

## 評判情報検索のための意見文収集に関する研究

清水 隆太<sup>†</sup> 東 基衛<sup>†</sup>

早稲田大学院 理工学研究科 経営システム工学専攻<sup>†</sup>

### 1. 研究の背景・目的

近年、インターネットが急速に広まり、Web 上では誰もが自由に情報を発信・公開できるようになった。またこれに伴い Web 上には人の意見や評価といった評判情報が多数存在するようになり、これらの評判情報を効率的に収集・分析する方法への期待が高まってきている。その中で、本研究では、商品の評判情報に着目して、「商品購入の際の情報収集の効率化」を研究目的とし、Web 上に存在する多量の情報の中から個人が発信する評判情報を収集し提供する「評判情報収集システム」を提案した。

### 2. 従来研究

従来研究として、藤村らの「文の構造を考慮した評判抽出手法」[1]がある。この研究では、評価表現辞書を作成し、商品の意見文群から各意見文の評価視点の抽出や肯定否定分類などを行い、評判情報をユーザに提示している。しかし、研究の焦点が「特定商品の意見文群から対象商品の評判を分析すること」にあてられており、どのように「特定商品の意見文群」を収集するかについては、今後の課題となっている。

### 3. 現状分析

本研究の中心である特定商品の意見文収集方法について調査を行った。調査 1 として従来研究で問題とされた、対象商品の情報収集方法について実際に意見文収集を行い、実験を試みた。調査 2 では、アンケート調査を行いユーザが評判情報の取得する際に最も期待するサイトについて調査を行った。

調査結果を以下に示す。

調査 1 では、クエリとして PC の商品名を入力し抽出された意見が対象商品に対するものか否か適合率を調べた。結果、平均 67.5% が対象商品の意見文であったが 3 割以上は対象商品の意見文でなかった。また、商品クエリによっては十分な量の意見文が得られず適合率、収集量において改善の必要性がある。

実験 2 では、アンケート結果から評判情報を最も期待するサイトを調査した。その結果 1 位は価格比較サイトのレビュー、2 位は掲示板、3 位が blog となった。しかし、従来手法では商品の表記の揺れに対応できておらず、また検索式も簡単なものであることから、価格比較サイトや掲示板などからの情報収集ができておらず、この点においても改善の必要性がある。

### 4. 研究アプローチ

#### 4.1 要件

研究目的・従来研究・現状分析より、本研究の要件を以下のようにまとめた。

- i. 商品の意見文の収集適合率を良くすること
- ii. 意見文の収集量を増やすこと
- iii. 収集する意見文からノイズを除去すること
- iv. 有用性の高いページから意見文を抽出すること
- v. 意見文を多く収集できるように条件式に改善すること
- vi. 商品の名称の表記の揺れに対応できること

#### 4.2 アプローチ方法

研究アプローチとして、「Web 上から意見文を収集する際のクエリの改良」「収集された Web ページからの有用性判定・ノイズ除去」を行った。これにより要件の実現を図り、対象商品の意見文の収集適合率を上げ、さらに収集意見文数を増やした。

### 5. 提案システム

#### 5.1 提案システム概要

以下にシステムの処理の流れを示す。システムはサーバ側に置かれ、ユーザはサイトにアクセスし、本システムを利用する。①ユーザは input として商品名・その商品が属するカテゴリ・検索したい商品の型番を選択する。②入力された商品名に対して、商品データベースから商品の正式名称や異表記等を取得する。③クエリとデータベースの情報からクエリを生成する。④生成したクエリで既存の検索エンジン(yahoo)により検索する。⑤検索された Web ページを解析し有用性判定を行う。不要と判断されたページを除去する。⑥さらにノイズ除去モジュール

Opinion Collection System for Commodity Reputation Retrieval

<sup>†</sup> Ryuta Shimizu, Motoei Azuma

Dept. of IMSE, Graduate School of Sci. & Eng., Waseda Univ

によりノイズと判断されたページを除去する。  
 ⑦あらかじめ作成しておいた評価表現辞書からドメインに従って評価語を取得し近接演算に用いる。  
 ⑧近接演算処理し各ページから意見文を抽出する。  
 ⑨適性値判定し意見文群をフィルタリングし意見文を取得する。  
 ⑩本システムにより抽出された意見文群を既存の評判情報分類・分析システムにインプットとして入れ、評判情報を取得する。

