社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－NL－164（28）
2004／11／6

嗜好を考慮した評判情報検索手法

矢野 宏実† 目良 和也‡ 相沢 輝昭‡

†広島市立大学大学院情報科学研究科
‡広島市立大学情報科学部

E-mail: †yano@nlp.its.hiroshima-cu.ac.jp, ‡{mera,aizawa}@its.hiroshima-cu.ac.jp

あらまし 本論文ではWeb上の飲食店の評判情報を自動抽出し，個人的嗜好に沿って順位付けするための手法を提案する．この手法は店舗の情報を得る部分，評判情報から評価情報を抽出する部分，評価情報と個人的嗜好情報をもとに各店舗を評価する部分から成る．本論文では，特に評判情報から評価情報を抽出する部分を扱う．この部分は，Web文書の構文的特徴を考慮した抽出ルールによって，対象とその属性表現だけでなく，属性の程度や情報の信頼度を表す副詞やモダリティ表現も抽出する．本手法を実際の店舗情報に適用した結果，評価文の抽出精度は79.70%,評価情報の抽出精度は66.18%であった．

キーワード 属性，信頼度，嗜好，評判情報検索

Opinion Information Retrieval Method with Personal Taste Information

Hiromi YANO† Kazuya MERA‡ and Teruaki AIZAWA‡

†Graduate School of Information Sciences, Hiroshima City University
‡Faculty of Information Sciences, Hiroshima City University

3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194 Japan

E-mail: †yano@nlp.its.hiroshima-cu.ac.jp, ‡{mera, aizawa}@its.hiroshima-cu.ac.jp

Abstract In this paper, we propose a method to retrieve opinion information for food stores automatically from WWW and to evaluate the information based on personal taste information of the user. This method consists on three sections; store information retrieval section, opinion information extracting section, opinion information evaluating section. In order to evaluate opinions for foods, both the degrees of attributes and the reliability of the opinions are important. Therefore, our method calculates such information from grammatical features such as adverbs and aspects. The precision of extracting opinion sentences method and extracting opinion information wa s 79.70 percent and 66.18 percent, respectively.

Keyword attribute, reliability, taste information, opinion information retrieval

1. はじめに

インターネット上には多くの情報が存在し，その中には様々なものに対する評価を掲載したサイトがある．これらから必要な情報を適切に収集できれば,対象を評価するうえで非常に有効な情報となる．しかし，対象に関する用語を用いて検索を行なっても,自分が必要としない情報まで数多く抽出されてしまう.例えば，広島のおいしいラーメン屋を探すため，“ラーメン広島”をクエリとして検索を行なっても，得られた105000 件から自分の好みに合った店を探すのは困難である．

そこでWebサイト内容を解析し，タグ情報や構文構造をもとに対象に関する知識を獲得する手法が提案されている．佐藤[1]はタグ情報をもとに，全国の水族館などの施設の名前，住所，電話番号などを自動収集する手法を提案している．しかし，この手法では収集する対象が定型的なものであるため，自然言語で記述された評判や意見の収集には適していない．

一方，立石ら[2]は，商品名（評価対象）の近くに評価を表す語である評価語があるかどうかで評判情報かどうかを判定して意見の自動収集を行う手法を提案している．この手法では，収集した意見を各属性ごとに肯定／否定の２値に分類し，意見の全体数から情報の信頼度，肯定／否定の割合から対象の満足度を算出している．しかし，この手法では，飲食店のように味覚に深く関係する対象を評価する際，同じ評価語でも個

人の嗜好によってその良し悪しが異なるような状況に対処できない．また，味覚のような曖昧なものを評価するには程度の情報（「とても～だ」や「～と思う」など）が必要だが，これらにも対処することができない．

