Webからの飲食店舗の評判情報抽出

高尾 美代子，酒井 浩之，増山 繁
豊橋技術科学大学 知識情報工学系
takao@la.cs.tut.ac.jp, sakai@tut.jp, masuyama@tut.jp

1　はじめに

外食する際に，Web上のグルメレビューサイトを参考にして外食先の飲食店舗の決定を行う場合は多い．しかしながら，Web上にあるYahoo!グルメ[*1]や，ぐるナビ[*2]などのグルメレビューサイトに掲載されている飲食店舗の中には，口コミ(以下，評判情報)が掲載されていない場合が約半数を占めており，それらに対しては評判情報を参考にすることが難しい(表1)．それに加え，Web上から飲食店舗の評判情報を検索しようとしても，飲食店舗の評判情報以外が記載されているWebページや，飲食店舗に関する広告のみが記載されているWebページが多いために，目的の店舗の評判情報が検索できない場合や，検索に時間が掛かってしまう場合がある．既存の類似システム[*3]があるが，検索対象がブログ情報と，投稿された口コミのみに限定されているため，目的の店舗の評判情報が必ずしも得られない場合や，評判情報でないブログが検索結果とされる場合がある．そこで本研究では，外食する際の飲食店舗選択支援の情報として，飲食店舗の評判情報を，Web上から自動的，かつ，正確に抽出することを目的とする．

表1　グルメサイトの飲食店舗に対するレビュー率(愛知県)

サイト名	店舗登録数(件)	レビュー数(件)	レビュー率(%)
Yahoo!グルメ	32,924	17,772	53.9
食べログ	34,937	14,092	40.3
ぐるなび	2,085	1,092	38.9

2　関連研究

矢野ら[矢野ら 04]は，まず，Web上から飲食店の店舗情報を取得し，その後，店舗情報から形態素解析によって評価文を選別することで評判情報を検索する手法を提案している．評価情報の精度は66.18 %となっている．矢野らの手法では，Web上から飲食店の店舗情報を取得する必要があるが，本提案手法では，店舗名を入力するのみで評判情報を抽出することができる．また，矢野らの手法は嗜好を考慮した評判情報検索手法となっているため，評価文を選別する際には「あっさり」や「こってり」などの味覚の評価のみとなっている．それに対し，本提案手法では，味覚の評価だけでなく，飲食店の雰囲気や従業員の態度まで，飲食店にまつわる幅広い評判を抽出することが可能となっている．

また，飲食店の情報を収集する研究として，山下ら[山下ら 07]，浪岡ら[浪岡ら 09]の提案手法があるが，自動的に情報を収集するものではない．前者は，オリジナルのブログインターフェースを用いてユーザからの飲食店舗の情報を蓄積し，飲食店舗の推薦を行うものであるため，本研究とは目的が異なっている．後者についても，健康管理のための飲食店舗情報検索を目的としており，本研究とは目的が異なっている．

これらに対し，本研究ではWeb全体から自動的に飲食店舗のあらゆる評判情報を抽出することを目的としている．

3　評判情報抽出手法

本研究では，Webページを飲食店舗の評判情報であるページと評判情報でないページ(以下，非評判情報)に分類を行うことで，飲食店舗の評判情報を抽出する手法を提案する．評判情報を部分的に含むページは，評判情報であるページとする．

3.1　前処理

提案手法を適用するための前処理として，共起語と共起表現の抽出を行う．本研究において，共起語とは，飲食店舗名の近辺(同一文中に限定しない，前方もしくは後方もしくは前後部分の2〜7語以内．)に出現する語(名詞以外の形態素)と定義する．共起表現は，評判情報であるWebページ・非評判情報であるWebページの双方に含まれる共起語のうち，評判情報であるWebページにおける出現確率と，非評判情報であるWebページにおける出現確率の比が2倍以上のものと定義する．ここで，共起表現の詳細な抽出方法を以下に示す．

[*1] Yahoo!グルメ (http://gourmet.yahoo.co.jp/restaurant/)
[*2] ぐるナビ (http://www.gnavi.co.jp/)
[*3] 食来エンジン Coocle , http://www.coocle.jp/