以下に提案システムの概要図を示す。

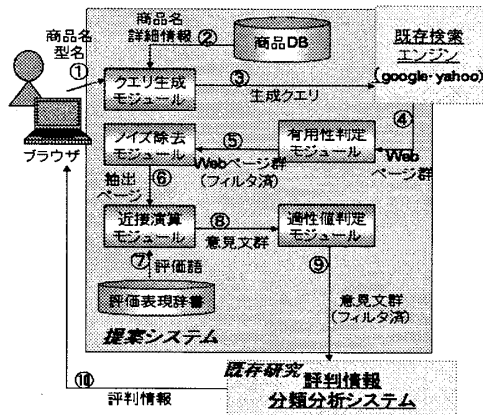


図1: 提案システム概要

## 5.2 クエリ生成

入力された商品名・型番から商品特定し、対象商品・型番の異表記名を商品データベースから取得する。これらの情報をクエリに含めることで、表記の揺れに対応する。さらに、条件式を改良することで、掲示板や価格比較サイトを含め、幅広いページからの情報収集を行う。具体的にはクエリとして次の条件を追加する。

+intitle: レビュー or レポート or 口コミ or クチコミ  
 +inurl: blog or bbs or review

また、クエリを検索エンジンに送信する際には、意見文を必要量集められない場合を除き、商品異表記名称を用いないなどの配慮を行うことで、適合率を下げずに意見文収集量を増やす。

## 5.3 有用性判定・ノイズ除去

既存検索エンジンから検索結果を取得し、Web ページに対し次の有用性判定処理を行う。この処理では、有用性判定の関連研究である福島らの「Web ページの信頼性の自動推定」[2]を利用する。この研究では実験により有用であるサイトの特徴を40個の尺度として決定しており、この指標の一部を利用する。さらに評判情報の情報源として有用と考えられる判定基準を新たに設定し、取得した Web ページが評判情報の収集源として有用であるか否かの判定を行う。

次に、ノイズ除去として、現状分析で最もノ

イズとして抽出されていたリンクテキスト文を除去する処理を行う。具体的には取得した Web ページ内に<a href="...">と</a>で囲まれた部分を削除する。

## 6. 評価実験

### 6.1 実験・評価方法

本研究の有効性を検証するために、検索対象を PC に絞り、プロトタイプを実装し、評価実験を行った。被験者として24名に利用してもらった。評価方法としては、ユーザに評判情報を取得したい商品を選択してもらい、本システムにより収集された Web ページが対象商品に関するページであるか否かを判定してもらい適合率を検証した。

### 6.2 実験結果

提案システム・既存システム・フィルタリング前の提案システムの3手法における Web ページの適合率について以下に結果を示す。

表1: 検証結果 (適合率)

手法	適合率
提案システム(フィルタリング前)	59.3%
提案システム(フィルタリング後)	81.1%
既存システム	61.5%

表1から提案システム(フィルタリング後)が最も適合率が高く、提案システムの有効性が示された。さらにフィルタリング前後において適合率が向上していることからフィルタリングモジュールの有効性も確認できた。

以上の結果からクエリの改良・Web ページのフィルタリングを行うことで評判情報検索システムのための意見文収集が効率的に出来る事が証明できた。

## 7. 結論と今後の課題

本研究では Web 上からの効率的な意見文収集の方法を提案した。得られた実験結果から Web ページのフィルタリングを行うことで、対象商品の意見文収集における適合率を向上させることが出来た。

今後の課題としては、商品の表記の揺れに対応するために作成した商品データベースや、評価表現辞書をどのように自動で構築するのか。また意見文収集数・適合率をどのように向上させていくことなどが挙げられる。

### 参考文献

- [1] 藤村滋他, 文の構造を考慮した評判抽出手法, 電子情報通信学会, 第16回データ工学(2005)
- [2] 福島隆寛, Web ページの信頼性の自動推定, 知能と情報, Vol19 No.3(2007)