そこで本研究では従来研究をもとに，飲食店の情報を対象とした評判情報自動抽出手法および個人的嗜好を考慮した評判情報評価手法を提案する[3][4][5]．収集される評判情報は，基本的に評価対象とそれに対する属性表現からなる．本研究では，自然言語で記述された文書から，これらの評判情報を抽出する．しかし，Web 上に存在する文書は話し言葉のようなくだけた口調も多く，表記の多様性や格要素の省略などの影響もあるため，文書中から評価文を適切に抽出することは難しい．

そこで，本研究では，このような表現に適した評価文抽出ルールを作成することで，Web 文書上の評価文の特徴を捉え，適切に評価文を抽出することを目指す．

そして，抽出した評価情報と事前に与えられた個人の好みに関する情報（嗜好情報）をもとに店舗の総合評価度を求める．総合評価度の値が大きいほど嗜好と一致しているものとみなし，降順に順位付けしたものを最終結果としてユーザに提示する．

2 章ではシステム全体の構成及び各部の処理について説明する．3 章では Web 文書からの評判文の抽出手法と評判文から評価情報を抽出するための手法について説明する．そして 4 章でこれらの手法について評価実験を行なった結果を示す．

2. 評判情報検索システム

本手法に基づいて作成するシステムは図 1 のような 3 部構成となっている．まず，入力である Web 文書に対して，店舗情報抽出部で店舗名とその店舗に関する情報を抽出し，評価情報抽出部に送る．評価情報抽出部は，評判に関する文書から評価表現として適切なものを抽出し，総合評価度の計算に必要なデータ集合を評価度算出部へ送る．評価度算出部では，ユーザから与えられた嗜好情報を参照して店舗の総合評価度を求める．出力は総合評価度の高い順に店舗情報の一覧を表示する．

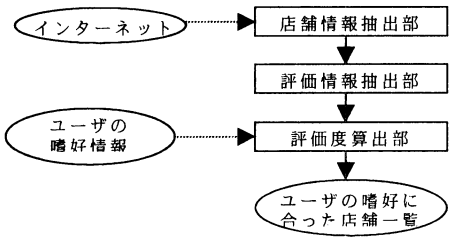

図 1　評判情報検索システム

2.1. 店舗情報抽出部

店舗情報抽出部は Google[6]で「ラーメン　広島」をクエリとする検索で得られたラーメン屋に関する記述のあるサイトを入力とし，店舗名と店舗情報を対応付けたものを出力とする．店舗情報とは飲食店に関する情報を含む文書のことで，飲食店の評判情報の他に店舗名や住所，営業時間，定休日，電話番号といった付加情報を含んでいる．店舗情報の例を図 2 に示す．

店舗情報抽出部の処理手順を図 3 に示す．まず，Web から広島のラーメン屋を検索し，Wget で Web ページをローカルのハードディスクに保存する．次に，ラーメン屋の店舗名を HTML タグによる表構造に注目して抽出する．そして，店舗名をクエリに加えて再検索を行い，その店舗の店舗情報を HTML タグによる表構造と正規表現で記述されたルールによって取り出し，「店舗 A から別の店舗 B が文章中に表れるまでは A に関する話題である」というヒューリスティックや付加情報をもとに店舗名と店舗情報の対応付けを行う．

店舗名　：陽気
住所　　：広島市中区江波南○－○－○
営業時間：16:30～24:00
定休日　：1,12,13,26 日
電話番号：082－○○○－△△△△
評判情報：豚骨ベースだが野菜もたっぷり使っているのでまろやかで甘みもある。・・・

