

競合製品や競合するサービスに関する オントロジーの構築とその利用

鈴木健之[†] 丸山広[†] 中村太一[†]
東京工科大学[†]

競合する他社の製品やサービスの名称のオントロジーを構築し、分析対象とする情報にアノテーションを付与することが商品の評判を獲得するために有効な手段であると考えられる。

本研究は Web 上の情報から比較表現を手掛かりに競合する他社の製品やサービスの名称の候補を抽出し、抽出した候補に自社の製品やサービスの仕様表現を持っているかによって競合する他社の製品やサービスの名称を判定し、自動で抽出した。その結果、仕様表現による判定で、抽出した候補に比べ適合率が 20%程度向上した。

1. はじめに

製品やサービス(以下、商品)を開発し提供する企業にとって自社の商品に対する顧客の評判を把握することは、商品の品質向上や問い合わせしてきた顧客への対応、及びマーケティング企画を立案する(以下、商品力向上)上で重要である[1][2]。

商品の評判は、商品名を表す対象表現、商品の機能や仕様、及びデザインを表す仕様表現、商品名や仕様の評価を表す評価表現の 3 要素で構成されていると定義する[3][4][5]。

評判を構成する要素のうち、仕様表現と評価表現の組みや、対象表現と仕様表現の組みのオントロジーを構築し、分析対象の情報に

アノテーションを付与し、分析することが評判の獲得には必要である。

他方、自社の商品と競合する他社の商品の評判も自社の商品の商品力を向上する上で重要な情報となると考えられる。

しかし、競合する他社の商品名は、正式名称だけではなく略称や愛称や、俗称があり、それらを人手で全て網羅することは困難である。

本研究は、特定の商品名に競合する商品名がテキスト中に出現するパターン(以下、共起パターン)を用いて自動で競合する他社の商品名の候補を抽出し、抽出した候補が自社の商品と同様の仕様表現を持っているかによって商品名候補にフィルタをかける。このようにして、競合する他社の商品名のオントロジーを自動で構築することを目指す。

Construction of Competing product and Competing Service with Ontology and The use of Ontology

[†]Kenshi SUZUKI [†]Hiroshi MARUYAMA

[†]Taichi NAKAMURA;

Tokyo University of Technology

2. 関連研究

Web 上の文章から評判を獲得したい商品名

と予め用意した評価表現辞書を用いて評判を獲得する研究がある[6]. ユーザが入力した商品名を元に Web 検索をするため, 特定の商品名に対する評判を獲得することはできる. しかし, 分野に合わせて評価表現辞書を構築する必要があるため, 商品名や評価表現のオントロジーを自動構築する本研究により, より低コストで評判を獲得できると考えられる.

Web 上から収集した情報に共起パターンを用いて仕様表現と評価表現を半自動で抽出する研究がある[7]. 抽出する表現は, 仕様表現と評価表現であり, 対象表現の抽出はしていない. 似たような対象表現と仕様表現の共起パターンを用いて対象表現を抽出することができれば, 同じ仕様を持つ商品名, すなわち競合する商品名を獲得できると考えている.

営業日報やコールセンターの応答履歴などの大量のテキスト情報を活用するために「実際に起きたこと」と「起きなかったことや予定していること」の区別に着目し, 事象の生起に関するアノテーションを付与する研究がある[8]. テキストマイニングのために付与すべきモノに関するアノテーションとして, 顧客, 製品, 自社, 競合, 協力者の5つを挙げており, これらのオントロジー構造を整理することが評判の獲得には必要であるといえる. 本研究は, 競合に関するアノテーションに着目し, 競合のオントロジーとして, 競合する他社の商品名が必要であると考え, このオントロジーを構築することを目指す.

我々は, Web 上の情報から特定の商品に関する対象表現, 仕様表現, 評価表現を自動で抽出し, 評判を獲得する研究に取り組んできた[3][4][5]. 評判を獲得するまでの流れを, 収集, 整理, 抽出, 分析という4ステップで取り組むことを提案し, 評判の獲得を目指している.

本研究は, 評判の獲得に, 競合する他社の商品名が利用できると考え, 従来の研究を基礎に競合する他社の商品名の獲得に取り組む.

