グルメサイトに対する対応分析を用いたレビュー分析

2013SE210 瀧将史　2013SE228 都築彰太
指導教員：河野浩之

1　はじめに

グルメサイトは 1996 年に誕生し，近年では特にグルメサイトを用いて自分にあった飲食店を探す人々が増加傾向にある．各グルメサイトの月間レビュー数に着目してみると食べログが約 16 億 4000 万，ぐるなびが約 11 億と莫大の量のレビューが投稿されていることがわかる．

本研究の流れとしてまずグルメサイトからテキスト（レビューや口コミなど）を抽出する．抽出する手段として API を使用し，テキストを集める．API はぐるなび API を使用する．そしてその収集したテキストを KHcoder を用いて抽出語リストを作成する．その結果を用いてクロス集計表を作成する．その結果より対応分析図を作成しお店への評価，考察を行い，新たな発見，意外な一面の発見していく．

本論文は全 5 章で構成され，第 2 章はデータマイニングを用いた推薦技術の先行研究，テキストマイニングに関する先行研究を比較していく．第 3 章では前章でとりあげた先行研究の課題に対しての解決方法とそれを解決するための提案をしていく．第 4 章ではアーキテクチャに基づき実験，評価・考察を行う．最後に第 5 章をむすびとし，今後の課題について述べる．

2　データマイニングを用いた推薦技術の先行研究

本章では知識源としてのマイクロブログを活用した先行研究とテキストマイニングに関する先行研究を紹介し, 先行研究の比較を行っていく．

2.1　知識源としてのマイクロブログを活用した先行研究

参考文献 [1] の研究ではブログデータをテキストマイニングし, その中に含まれる情報から日本酒の美味しい店を発掘することに成功した.Twitter 社が提供している Search API を利用して, 2012 年 12 月 21 日から 2013 年 6 月 12 日までの半年程度の期間において日本酒に関する 4,123,950 件のデータが分析対象となった. この約 412 万件のうち、日本酒を含むものは約 77 万件であった. 収集したデータを IBM Watson Content Analytics Version 3.5 に投入し, 日本酒の美味しい店を探すという観点から, 日本酒の銘柄、店名の言及、好不評を示す評価表現に関する調査を行った．10 以上の地域において日本酒の美味しそうな店の情報をツイートから取得し, 店を選んで実地調査を行った. 結果として, 有効性を評価するため, 参加者に満足度と再訪希望度を回答してもらった. 本手法で選択して実地調査した店は 12 店であるが、そのうち少なくとも 4 店は参加者が実際に再訪しており, 満足度の高い店を選ぶことができたと考えられる．

しかし, 膨大な量のつぶやきから調査, 評価をした為に, この研究から得られた技術的課題として, 自然言語処理の観点において, 語義曖昧性解消の必要があると分かった. 店名には「彩」や「南」など汎用性の高い文字が使われていることが多いためである．

2.2　テキストマイニングに関する先行研究

他には 参考文献 [2] の「テキストマイニングでご当地ラーメンを特徴ごとに分類」という記事がある．ラーメンは地域によって様々な特色を持ち, 多様な進化を遂げている．

そこで, ご当地ラーメンがどんな特徴を持っていて, 主流がどのようなラーメンなのか把握した. そして, 代表的なラーメンがどのようなニーズの人におすすめかも調べた. R 言語とツイッターを使用し, 各ご当地ラーメンに関するツイートを 1000 個づつ自動収集した. そこで得られたツイッターテキストを RMeCab を使用してデータマイニングした．その結果, キーワードの頻度を比較の指標にしてクロス集計表を作成した．視覚的にわかりやすい形にするため, コレスポンデンス分析で視覚化を行った．コレスポンデンス分析は対応分析とも呼ばれ, 列項目と行項目の相関が最高になるように両方ともを並べ替える事である. そして, クラスタリングで似た者同士をまとめてグループに分類した．

例えば九州のラーメン, 熊本と博多, 鹿児島はとても近い位置にあったが, これは豚骨との相関が他のラーメンと比べてとても高いことを示す. 逆に豚骨から最も遠い札幌ラーメンは豚骨との相関がとても低いということである．