図 2　店舗情報例

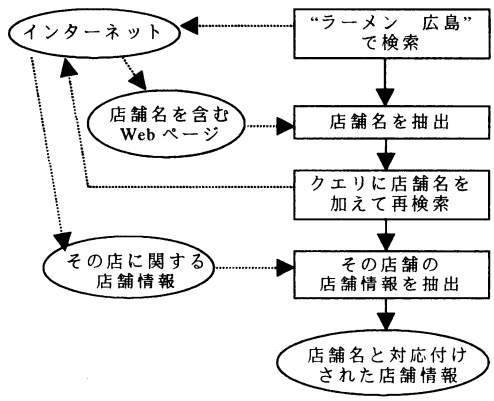

図 3　店舗情報抽出部

2.2. 評価情報抽出部

評価情報抽出部は，評価文の判定と評価情報の抽出という 2 つの処理で成る．店舗情報抽出部で抽出した

1 店舗分の店舗情報を入力とし，店舗情報から評価文を選別，評価文から取り出した＜評価対象語・評価語・様相＞の 3 つ組のデータを出力とする．この 3 つ組のデータを評価情報という．評価対象語とは属性を持つ対象で，「ラーメン」「スープ」「味」などがある．評価語とはラーメンを評価した結果を表す属性表現で，「おいしい」「あっさり」「辛い」などがある．様相とは副詞，否定，二重否定，モダリティ表現といった構文情報のことである．様相は，評価度算出の際に属性の程度と信頼度を得るための情報となる．

評価文抽出の処理の流れを図 4 に示す．まず，店舗情報を受け取り，茶筅[7]を用いて形態素解析を行う．次に，1 文ずつ評価文として適切かどうかを判定し，適切と見なしたものについてのみ評価情報を取り出す．

評価文の判定および評価情報の抽出については 3 章で詳細を述べる．

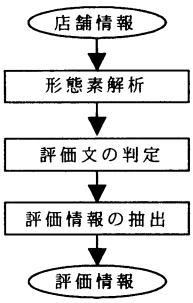

図 4　評価情報抽出部

2.3. 評価度算出部

評価度算出部では，評価文抽出部で得られた 1 店舗分の評価情報を入力とする．これをもとに，各属性に対する程度と信頼度を計算する．さらに，この店舗がどのくらいユーザの嗜好と一致しているかを表す総合評価度を求める．

評価文から評価情報を取り出して，総合評価度を得るまでの流れを図 5 に示す．

まず，評価情報を属性ごとに統合しても検定を行い，各属性の属性値とその信頼度を求める．属性値とは属性の強度のことで，様相の副詞に割り当てられた吉江ら[8]の発話意図抽出手法で用いられている肯定値変化倍率の値に従う．また，信頼度とはその属性値の確かさで，様相のモダリティ表現に割り当てられた青山[9]の肯定／否定モダリティ表現の確信度の値に従う．

次に，属性値に，あらかじめ定義しておいたメンバーシップ関数を各属性ごとに適用して満足度を求める．この満足度と先ほどの信頼度を掛けて属性評価度を求める．そして，この値と，事前にユーザに各項目を一対比較してもらう事で得た嗜好値でファジィ積分[10]を行って総合評価度を得る．

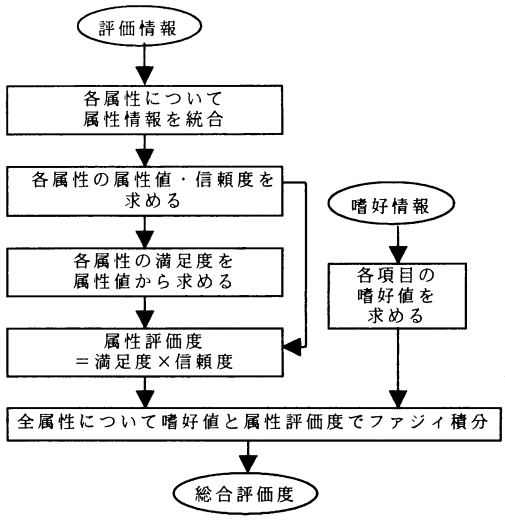

図 5　総合評価度算出

3. 評価情報抽出手法

3.1. 処理の流れ

店舗情報抽出部から受け取った店舗情報に含まれる評判情報から評価文を選別して評価情報を得るまでの流れを具体的に説明する．処理の流れを図 6 に示す．

まず，評判について記述した文書を受け取り，茶筅[7]を用いてその形態素解析を行なう．次に区切り判定ルールを適用して，どこまでを 1 文とみなすかを判定する．そして，区切られた 1 文ごとに妥当性判定ルールを適用してその 1 文が評価文として適当であるかを調べる．適当とみなしたものについて，解析結果修正ルールで形態素解析結果の誤りを正し，評価文抽出ルールを適用して評価情報を取り出す．各々のルールは店舗情報 58 件（総文数 2246 文，評価文数 494 文）を解析し、解析誤りや語の並びについて頻出パターンを集めることで作成した．

