D-031

評判情報利用によるネットオークションの商品選定支援システムの提案

A Goods Selection Support System of the Online Auction by Opinion Use

西村 圭亮† 湯浅 将英† 大山 実†
Nishimura Keisuke Yuasa Masahide Ohyama Minoru

1. はじめに

急速に成長するインターネット技術により情報が多様化する中，評判情報を扱ったサービスが増えている．現在注目されるサービスに，商品情報や商品の口コミ・評判情報を集めたサイト(kakaku.com[1])がある．さらに，Blog(Weblog)と呼ばれる日記形式に近い Web サイトの登場により，日記のみならず商品の評判や感想を含むサイトが登場してきた[1]．これらのサービスはユーザにとって商品購入時の意志決定支援に大きな役割をもっており，重要なサービスとなってきている．

一方，商品情報を扱ったサービスにネットオークションがある．これは近年のインターネットの普及により利用者数が大きく増加し，より身近な存在になってきている．

しかし，オークションにおいて商品を選ぶ際，購入時の参考となる情報はユーザ自身が集める必要があり，時間制限があるオークションでの素早い選定ができない場合がある．そこで上述したように商品購入時の意思決定支援に評判情報を用いれば効率的な選定が可能である．

本研究では，評判情報を用いたネットオークションの商品選定支援システムを提案する．評判情報の抽出に，肯定・否定の評判があらかじめ登録された評価表現辞書を用い，その賛否分類に SVM を用いた．本稿では，実際に作成したシステムの概要と評判の賛否分類手法及び分類精度の実験について述べる．

2. 試作システムの概要

本研究では，ネットオークションの商品の評判情報を収集し，賛否分類された評判のページを提示するシステムを作成した．

2.1 システムの流れと特長

実際の処理の流れを図 1 に沿って説明する．まず，ユーザは探したい商品のキーワードを入力する．商品情報収集モジュールは，そのキーワードから検索された商品情報を取得する．その際，複数のオークションサイトから商品情報を取得することで統合的に商品情報を網羅し，ユーザに商品情報のリストとして表示する．目的の商品の評判を見るには，商品情報リストページの商品タイトルをクリックする．それにより，商品タイトルが評判情報収集モジュールへクエリとして渡される．評判収集モジュールは WEB 上からページを収集し，収集されたページは解析モジュールに渡される．最終的に，ユーザに

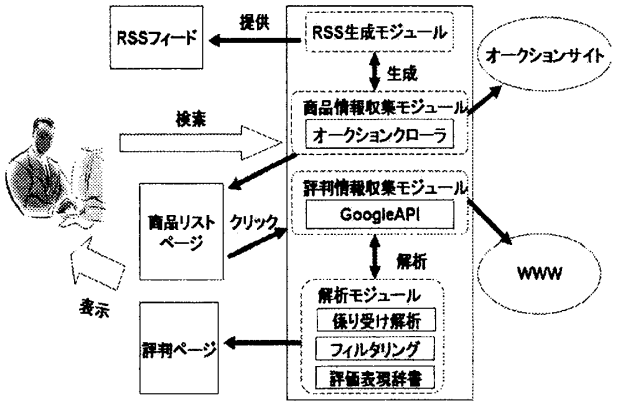

図 1: システムの構成と流れ

は賛否分類済みの評判ページが提示される．

2.2 システムの全体構造

(1) **商品情報収集モジュール** ：オークションサイトの商品情報のクロールには，複数のオークションサイトから商品情報を一括して収集するクローラを作成し，用いる．

(2) **評判情報収集モジュール** ：評判ページの収集には，Google の検索情報が利用できる GoogleAPI を用いる．

(3) **解析モジュール** ：このモジュールは評判情報の賛否を判定する．(2)で収集された HTML は後節で述べる係り受け解析のため，タグを全て取り除き，1 文ずつに区切る．区切られた 1 文は，Cabocha[2]により係り受け処理される．係り受け処理された 1 文(以下，係り受け文)は，非評価文の除去のため品詞によるフィルタリングをする．品詞は評価表現によく使用される形容詞，名詞-形容動詞互換，直前の単語の賛否を反転させる，助動詞　特殊・ナイ，助動詞　特殊・ヌの 4 種以外は取り除いた．フィルタリングされた係り受け文は，評価表現辞書に登録された肯定・否定表現とのマッチングし，賛否を分類する．

(4) **RSS 生成モジュール** ：このモジュールは(1)のクローラから収集した商品情報を RSS の形で生成するモジュールである．生成された RSS は，将来的に RSS リーダと連携させ，本システムの商品情報リストページを表示することが可能である．

3. 評判の賛否分類手法

本節では，2.2(3)で述べた評価表現の賛否分類に機械学習手法である SVM を用いた．SVM の実装には Libsvm[5]を用い，Kernel 関数には RBF を用いた．評判として扱う素性には，1 文を単位とする単語の係る関係を表した単語 N-gram(以下，素性)を採用した．関連する手法として藤村ら[3]の研究がある．本素性では Unigram から Trigram を採用し，また係り受け文は，文を最低限構成する名詞，