共起表現の抽出方法

Step 1.
正例・負例，それぞれについて，共起語の出現確率を求める．

Step 2.
正例・負例の双方に含まれる共起語のうち，評判情報である Web ページにおける出現確率と，非評判情報である Web ページにおける出現確率の比が 2 倍以上のものを共起表現として抽出する．■

ここで，共起表現の抽出対象となるデータ集合は図 1 の通りである．

Yahoo!検索 API を用いて取得した Web ページ 2,000 件のうち

- 正例：評判情報であるページ 100 件
- 負例：非評判情報であるページ 100 件

図 1　抽出対象となるデータ集合

共起表現の抽出例として，飲食店舗名が「すき家」の場合の共起表現を以下に示す．下線部のそれぞれが共起表現となる．

図 2 に例示したように，飲食店舗の評判情報には特徴

- すき家は やっぱり 美味しい
- 今日の お昼 は 久しぶり にすき家

図 2　共起表現の例

的な共起表現がある．そこで，共起表現が含まれていない Web ページを除去することによって，Web 検索時のノイズである広告やアフィリエイトの影響をできる限り小さくできると考え，前処理として共起語と共起表現の抽出を行った．

3.2　提案手法の概要

提案手法の概要について述べる．

評判情報である Web ページと非評判情報である Web ページの分類法

Step 1. **共起表現による分類**
分類したい Web ページ内に共起表現が含まれていれば Step 2 へ．共起表現が含まれていなければ非評判情報と判断する．

Step 2. **SVM による分類**
SVM^{Light}[*4]を用いて Web ページを分類．SVM で正例と判断された Web ページを評判情報と判断する．SVM の訓練データと素性については 3.3 節で述べる．■

[*4] SVM^{Light}(http://svmlight.joachims.org/)

3.3　SVM の訓練データと素性

本提案手法の Step 2 では，共起表現が含まれていた Web ページに対して，SVM による分類を行う．本提案手法では，SVM の学習に図 3 と図 4 に示す 2 種類の訓練データを用いた．それぞれを Yahoo!グルメ訓練データ，Web 訓練データとと呼ぶ．

- 正例：Yahoo!グルメから自動的に取得してきた評判情報であるテキストファイル 500 件
- 負例：Yahoo!検索 API を用いて取得した Web ページ 2,000 件のうち，非評判情報であるページ 500 件

図 3　訓練データ：Yahoo!グルメ訓練データ

Yahoo!検索 API を用いて取得した Web ページ 2,000 件のうち

- 正例：評判情報であるページ 100 件
- 負例：非評判情報であるページ 100 件

図 4　訓練データ：Web 訓練データ

また，SVM の素性は，酒井ら [酒井ら 06] の手法を用いて抽出した語を用いた．SVM で素性として用いられる品詞については，4.1.2 節に示す．ここでは，品詞が動詞と形容詞の場合の素性例を以下に示す．

食べ	思い	美味しい	行き	おいしい
落ち着い	すごく	多く	連れ	辛い

4　評価実験 1　効果的な共起表現の抽出範囲と素性判定実験

本研究の評価実験として，2 つの実験を行った．1 つ目として，効果的な共起表現の抽出範囲と素性判定実験について述べる．

4.1　実験内容

本手法において効果的な共起表現の抽出範囲と SVM の素性を判定するために，共起表現の抽出範囲パターンを共起パターン 1 から共起パターン 3，SVM の素性を抽出する際の品詞パターンを品詞パターン 1 から品詞パターン 8 まで変化させ，それぞれのパターンを，共起パターン 1 の場合に品詞パターン 1，共起パターン 1 の場合に品詞パターン 2…のように組み合わせ，Web ページを評判情報と非評判情報に分類する実験を行った．共起表現の抽出範囲パターンを 4.1.1 節に，SVM の素性を抽する際の品詞パターンを 4.1.2 節に示す．

4.1.1 共起表現の抽出範囲

共起表現を抽出する際の，抽出対象範囲のパターンを以下に示す.
ここで,
前方共起表現：飲食店舗名より前方に出現する共起表現
後方共起表現：飲食店舗名より後方に出現する共起表現
前後共起表現：飲食店舗の前方と後方に出現する共起表現
共起語数：飲食店舗名から共起表現までの語数
と定義する.