3.2. 評価文抽出

3.2.1. 解析結果修正ルール

解析結果修正ルールは，評価文の選別をしやすいように形態素解析結果に修正を加えるものである．ルールは次の 4 つである．

- 形態素解析結果の誤り訂正

評価対象語・評価語のなかには，形態素解析に誤りがあるものがある．その中で，常に同じ誤りを生じるものを事前に登録しておき，修正を行なう．

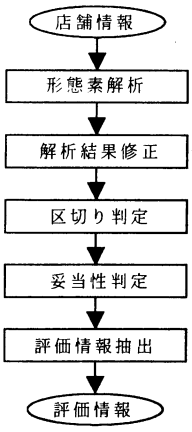

図 6　評価情報抽出部詳細

（例）とんこつ・醤油　→　とんこつ醤油
（形態素解析の結果，・のところで分けられてしまう）

- 文中の指示語の補完

特定の指示語（「これ，それ」「この，その」）と「どれ・も」という表現を含み、かつ，それが接続助詞，評価対象語より前にあるとき，これまでの文で最後に出現した評価対象語を代入する．

（例）スープはかなりこってりしています。
でも、それがまたよくて、
→　スープがまたよくて、

ただし，次の例のように，接続助詞，評価対象語より後ろの代名詞は，これまでの文の評価対象語を受けているとは限らないので，修正しない．

（例）縮れた麺がこのスープに合います。

- 表現の統一

これは評価度算出部での処理を簡単にするための処理である．

（例）トンコツ　→　とんこつ

- 読点の除去

これは，読点が含まれていると評価情報抽出ルールで評価情報をうまく取り出せないためである．

（例）チャーシューは、とても柔らかくて、好みでした。
→　チャーシューはとても柔らかくて
好みでした。

3.2.2. 区切り判定ルール

評価情報の抽出は 1 文ずつ行なう．しかし，1 文が長すぎたり，短かすぎたり，文中に評価情報の取り出しに不適切な部分を含んでいたりすると評価情報を適切に取り出せない．そこで，区切り判定ルールを用いて文を適切なところで区切って，評価情報を取り出しやすい長さの文にしたり，不適切な部分を除外しやすくする．

このルールでは，主に句点や空白，接続助詞，比較表現で文を区切っている．

ルールの詳細と適用例を以下に示す．適用の結果区切られる個所を／で示す．

- 句点，！などの記号，空白
（例）うまい！／そう思いました。／
- 「て」「で」「し」「ば」「ながら」「ので」「から」「と」以外の接続助詞
（例）美味しいのだが／途中で飽きてしまった。

ただし，順接の接続詞の場合話題が継続していることが多いため（例：チャーシューが柔らかくておいしい。），ここで区切ると後半の評価対象語が評価情報抽出ルールの補完ルール（3.3 参照）によって評価対象語が変化する可能性があるので，ここでは区切らない．

- 「より」「ほど」（比較表現）
（例）見た目よりも／やや薄い。

3.2.3. 妥当性判定ルール

妥当性判定ルールは，区切り判定ルールで区切られた 1 文が評価文として適切かどうかを判定するルールである．ここで、評価文として妥当であるとは，その 1 文がラーメンの実際の味や評判の評価を表す内容になっている，ということである．判定手順とルールの適用例は以下のようになっている．

1.　1 文中に評価語が一つ以上存在するか
（例）ベースは鶏でしょうか、
（評価語が無いので評価文ではない）
味はこってりしています。
（評価語「こってり」があるので評価文）

