = 0.5$，デジタルカメラ)

手法	正解率	Precision@5	MAP@5
TF	0.521	0.386	0.445
TF（補正）	**0.586**	**0.417**	**0.483**
TF（パタン）	0.536	0.365	0.434
TF（パタン, 補正）	**0.586**	0.395	0.471

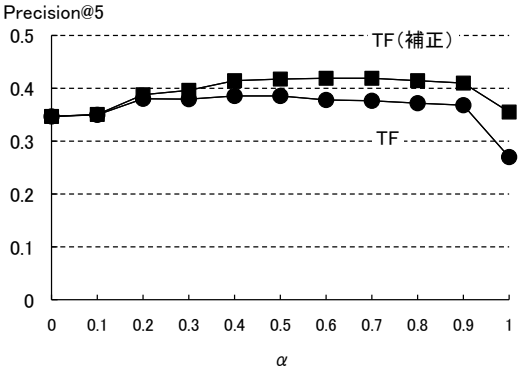

図 3: 重みと性能の関係（デジタルカメラ）

本実験ではレビュー間の類似性を測る際にレビュー中のすべての名詞を用いた。しかし、より正しく類似性を測るには、商品と関連の深い特徴量を適切に抽出する方法を検討すべきである。

参考文献

[1] フィリップ・コトラー. コトラーのマーケティング・コンセプト. 東洋経済新報社, pp.42–44, 2003.

[2] 渡辺尚吾, 乾孝司, 山本幹雄. 商品カテゴリ情報に着目した教師データ収集による商品名抽出手法. 第 25 回人工知能学会全国大会, 2011.

[3] S. Li,C. Lin,Y. Song, and Z. Li. Comparable Entity Mining from Comparative Questions. In *Proc. of the 48th ACL*, pp.650–658, 2010.

[4] 山崎義隆, 乾健太郎, 松本裕治. 競合事物間における比較関係認識. 情報処理学会自然言語処理研究会, pp.1–7, 2011.

[5] 佐藤敏紀, 奥村学. blog からの比較関係抽出. 情報処理学会自然言語処理研究会, pp.7–14, 2007.

[6] L. Page, S. Brin, R. Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Stanford InfoLab*, 1999.