3. 評判獲得の枠組み

3.1 Web 上の情報の特徴

近年, インターネットの普及により商品に関する評判が Web 上に発信されるようになった. Web 上の情報には, 企業のコールセンターでは獲得できない自社の商品と競合する他社の商品を比較する情報が存在している. このため, 企業は Web 上の情報を顧客の評判を獲得するための情報源として捉えるようになった.

しかし, 大部分の Web 上の情報には W3C(World Wide Web Consortium)の Tim Berners-Lee が提唱する Semantic Web のようにコンテンツにメタデータが付与されていないため, Web 上の情報は未整理な状態であるといえる.

そのため, 未整理な Web 上の情報から評判を獲得するためには, 仕様や評価に関するオントロジーを構築し, Web 上の情報にアノテーションを付与することが必要となる.

本研究は, Web 上の情報から評判を獲得するためには, どのようなオントロジーを構築し, アノテーションを付与すべきかを検討する.

3.2 構築すべきオントロジー

Web 上の情報から評判を獲得するためには, 対象表現, 仕様表現, 評価表現に関するオントロジーに加え, 競合する他社の商品に関する

るオントロジーを構築することが必要であると提案する。

競合する他社の商品名のオントロジーを利用し、Web 上の情報に競合する他社の商品名のアノテーションを付与することができれば、Web 上の情報の特徴である自社の商品に競合する他社の商品との比較に関する評判を獲得することができるのではないかと考えたからである。

一般的に、他社の商品が自社の商品に競合しているかどうかの判断は企業によって異なるが、企業は自社の商品と同様の機能や仕様を持つ商品を開発し提供する他社を気かけ、常に調査をしていると考えることができる[9]。

本研究は、特定の企業が開発し提供している商品と同じ仕様や機能を持つ他社の商品を競合する他社の商品であると定義し、このオントロジーの構築を目指す。

3.3 構築すべきオントロジーの整理

対象表現、仕様表現、評価表現、競合する他社の商品名以外に評判の獲得に有効であると考えることができるオントロジーについて検討する。

はじめに、評判が肯定であるか、否定であるかというオントロジーを構築することが、従来研究で多く取り組まれてきた。しかし、語の単位で肯定か否定かというアノテーションを付与してしまうと以下のような文単位の評判に対応することができない。

例:“A 社の車の性能は決して良いわけではないかもしれない”

そこで、肯定、否定のオントロジーを語の単

位で構築したとしても、アノテーションを付与する際に付与した情報から 1 文単位で肯定か否定かの意味を判定する必要がある。

次に、時系列情報のオントロジーについて検討する。評判がいつ話題になったのかを知る上で時系列情報は、重要である。しかし、時系列情報といっても、明日や今日といった特定の日だけを表しているや何ヶ月間や数週間など特定の期間を示すものがある。これらを分類しオントロジーを構築することが評判の獲得には必要である。

最後に、程度表現のオントロジーについて検討する。程度に関するオントロジーは、評判の度合いを調べる上で必要となる。その度合いがどの程度であるかを定量的に表す方法は、従来研究で取り組まれている[10]。

このように複数のオントロジーを構築し、アノテーションを分析テキストに付与することが、評判を獲得する上で必要となる。

これらのオントロジーを小規模であっても考えられる限り構築し、全てのオントロジーのアノテーションを付与したときに、評判の獲得にどのような影響が出るのかということ調べてみる必要があるのではないかと考えている。

4. 問題点と課題

4.1 競合する他社の商品名候補の獲得

Web 上に存在する競合する他社の商品名は、正式名称だけではなく略称や愛称や、俗称があり、それらを人手で全て網羅することは困難である。また、Web 上には比較に関する情報だけが存在しているわけではない。そこで、自社の商品名と競合する他社の商品名を比

較する表現を手がかりに、自社の商品名と競合する他社の商品名の共起パターンを利用することで、Web 上の情報から比較に関する情報を収集し、収集した情報から競合する他社の商品名の候補を獲得する。

4.2 共起パターンを利用して獲得した競合する他社の商品名候補の判別

自社の商品名と競合する他社の商品名の共起パターンを利用して獲得した競合する他社の商品名候補には、競合する他社の商品名以外の語が含まれていることが予想される。そこで、競合する他社の商品名の候補が自社の商品と同様の仕様表現を持っているかで競合する他社の商品名かどうかを判別する。