2.　1.で評価文とみなされたものであっても、次のような文は評価文としない

- 負例辞書に登録された語を含む
負例辞書にはラーメンの評価に用いられない語（「楽しい」「遠い」等）が登録されている．
（例）明るくて清潔な店内に、
- 「以前は」「前は」等を含む（過去の情報）
（例）以前はもっと濃厚だった。
- 「一見」「見た目は」等を含む（外見）
（例）見た目はこってりですが、
- 仮定・提案・要求

(例)油を少なめにすれば美味しくなるでしょう。

- 主部（ハ格，ガ格，モ格のみを対象とする）があるとき，そこに評価対象語を含まない
  (例) 店員の対応は普通でした。
- 「私は」「個人的に」等を含む
  (例) 私はあっさり好みですが、
- 「まるで」「（～の）よう（な）」を含む（直喩）
  (例) 黄身のようなコクが楽しめる。
- 「～（という）より（は・も）」を含む（比較）
  (例) 塩辛いというよりも、
- 連体化「の」以前に評価対象語・評価語のどちらも無いとき，それ以降に評価対象語のないもの
  (例) サービスのいいお店でした。
  （評価文とみなさない）
  こだわりのうまいラーメンをだす。
  （評価文とみなす）

なお，負例辞書とは，ラーメンの評価に用いられない語を登録したものである．ここに登録された語はラーメンの味や評判以外のものを評価する語なので，こららを1文中に含むものは評価文ではない可能性が高い．負例辞書に登録されている語の一覧を図7に示す．

楽しい，好き，暖かい，寒い，陽気，強引，若い，年配，遠い，近い，清潔，明るい，暗い，広い，狭い，わかりやすい，わかりにくい，黒い，真っ黒、黄色い

図7　負例辞書

3.3. 評価情報抽出

抽出した評価文から評価情報を取り出すには，評価情報抽出ルールを用いる．これは文中の評価対象語と評価語の関係を調べて適切に評価情報を取り出すためのルールである．その一部を図8に示す．ルールの内容と適用例は以下の通りである．

- 評価文に評価対象語と評価語の関係を調べるルールを適用し，「評価対象語」「評価語」「様相」を取り出す
  (例) チャーシューはまったく臭みがない。

適用ルール
＜評価対象語＞＜格助詞｜係助詞＞＜副詞＞*＜評価語＞＜否定語＞*

評価対象語　チャーシュー
評価語　臭み＋否定
様相　まったく

- 評価文中に評価語が存在しないとき，評価語などから推測する
  推測に使う語とそれから推測される評価対象語の組み合わせを表1に示す．表中に該当するものが無い場合は評価対象語を「ラーメン」と補完する．
  (例) 結構コシがあります。
  評価対象語　麺
  評価語　コシ
  様相　結構

- ＜評価対象語＞＜格助詞｜係助詞＞＜副詞＞*＜評価語＞＜否定語＞*
- ＜評価対象語＞＜の（連体化）＞＜評価語＞＜評価対象語｜評価語＞
- ＜評価語＞＜係助詞｜の（連体化）＞＜評価語｜否定語＞
- ＜評価語＞＜格助詞｜の（連体化）＞＜評価対象語｜動詞－ある＞＜評価語＞*＜評価対象語＞*
- ＜副詞＞＜評価語＞＜評価対象語＞*

図8　評価情報抽出ルール（一部）

表1　評価対象語の推測

評価語等	推測される評価対象語
こく,飲む,飲み干す,あっさり,こってり,すっきり	スープ
こし,茹でる,縮れ,太い,細い,ストレート,小麦（粉）	麺
多い,少ない	量
高い,安い,円	値段
厚い,トロトロ,ばさばさ,枚	チャーシュー

4. 評価実験

評価情報抽出手法で用いた各種ルールの作成に使用した店舗情報58件（Closed test）と使用しなかった店舗情報89件（Opened test）．に本手法を適用した結果を表2に示す．

