ストーリーを持ったアイテムに対する レビュー文からのあらすじ除去に関する基礎検討

A Basic Study of Deleting Stories from the Reviews to the Items with Stories

池田　郁　　土方　嘉徳　　西田　正吾
Kaori Ikeda　Yoshinori Hijikata　Shogo Nishida

大阪大学基礎工学部基礎工学研究科
Graduate School of Engineering Science, Osaka University

Recently,many commercial web sites provide a function that users can write and see reviews on items(e.g. Amazon.com and Kakaku.com) . A number of people see those reviews when they want to know the information about their interested items. We can evaluate the items using the valuable reviews. However,there are some reviews that have "storys" as well as valuable information. (In this paper,we call items' contents "storys".) If you know the plots from reviews, you might feel disappointed or don't want to read or see the item anymore.
In this paper,we propose the system that eliminates storys from reviews and accents valuable information.

1.　はじめに

近年，WWW 上には，Amazon.co.jp [*1] や価格.com [*2]などのように，アイテムに対してレビューを簡単に作成・閲覧できるサイトが存在する．あるアイテムについての情報を得たい場合，レビューサイトを閲覧する人も多い．レビューに書かれた他者の意見は，対象となっているアイテムの良し悪しを判断するのに大いに役立つ．だが，小説や映画などのストーリーを持ったアイテムに対するレビューには，レビュアの意見と同時に，あらすじが書かれている場合がある．あらすじとは，ストーリーの一部分のことをいう．レビューによりあらすじがわかってしまうと，実際に小説や映画を見た時の楽しみが減ってしまう．実際，様々なアンケートをインターネット上で行っているサイト[*3]では，コンテンツのネタばれに関するアンケートを行っており，50%以上の人がレビューによってストーリーの内容を知りたくないと回答している．

そこで本研究では，ストーリーを持ったアイテムに対するレビューについて，あらすじを自動的に除去し，意見部分を表示するシステムを提案する．ここで，あらすじとはストーリーの一部についての記述を指し，意見文とはレビュアのアイテムに対する感想や評価を指す．

2.　システム

本システムは図 1 のようになっている．本システムはクライアントの要求により動作を行う Web アプリケーションとして実装を行った．このシステムでは，ユーザから入力としてレビューを見たいアイテム名とウィンドウ幅を与えてもらう．それに対して, 入力されたアイテムにつけられた全レビューについて，あらすじを除去し意見文を強調するような表示になるよう HTML を書き換え，出力としてユーザに返す．ユーザはこれをブラウザで見ることができる．システム内部では，受け取った入力に対しレビュー文抽出，形態素解析，あらすじ・意見部分の抽出という３つの処理を行って出力をしている．この３つの処理について以下で述べる．

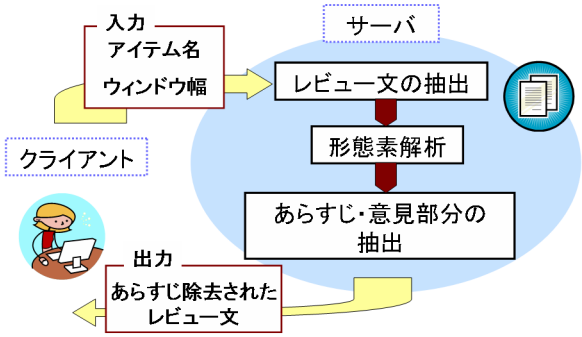

図 1: システムの概要

2.1　レビューの抽出

本システムでは，アイテムに関する情報（アイテム名，アイテム ID，著者名，発行年月日，つけられたレビュー文集合）をあらかじめアイテム情報データベースに登録しておく．これに対しアイテム名を入力として与え，レビュー文集合を得る．

2.2　形態素解析

レビュー文のような人間の手で記述された文章は表現が多種多様であり，計算機で扱うことは困難である．そこで，まず文章に対し，意味のある要素（これをトークンと呼ぶ）ごとに分解する処理を行う．この処理は形態素解析と呼ばれる．本システムでは，レビューの抽出により得られたレビュー文全てに対してこの形態素解析処理を行う．

2.3　あらすじ部分・意見部分の特定

見せたい部分を強調させたり，見せたくない部分を隠すためには，それぞれの部分の特定が必要である．

ストーリーを持ったアイテムに対するレビューを調査すると，あらすじ部分には，登場人物の名前が頻繁に出現している．意見部分については，感想・評価を示す単語（「すばらしい」「感動した」など）が出現している．本研究では，このことに注目

連絡先: 池田　郁，大阪大学大学院基礎工学研究科，大阪府豊中市待兼山町 1-3，TEL:06-6850-6383，FAX:06-6850-6341，ikeda@nishilab.sys.es.osaka-u.ac.jp

*1　http://www.amazon.co.jp
*2　http://kakaku.com
*3　http://www.enquete.ne.jp/hundred

し，あらすじ部分・意見部分を特定する．

具体的には，まず，人名を集めて人名辞書として保存する．同様に，感想・評価を表す語を集めて意見辞書として保存する．そして，レビュー文に対し形態素解析の処理を行ったあと，各トークンと辞書とを比較し，人名辞書に合致する要素を「人名」，意見辞書に合致する要素を「意見語」と特定する． さらに，その語の前後に幅を設け，その部分をあらすじ部分，意見部分と決定する．この幅のことをウィンドウ幅と呼び，ユーザの好みを反映できるように入力として与えてもらう．