5 解決策

5.1 比較表現を用いたデータの収集

Web 上に存在する情報は日々増え続けており、この膨大な Web 上の情報全てに特定の商品に競合する他社の商品名が含まれているわけではない。

そのため、Web 上の情報全体から競合する他社の商品名を抽出することに比べ、比較に関する情報から競合する他社の商品名を抽出するほうが、競合する他社の商品名を抽出することができる。と考える。

本研究では Yahoo 検索のフレーズ検索に比較表現と特定の商品名を利用することで、Web 上の情報から特定の商品と競合する他社の商品名を比較する情報を収集する。

我々が定義した11の比較表現を表1に示す。また、検索に利用するフレーズの例を以下

の例1から例3に示す。

例1 ウイルスバスターVS

例2 ウイルスバスターと

例3 VS ウイルスバスター

表1 比較表現一覧

VS	と	に比べ
か	や	または
よりは	から	よりも
より	の方が	

5.2 収集した情報の整形と語の抽出

収集した情報から共起パターンを利用して、競合する他社の商品名を抽出するためには、日本語の自然言語処理技術を用いて語を抽出する必要がある。しかし、Web 上から収集したテキストには日本語として不適切な情報が含まれるため、整形が必要となる。具体的には、空欄、顔文字などに使用される括弧を始めとする記号、HTML タグは不要な情報として削除する。

整形したテキストから競合する他社の商品名を抽出するために、まずは言語において意味を持つ最小の単位(形態素)に分割する。この形態素解析にはデータを処理しやすい xml 形式にするために日本語係り受け解析器「南瓜」を用いた。

競合する他社の商品名は、複数の形態素から構成される場合があるため、形態素の結合処理を行う。例えば、名詞と数詞で構成される「ノートンアンチウイルス 2007」がある。また、茶釜はアルファベットを1文字ずつに分割してしまうため、アルファベットの連続も結合する。

定義した結合ルールを以下に示す。

1. 名詞の連続
2. 名詞+数詞
3. アルファベットの連続
4. 未知語+数詞

5.3 共起パターンによる競合する他社の商品名候補の抽出

形態素に分割した Web 上の情報から以下の例1, 例2に示す共起パターンに 5.1 の検索に用いたキーワードを“比較となるキーワード”として代入し, この共起パターンを抽出する. 抽出した共起パターンから競合する他社の商品名の候補を抽出する.

例1: “商品名”+“比較となるキーワード”
+“競合する他社の商品名の候補”

例2: “競合する他社の商品名の候補”+“比較となるキーワード”+“商品名”

なお, 企業は自社の商品名を既に知っているかと仮定し, 自社製品に関するキーワードは候補として抽出しない.

5.4 競合する他社の商品名の判別

共起パターンにより抽出した競合する他社の商品名の候補には, 競合する他社の商品名以外の語が含まれていると考える. そこで, 競合する他社の商品名を判別する方法として自社の商品が持つ仕様表現(以下, 固有表現)と同様の固有表現を持つ候補を, 競合する他社の商品として抽出する.

具体的には Yahoo 検索のフレーズ検索で, “競合する他社の商品名”+の+“固有表現”と

いうフレーズで検索し, 検索が HIT するかしないかで, 自社の商品と同様の固有表現を持っているかを判別する.

6. 実験

6.1 前提条件

自社の商品名を“ウイルスバスター”として 11 の比較表現で 7 月 4 日 13:20 に実験対象となる情報を収集した. 比較表現を商品名の前後に入れて検索したため(例:ウイルスバスター-VS, VSウイルスバスター), 合計 22 回のフレーズ検索をした.

6.2 結果

6.2.1 データの収集結果

Yahoo 検索の検索結果を表示するページを合計 22 ページ収集した. なお検索では, 100 件の結果を 1 ページに表示するオプションを用いた. 22 ページに表示される検索結果から重複する結果を除いたところ検索に HIT した合計の件数は 1525 件であった. この結果を形態素解析したところ, 154866 の形態素が抽出され, そのうち異なり語は, 11343 語であった.

この情報から人手で正解となる競合する他社の商品名を抽出したところ, 171 語の正解を獲得することができた. この正解データのうち, 収集したデータを形態素解析した結果に出現した形態素の数は 75 語であったため, 自然言語処理により獲得できる正解を 75 語とした. 正解の一部を表 2 に記載する.