共起パターン 1　前方共起表現で，共起語数を 2〜7 に変化.
共起パターン 2　後方共起表現で，共起語数を 2〜7 に変化.
共起パターン 3　前後共起表現で，共起語数を 2〜7 に変化.

4.1.2 SVM に用いる素性パターン

SVM の素性を抽出する際の，品詞パターンを以下に示す.

品詞パターン 1　動詞と形容詞
品詞パターン 2　動詞と助動詞を合わせた複合語
品詞パターン 3　形容詞と助動詞を合わせた複合語
品詞パターン 4　形容詞と助詞と動詞を合わせた複合語
品詞パターン 5　名詞と助詞と形容詞を合わせた複合語
品詞パターン 6　名詞と助詞と動詞を合わせた複合語
品詞パターン 7　形態素バイグラム，又は，トライグラム
品詞パターン 8　単語バイグラム，又は，トライグラム

さらに，SVM の訓練データに用いた，訓練データパターンを以下に示す.

訓練データパターン 1　Yahoo!グルメ訓練データ
訓練データパターン 2　Web 訓練データ

4.2 実験結果

実験結果を表 2，表 3 に示す.

4.3 考察

表 2 より，共起表現の抽出範囲として，前方共起表現を用いる場合よりも，後ろ共起表現・前後共起表現を用いる方が精度が高くなった．これより，Web ページの飲食店舗の評判情報では，飲食店舗名の前部分よりも，後ろ部分に評判情報を記述することが多いことが分かる．また，表 3 に例示したように，共起語数に関しては，共起語数が少ないほどノイズが軽減され，精度が高くなった．しかしながら，抽出される評判情報のデータ数は，共起語数が少なくなるにつれ減少し，共起語数 3 以下では評判情報が抽出されない場合や，抽出されたデータが全て非評判情報である場合が多くなった．以上より，共起語数は 4 以上 6 以下のパターンを用いたほうが効果的と言える.

SVM の素性については，飲食店舗によって精度の最も高い素性パターンが異なる結果となった．しかし，パターン 1 の動詞と形容詞のみの場合に比べ，パターン 7 の形態素バイグラムを用いる場合や，パターン 8 の語バイグラムを用いる場合の方が精度が高くなりやすい傾向が見られた．以上より，SVM の素性品詞には，動詞と形容詞のみの場合よりも，バイグラム，又は，トライグラムを用いた方が効果的と言える.

5 評価実験 2　Web ページの部分抽出の有効性判定実験

本研究の 2 つ目の評価実験として，Web ページからの部分抽出の有効性判定実験について述べる.

5.1 部分抽出

本研究において，部分抽出とは，飲食店舗名が含まれている最小のタグ部分 (tr,td,p,div) を抽出することと定義する．飲食店舗名が「なご壱」である場合の部分抽出例を以下に示す.

部分抽出前

```
<html> <body>
<div class=”title”> 今日のご飯</div>
<div class =”body”>今日はなご壱に行ってきました！めっちゃ美味しい！天むす最高！！</div>
</body> </html>
```

部分抽出後

```
<div class =”body”>今日はなご壱に行ってきました！めっちゃ美味しい！天むす最高！！</div>
```

5.2 実験内容

Web ページの部分抽出の有効性を判定するために，Web ページから部分抽出を行った場合・部分抽出を行わない場合それぞれについて，共起表現・素性抽出時の抽出パターンを組み合わせ，Web ページを評判情報と非評判情報に分類する実験を行った.

5.3 実験結果

実験結果を表 4，表 5 に示す.

5.4 考察

表 4 より，部分抽出を行わない場合に比べ，部分抽出を行う場合の方が，精度が高くなった．これは，部分抽出を行ったことにより，広告やアフィリエイトなどのノイズを除去できたためである．また，表 5 より，部分抽出を行った場合でも，評判情報の抽出件数は数件程度しか減少せず，抽出件数が増加する場合も見られた．以上より，部分抽出は評判情報抽出において効果的と言える.