2.3.1 辞書の概要

本システムでは，上述の意見辞書，人名辞書として次のものを利用している．

- **意見辞書:** 意見辞書については，evaluative expressions[1] を使用している．これには約 5200 語が収録されている．レビュー文のような，比較的自由で堅苦しくない形態で書かれる文章において使用される表現（「ヤバイ」「ショボい」など）も収録されている．

- **人名辞書：** 人名辞書については，既存のもので語数が十分だと思われるものが存在しなかったので，WWW 上から収集を行った．

  特に，日本人の名前は豊富にあるため，日本の男女の名前上位 40000 件を収録している ”同姓同名辞典 ”[2] を辞書として保存した．名字については，約 24000 件を収録している ”日本の苗字七千傑 ”[3] を辞書として保存した．ストーリーの登場人物には外国人が登場する可能性があるため，外国人の姓・名もあわせて約 16000 語を保存した．これらは，鷹書房弓プレス社発刊の ”英米人の姓名 ”[4] や WWW 上の辞書から収集した．この中には，イギリス人・フランス人・ドイツ人などの名前が含まれている．

  実際に辞書を使用してみると,「愛」「勇気」などの一般語が人名と判定される例が多くあった．このため，人名辞書に残っている語の中から，国語辞典に姓名以外として掲載されているものを一般的な名詞とみなし，これを除いた．これには yahoo!辞書 *4で使用されている国語辞典「大辞林」[5] を使用した．

  また，レビュー文においては，ストーリーの登場人物に関して「主人公」など特定の言葉で表現される場合も多い．よって，このような特定の語を役名と定義し，役名についても人名辞書に収録した．役名には，実際にレビュー文で頻繁に使われていた「主人公」「彼」「彼女」などの語を採用している．

2.4 結果の表示

あらすじ部分・意見部分の結果の表示の方法は，ユーザの満足度に大きく関与する部分である．いくら正確にあらすじ部分を特定できても，その表示の仕方が悪ければユーザは満足しない．本研究では現在，３つの表示の方法を考えている．色分けによる表示，背景色と同化させる表示，タブによる表示である．

現在，本システムでは，結果がわかりやすいという理由から，第一に挙げた色分け表示を採用しているが，今後他の方法も実装し，ユーザに使ってもらうことでその評価をシステムに反映したいと考えている．

図 2 に，提案システムの出力例を示す．この例では，Amazon.co.jp のレビューに対し本システムによる処理を行い，結果を表示している．出力の際は，レビュー内でのあらすじを青，意見部分を赤でマーキングして表示している．図 2 では，「ダンブルドアがスネイプに殺された」というあらすじ部分がマーキングによりうまく隠されている．

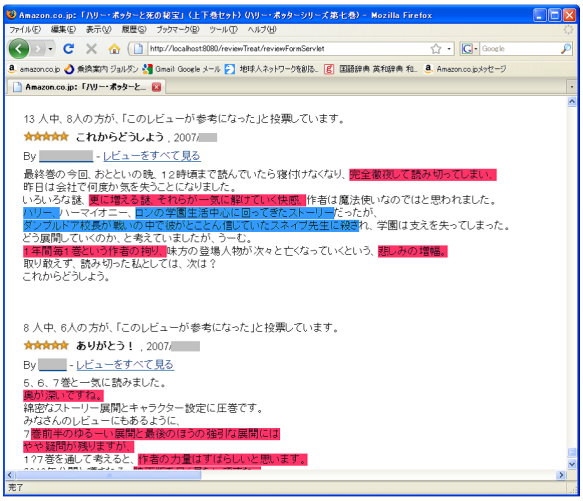

図 2: 提案システムの出力例

3. おわりに

本研究では，ストーリーを持ったアイテムに対するレビュー文について，あらすじ部分を自動的に除去して, 意見部分を強調して表示するということに着目している．あらすじ部分・意見部分を特定する方法として，形態素解析による品詞分類と人名辞書・意見辞書を用いてレビュー文から人名・意見語を探す．次に，人名・意見語をもとにウィンドウ幅を用いてあらすじ部分・意見部分を特定する．そして，適切なインタフェースを用いて，あらすじ部分を隠して意見部分を表示するようなアプリケーションの作成を目的とし，実装を行った．

今後の予定として，あらすじを表す語をより多く特定できるようにすることと，インタフェースの検討が挙げられる．具体的には，比較的きれいな文である商品紹介文に対して係り受け解析を行い，その結果により人名を特定を行いたいと考えている．また，結果の見せ方の実装を進め，実際にユーザに使用してもらい，インタフェースの評価を行う．そしてその評価をシステムに反映させる予定である．

参考文献

[1] evaluative expressions, http://www.syncha.org/evaluative_expressions.html

[2] 同姓同名辞典, http://www.douseidoumei.net/00/mei01.html

[3] 日本の苗字七千傑, http://www.myj7000.jp-biz.net/freq.htm

[4] 木村正史, 英米人の姓名, 鷹書房弓プレス,1980.

[5] 松村明, 大辞林, 三省堂, 第二版,1995.

*4 http://dic.yahoo.co.jp