表 2 正解データ一部

1	カスペルスキー
2	ウイルス警備隊
3	ウイルスワクチン
4	ウイルスブロック
5	ウイルスドクター
6	ウイルスチェイサー
7	ウイルスセキュリティ
8	マカフィー
9	ノートン アンチウイルス
10	ゾヌ

6.2.2 共起パターンを利用した競合する他社の商品名の獲得

商品名をウイルスバスターにし、5.3 の例1、例2の共起パターンに商品名としてウイルスバスターを代入し、共起パターンを抽出した。抽出した共起パターンから、競合する他社の商品の候補を抽出し、候補に“ウイルスバスター”、“トレンドマイクロ”、“TRENDMICRO”、というウイルスバスターを開発、販売しているトレンドマイクロ株式会社の商品に関するキーワードが出現しているものは削除する。この抽出結果をまとめたところ、154 語が抽出された。

6.2.3 固有表現を用いた商品名の抽出結果

ウイルスバスターの固有表現である“パターンファイル”と“ファイアウォール機能”を用いて 6.2.2 で抽出した競合商品名候補をフィルタリングする。

具体的には、“競合商品名候補”+の+“固有表現”というフレーズで Web 検索し、HIT の有無で競合商品名を判別する。

“パターンファイル”と“ファイアウォール機能”の 2 語で検索したため、一方の検索結果は HITしたが、他方では検索に HITしないということがありと推測できた。そのため、“パターンファイル”と“ファイアウォール機能”の検索結果のうち両方に出現した語を競合する他社の商品名とした場合(以下、この条件を AND 検索での結果と呼ぶ)と、“パターンファイル”と“ファイアウォール機能”の検索結果のうち片方だけでも出現した語を競合する他社の商品名であると定義した場合(以下、この条件を OR 検索での結果と呼ぶ)の抽出結果を調べた。

AND 検索では、22 語が抽出された。OR 検索では、53 語が抽出された。

6.3 評価

抽出された結果を再現率と適合率、及び F 値を用いて評価する。

再現率とは、収集した情報(図1の Web page:1)からシステムが抽出した競合する他社の商品名(図1の System:S)を人手で作成した正解データ(図1の正解:H)がどれだけ網羅しているかを表した値である。

適合率とは、収集した情報(図1の Web page:1)からシステムが抽出した競合する他社の商品名(図1の System:S)にどれだけ人手で作成した正解データ(図1の正解:H)が含まれているかを表した値である。

以下に図1の内容から再現率と適合率を求める式を記載する。

$$\text{再現率(Recall ratio)} \quad r = T/H$$

$$\text{適合率(Precision ratio)} \quad p = T/S$$

F 値とは再現率と適合率の調和平均であり、F 値が高いほど性能が良い。この F 値の式を以下に示す。

$$F \text{ 値} = 2 * (r * p) / (r + p)$$

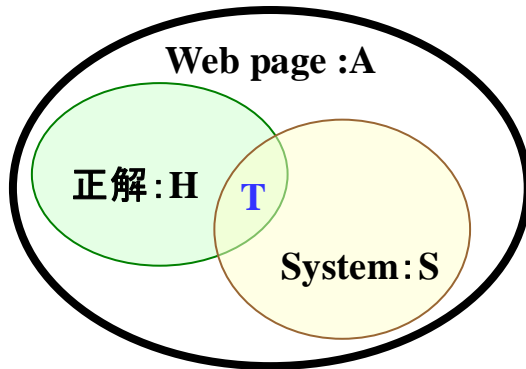


図1 再現率と適合率

6.3.1 共起パターンを利用した競合する他社の商品名の獲得

共起パターンで抽出された語数は 154 語であり、その中に正解データ 75 語のうち 37 語を含んでいた。そのため再現率は、49.33%であり適合率は 24.03%であり、F 値は、0.32 であった。

6.3.2 固有表現を用いた商品名の抽出結果

AND 検索の結果をまとめたところ、検索に 22 語が HIT し、その中に正解データ 75 語のうち 10 語が含まれていた。そのため、再現率は 13.33%、であり適合率は 45.45%であり、F 値は 0.21 であった。他方、OR 検索では 53 語が HIT し、その中に正解データ 75 語のうち 22 語が含まれていた。そのため、再現率は 29.33%であり、適合率は 41.51%であり、F 値は、0.34 であった。