6 おわりに

本研究では，共起表現による分類と SVM による分類を併用することで，Web からの飲食店舗の評判情報の抽出を試みた．その結果，それぞれの分類を単独使用するよりも抽出精度は向上し，飲食店舗 5 件での精度は，

表 2　実験結果 1-1：全店舗の平均値

共起表現	共起語数	SVM の品詞パターン	SVM の訓練データパターン	精度
なし	なし	パターン 1(動詞・形容詞)	パターン 1(Yahoo!グルメ訓練データ)	52.2 %
前方共起	語数 6	パターン 7(形態素バイグラム)	パターン 2(Web 訓練データ)	57.14 %
後方共起	語数 6	パターン 7(形態素バイグラム)	パターン 2(Web 訓練データ)	83.3 %
前後共起	語数 6	パターン 7(形態素バイグラム)	パターン 2(Web 訓練データ)	60 %

表 3　実験結果 1-2：店舗名「あんかけ亭」についての結果

共起表現	共起語数	SVM の品詞パターン	SVM の訓練データパターン	精度	評判情報の抽出件数
後方共起	語数 6	パターン 7(形態素バイグラム)	パターン 2(Web 訓練データ)	80.0 %	8 件
後方共起	語数 4	パターン 7(形態素バイグラム)	パターン 2(Web 訓練データ)	83.3 %	5 件
後方共起	語数 2	パターン 7(形態素バイグラム)	パターン 2(Web 訓練データ)	100 %	1 件

表 4　実験結果 2-1：全店舗の平均値

部分抽出	共起表現	共起語数	SVM の素性パターン	SVM の訓練データパターン	精度
なし	後ろ共起	語数 5	パターン 8(語トライグラム)	パターン 1(Yahoo!グルメ訓練データ)	72.2 %
あり	後ろ共起	語数 5	パターン 8(語トライグラム)	パターン 1(Yahoo!グルメ訓練データ)	100 %

表 5　実験結果 2-2：店舗名「あんかけ亭」についての結果

部分抽出	共起表現	共起語数	SVM の素性パターン	SVM の訓練データパターン	精度	評判情報の抽出件数
なし	前後共起	語数 5	パターン 8 (語バイグラム)	パターン 1(Yahoo!グルメ訓練データ)	88.9 %	8 件
あり	前後共起	語数 5	パターン 8 (語バイグラム)	パターン 1(Yahoo!グルメ訓練データ)	100 %	11 件
なし	前後共起	語数 5	パターン 8 (語トライグラム)	パターン 2(Web 訓練データ)	85.7 %	12 件
あり	前後共起	語数 5	パターン 8 (語トライグラム)	パターン 2(Web 訓練データ)	100 %	10 件

平均 88.8 %程度となった．さらに，それぞれの分類に加え，部分抽出を行うことで，抽出精度をより向上させることができ，平均の精度が 100 %であるパターンも多く見られた．しかしながら，本研究では評判情報を正確に抽出することを目的としたため，精度のみを考慮し，再現率を考慮していない．従って，今後の課題として，再現率を考慮した改良手法の提案を行いたい．

参考文献

[酒井ら 06] 酒井 浩之, 梅村 祥之, 増山 繁, 交通事故事例に含まれる事故原因表現の新聞記事からの抽出, 自然言語処理, pp.99-123, 2006

[矢野ら 04] 矢野 宏実, 目良 和也, 相沢 輝昭, 嗜好を考慮した評判情報検索手法，情報処理学会，pp.165-170, 2004

[山下ら 07] 山下 晃弘，川村 秀憲，山本 雅人，大内 東，ブログによる情報収集と推薦技術を用いた飲食店情報サイトの構築，情報処理学会，pp.133-138, 2007

[浪岡ら 09] 浪岡 潤，澤井 政宏，久保 洋，RoR を用いた健康管理のための飲食店情報検索システムの構築に関する研究，SVBL 年報，pp.81-82, 2009