†東京電機大学 情報環境学部 情報環境工学科
1. http://www.kakaku.com/

形容詞，動詞，未知語の品詞だけを採用し，同一単語における活用の曖昧性をなくすため，語活用は基本形を採用した．表1に1例を示す．

4. 分類実験

本実験では，3節で述べた手法を検証する．

4.1 準備

実験で用いるデータセットは多くの評判情報を含む掲示板としてkakaku.comの口コミ掲示板を利用した．記事はカメラ関連の掲示板から97266文を収集し，2.2(3)の係り受け及びフィルタリングし，肯定26545文，否定25113文を得た．ここで，肯定6081文，否定4210文を学習用とし，4.2で述べる方法で分類した．さらに学習用データを訓練集合，検査集合に分割した．本実験では，3種類の実験をした．実験1では事例数の変化による分類精度を調べた．実験2では，実験1で学習したSVMが異なる商品ドメインにおいて有効であるか実験した．また，kakaku.comの掲示版では記事ごとにユーザの手によって肯定・否定の分類がなされている．実験3ではこの肯定・否定の評価を利用し分類した．

4.2 学習に用いる素性

学習に用いる素性は，学習に悪影響を及ぼすことが懸念される非評価文や間違った係り受け文などを人手で取り除いた．また，予備実験から学習単語数が多すぎると学習が困難となるため，肯定・否定表現において出現数が少ない単語，両表現の頻度に偏りがない単語は，各グラムで統計をとり，x^2検定により有意水準5%を満たす素性以外は取り除いた．その結果，肯定2234文，否定1545文の全3779文を得た．実験1では，これを用いた．

(1) 実験1

この実験は5分割交差検定とし，学習に用いる訓練集合はそれぞれ，300，600，1000，1500，2000をランダムにサンプリングした．また，Kernel関数のパラメータは，入力素性の影響を受けるため適宜最適なものを設定した．

(2) 実験2

この実験では，PC関連，家電関連を4.1と同様の方法で収集した．また，検査集合の素性数は実験1に合わせた．

(3) 実験3

この実験では，実験1，2と異なり，人手を用いない分類を検討するために実施した．まず，kakaku.comに書き込まれている評価を利用して上述と同様に統計を取った．その際，同一表現において多く言われる表現の賛否を正しいと仮定し，その賛否を用いた．ここで，分類実験をする上でN-gramの最後のグラムに比較的賛否が決定できる単語が多い知見を得たため，学習に用いる素性の賛否は最後のグラムの賛否を用いた．データセットは4.1で得た肯定26545文，否定25113文のうち，肯定7330文，否定4979文を用い上述と同様に検定した．有意水準50%を満たす素性以外は取り除き，肯定2178文，否定4477文を得た．他の条件は実験1と同様とした．

4.3 考察

実験1では，2000文の学習で82.3%の精度を得られた．また実験2では，事例数2000文で家電関連が

表1：採用する素性の1例

素性	肯定表現	否定表現
Unigram	感じ	電池
Bigram	柔らかい/ 感じ	電池/ 良い
Trigram	柔らかい/ 感じ/ 良い	電池/ 良い/ ない

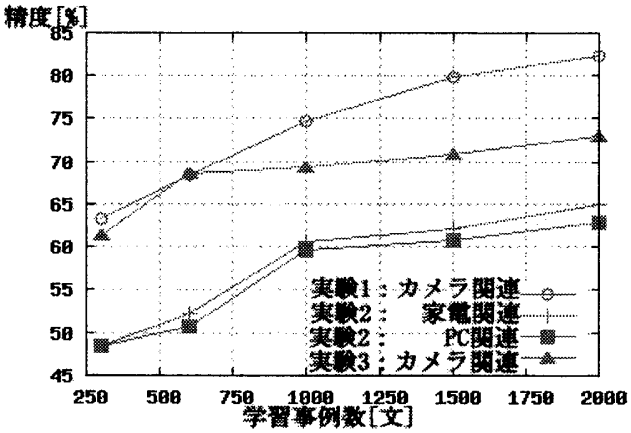

図2：学習事例変化による分類精度

65.0%，PC関連で62.9%の精度を得た．若干，前者の精度が高いのは実験1で学習したカメラ関連の評価表現が後者よりも多く含まれていたのが要因と考えられる．具体的には，カメラ，家電，PCに共通する表現として画質に関連する表現が多く述べられていたが，PC関連では液晶という表現が画質よりも多く使われていた，またキーボード，ファンと言ったPCドメイン固有の表現が家電固有ドメインよりも多く出現しており，それらが精度に影響したと考えられる．最後の実験3では，実験1ほど高い精度は示さなかった．この要因には，やはりN-gramの最後のグラムだけでは十分でないことが考えられる．

5. まとめ

本研究では，ネットオークションにおけるユーザの商品選定支援に評判情報を提示するシステムを提案した．また，試作システムの評価表現辞書の賛否分類の性能を示し，高い精度が得られることが分かった．しかし，非評価の考慮はしていないため今後は非評価も含めた肯定・否定・非評価の3値分類をしていきたい．また，解析モジュールにおける係り受け解析の時間コストが問題となっているため高速化も考えていく．

[参考文献]

[1] 奥村学，南野朋之，藤木稔明，鈴木泰裕 “blogページの自動収集と監視に基づくテキストマイニング” SIG-SW&ONT-A401-01, 2003

[2] 工藤 拓，松本 裕治 “チャンキングの段階適用による日本語係り受け解析” 情報処理学会論文誌 43, 6, 1834-1842, 2002

[3] 藤村滋，豊田正史，喜連川優 “文の構造を考慮した評判抽出方法” DEWS 2005, 6C-i8, 2005

[4] Chih-Chung Chang and Chih-Jen Lin “a library for support vector machines” 2001