6.4 考察

11 の共起パターン用いて抽出した結果、抽出できた語数は、154 語であり、正解を 37 語含んでいたため、49.33%であり適合率は 24.03%であったが競合する他社の製品名を抽出することができた。

予想通り、競合する他社の製品名以外の情報が抽出されていたため 11 の共起パターン用いて抽出した結果の適合率は、24.03%と低かった。そのため、自社の商品の固有表現によるフレーズ検索で競合する他社の商品名を判定したところ、AND 検索の結果及び OR 検索の結果、共に再現率は低下したが適合率は、上昇した。再現率は、母集団の規模が増える毎に調査することは難しくなるため、適合率を重視することが必要であると考え。そのため、自社の商品の固有表現による競合する他社の商品の判定は、効果があったといえる。

なお F 値が、AND 検索の結果では 0.21、OR 検索の結果では 0.34 であったため、固有表現による判定では OR 検索の性能が良かった。これは、今回収集した Web 上のデータに片方の固有表現しか使われていなかったことが考えられる。また、今回用いた固有表現は“パターンファイル”と“ファイアウォール機能”の二つのでしかないため、固有表現を更に用意し、OR 検索と AND 検索の結果ではどのような差が出るのかを調査する必要がある。

7. おわりに

本研究は、競合する他社の商品名のオントロジーを Web 上の情報から共起パターンと固有表現による絞込みで、構築できることを検証した。しかし、構築した競合する他社の商品名

のオントロジーの正解データに対する適合率は、改善する必要がある。

本研究は、11 の比較表現を利用したが、比較表現毎の抽出結果による比較表現の見直しが必要である。また、11 の比較表現以外にも利用できる比較表現が存在する可能性があるため調査を進めていくことが必要である。これらの成果として、共起パターンを整理し、再現率と適合率を共に上昇させることができると考えている。

自社の商品の固有表現を使ったフィルタでは、“パターンファイル”と“ファイアウォール機能”で有効性を示した。しかし、比較表現で抽出した結果に対し適合率は上昇したが、これは一般的なことである。だが、語の共起パターンから競合する他社の商品名の候補を抽出し、その候補から競合する商品名を自社の商品が持つ仕様で判定する方法を試行し、この判定方法が競合する他社の商品のオントロジーを構築する上で有効であることを検証した。

今後は、母集団の拡張や今回利用した比較表現以外の比較表現を調査し、利用することで、競合する他社の商品名が獲得できると考えられる。

8. 参考文献

- [1] テキストマイニング活用法—顧客志向経営を実現する
- [2] 産業構造審議会 新成長政策部会 中間報告:創造的産業組織の構築
- [3] 鈴木健之,丸山広,中村太一: Weblog から共起関係を利用して評判情報を把握する手法の提案”, 情報処理学会第 69 回(平成 19 年)全国大会, 5ZB-4, pp. 2-565 - 2-566(2007)
- [4] 丸山広, Web 情報活用基盤の研究 仕様と

評価の共起関係を用いた Web からの要求獲得手法 平成 18 年度東京工科大学大学院バイオ・情報メディア研究科修士論文 (2007)

- [5] Taichi Nakamura and Hiroshi Maruyama: "Extracting Opinions Relating to Consumer Electronic Goods from Web Pages", JOINT CONFERENCE ON KNOWLEDGE-BASED SOFTWARE ENGINEERING 2006(JCKBSE'06), IOS Press, 2006, pp.206-209(2006)
- [6] 立石健二,石黒義英,福島俊一: インターネットからの評判情報検索,情報処理学会研究報告,NL-144-11, pp.75-82 ,2001
- [7] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: テキストマイニングによる評価表現の収集, 情報処理学会研究報告,NL-154-12,pp.77-84,2003
- [8] 小泉敦子, 森本康嗣, 相菌敏子: テキストマイニングのための意味アノテーション, 人口知能学会
- [9] コトラのマーケティング・コンセプト pp.42-44
- [10] 細見格,「程度表現オントロジーの提案(1)コンセプトと設計」, 人工知能学会 セマンティックWebとオントロジー研究会, SIG-SWO-A603-02 ,2007.