表2　評価情報抽出手法適用結果

テスト名（総文数）	評価文の抽出（再現率／精度）	評価情報の抽出（再現率）	誤認評価
Closed test (2246)	462 (93.52%／86.91%)	394 (73.51%)	74
Opened test (2499)	436 (79.70%／79.71%)	362 (66.18%)	111

評価文の抽出は，システムが出した結果と事前に人手で抽出した結果が一致しているものを正解とみなす．評価情報の抽出に関しても同様だが，こちらは1文から複数の評価情報が取り出せる場合があるので，人手で抽出した結果と完全一致した場合を正解としている．

最右列の「誤認評価」は，評価文ではない文を評価文と認識したもののうち，各種ルールで除外できなかったものの数である．

評価文抽出の適合率と精度は以下のような計算で求めている．

再現率(%)＝「評価文の抽出」
／「人手で抽出した正解の評価文」×100

精度(%)＝「評価文の抽出」
／(「評価文の抽出」+「誤認評価」)×100

また，評価情報抽出の適合率は以下のようにして求めている．

適合率(%)＝「評価情報の抽出」
／(「評価文の抽出」+「誤認評価」)×100

Closed test に比べ，Opened test の適合率，再現率がやや低いのは，未知の評価語の存在と未登録の形態素解析結果誤りの出現が大きな原因であった．

Closed test，Opened test に共通する誤認評価の原因としては，メニュー名・店名・書名を評価文とみなしてしまう,省略された評価対象語の補完が誤っている，形態素解析結果が誤っている，などがあった．

5. おわりに

本論文では飲食店の評判情報を自動抽出し，個人の嗜好情報に沿った順位付けをするための手法を提案した．この手法は，Web からの店舗情報を抽出する部分と，評判を記述した文書から各属性に対する評価情報を抽出する部分と，抽出した評価情報から情報の信頼度や個人の嗜好情報をもとに各店舗の評価付けをする部分から成る．本論文では特に文書からの評判情報の抽出処理について詳説した．

今回対象としている飲食店の評価付けは味覚という曖昧なものを扱うため，評判情報として属性の強度と信頼度を抽出できるようにした．

本手法の評価情報抽出部について実験を行なった結果，評価文の抽出精度は 79.71%,評価情報の抽出精度は 66.18%となった．

今後の課題は，評価度計算手法の検討，属性値の信頼度の計算手法と評価度への影響の調査と共に，自然言語からの評価情報収集手法の強化や各属性に応じた満足度算出用メンバーシップ関数の作成を行なうことである．

文　献

[1] 佐藤理史，“ワールドワイドウェブを利用した住所検索，”情報処理学会論文誌，Vol.42，No.1，pp.59－67，Jan.2001.

[2] 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，”情報処理学会自然言語処理研究会，NL-144-11，pp.75－82，July.2001.

[3] 笠信太郎，“Web からの店舗情報抽出，”平成 15 年度広島市立大学大学院情報科学研究科修士論文，Feb.2004.

[4] 矢野宏実，“嗜好を考慮した評判情報検索のための評価文抽出手法，”平成 15 年度広島市立大学情報科学部卒業論文，Feb.2004.

[5] 安善奈津美，“嗜好を考慮した評判情報検索の評価計算手法，”平成 15 年度広島市立大学情報科学部卒業論文，Feb.2004.

[6] Google　http://www.google.co.jp/

[7] 松本祐治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸，日本語形態素解析システム『茶筅』Ver．2.3.3，奈良先端科学技術大学院大学，Aug,2003.

[8] 吉江誠，“真偽疑問文に対する返答発話の肯定／否定意図解析手法の改良，”平成 14 年度広島市立大学情報科学部卒業論文，Feb.2003.

[9] 青山広，“真偽判断と確信度，”計量国語学，第 21 巻第 1 号，pp.1－10，June.1997.

[10] 高萩栄一郎，“人間の感性とファジィ積分を使ったデータ検索システム，”経営情報学会秋季全国発表大会，Nov.2003.