

Title	評判情報分析のための製品属性異表記辞書の自動構築
Author(s)	劉, 朝いく
Citation	
Issue Date	2016-03
Type	Thesis or Dissertation
Text version	
URL	http://hdl.handle.net/10119/13614
Rights	
Description	Supervisor: 白井 清昭, 情報科学研究科, 修士

評判情報分析のための製品属性 異表記辞書の自動構築

北陸先端科学技術大学院大学
情報科学研究科

LIU CHAOYU

2016 年 3 月

修 士 論 文

評判情報分析のための製品属性 異表記辞書の自動構築

指導教員 白井清昭

審査委員主査 白井清昭
審査委員 東条敏
審査委員 飯田弘之

北陸先端科学技術大学院大学
情報科学研究科

1410047 LIU CHAOYU

提出年月: 2016 年 2 月

概要

日本語では、送り仮名の違い、長音の有無、字種の違い、あるいは全く異なる表現など、同じ実体が異なる文字列で表現されることがよくある。このような表現を「異表記」または「表記ゆれ」と呼ぶ。本論文では、製品属性を対象とした評判情報分析のための基礎的な知識として、製品属性の異表記辞書を自動的に構築する手法について述べる。

提案手法は大きく2つに分けられる。ひとつは、製品のウェブページなどから製品の仕様表を抽出し、さらに仕様表から製品属性を抽出して、初期の製品属性異表記辞書を構築する処理である。価格.com から取得した仕様表はフォーマットが決まっているため、属性と属性値は容易に抽出できる。一方、メーカーのページに掲載されている仕様表からは作ったルールにしたがって属性と属性値の組を取得する。仕様表から属性と属性値の組を抽出した後、同じ実体を表わす異表記の属性をひとつにまとめるために、凝集型クラスタリングアルゴリズムによって、属性のクラスタリングを行う。クラスタリング後、属性値を取り除いて、初期の製品属性異表記辞書を得る。もう一つは、製品に対するレビュー文を知識源とし、製品属性を抽出するパターンを獲得し、そのパターンを用いたマッチングによって異表記の製品属性を獲得する処理である。まず、あらかじめ決まったテンプレートを用いて、製品属性を抽出するパターンの候補を獲得する。テンプレートの中に、初期の製品属性異表記辞書に含まれる属性と評価語の間に出現する単語は最大に5個までである。初期の製品属性異表記辞書の中に含まれる属性を多く抽出できるパターンほど信頼性が高いとみなす。すると、マッチする文の数が3以上で、パターンにマッチしかつ初期の製品属性異表記辞書の中の属性が抽出される文の数とパターンにマッチする文の数の割り算の値は0.5以上のパターンを製品属性抽出パターンとして獲得する。得られたパターンをレビュー文集合に適用し、新たな製品属性を得る。最後に、仕様表から獲得した製品属性とレビュー文から獲得した製品属性を統合し、最終的な製品属性異表記辞書を得る。統合は、パターンマッチで獲得した属性を初期の製品属性異表記辞書に併合することで実現する。抽出するパターンによって得られる属性のうち、製品属性異表記辞書に既に含まれている属性の集合を K_j 、含まれていない属性の集合を U_j とおく。 K_j の要素は初期の製品属性異表記辞書における属性集合のいずれかに属する。 K_j の中で出現頻度が最大の属性集合を求め、その抽出するパターンは出現頻度が最大の属性集合の属性の異表記を抽出するためのパターンとみなす。そして、 U_j を最大の属性集合に追加する。

実験結果には、まず、仕様表から製品属性を抽出する手法を評価し、精度と再現率とも十分に高いことがわかる。次に、抽出された属性・属性値の組を評価する。精度は0.90となり、実用的な観点からも十分に高いことがわかった。次に、属性のクラスタリング手法を評価する。クラスタリングの Purity は、全体で0.829と高い。獲得されたパターンによって、初期の製品属性異表記辞書に登録されていない抽出された属性、そのうち、正しい異表記の属性とみなせるものの抽出精度は低く、改善の余地がある。

最後に、今後の重要な課題として、提案手法の実用的な評価が挙げられる。