しかし, この記事内で紹介されたツールでは最低限の分類分けだけに終わり，分析しきれなかった特徴を持つラーメンが他にも多数存在する．

2.3　先行研究の比較

前で述べた先行研究を比較する．日本酒の研究ではラーメンの研究に比べ，より満足度の高い店を発掘することが出来るということが分かった． ラーメンの研究では，ユーザーの趣味嗜好などの情報をもとに店を探すということは行わなかった． 単純なジャンル分け， 視覚化のみを行ったため，細かな情報を収集することが出来なかった．

日本酒の研究では， ツイッター上から日本酒の美味しいお店発掘することに成功した． 問題点により細かな情報を元に店を探す場合， 非常に困難であるということだ. 人手がほとんどかからず細かな特徴から店を探し出せるようになればさらに便利になる．次のページの表 1 に先行研究を比較したものを示す．

表 1　先行研究の比較

先行研究	結果	課題
データマイニングに関する先行研究	満足度の高い店を推薦	詳細な検索不可. 人手がかかる
テキストマイニングに関する先行研究	単純なジャンル分け, 視覚化	大まかなジャンル分け

3　感性分析技術を用いたレビュー評価のアーキテクチャ

本章では我々の提案するアーキテクチャを基にどのような流れで研究を行っていく仮説していく.

3.1　提案

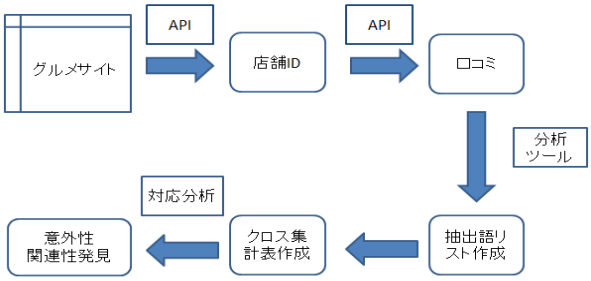

図 1　レビュー評価のアーキテクチャ

これから本研究のアーキテクチャについて記す．図 1 にアーキテクチャを示す. 先行研究は満足度の高いお店の推薦や，単純なジャンル分けを可能にした.

本研究では，4 種類のとんかつチェーン店に焦点を当ててそれぞれのレストランの評価を視覚的かつ数値を用いて行う．ぐるなび API を用いてレストランの口コミを抽出し，KHcoder で抽出語リストを作成し，クロス集計表にキーワードをまとめ，対応分析のグラフを作成する．最後にどういった特徴を持った店か，どのような意外性があるかグラフの座標をもとに考察していく.

3.2　ぐるなび API を用いた口コミ収集

本研究ではグルメサイトよりぐるなび API を用いて口コミ抽出をする．そこでどのグルメサイトを使用するか検討する．一つ目が「食べログ」である．月間ユーザー数が約 7265 万人（2016 年 6 月）である．日本国内では最大の勢力を誇っている．二つ目は「ぐるなび」である．月間ユーザー数は約 5200 万人と「食べログ」に劣るが，「ぐるなび」は API が存在する．「食べログ」「ぐるなび」ともに匿名投稿である．匿名であるということはユーザーの本音が聞きやすいメリットがあると考える.

続いてグルメサイトの三つ目に「Yelp」がある．月間ユーザー数が他より多く，これまで進出した国より日本でのユーザー数の伸びが多いことが特徴となっている．またレビューをする際実名投稿が原則となっていることから内容に責任が生じ，より確かなレビューや情報を入手出来ると考えられる．また API も存在する.

以上を踏まえて匿名投稿であり，三つの中で 2 番目月間ユーザー数が多く API が存在する「ぐるなび」よりテキストを抽出することにした．以下の表 2 は三つのグルメサイトを比較したものである.

表 2　グルメサイトの比較

グルメサイト	月間レビュー数	月間ユーザー数	投稿	API 有無
食べログ	約 16 億 4000 万	約 7265 万人	匿名	無
ぐるなび	約 11 億	約 5200 万人	匿名	有
Yelp	-	約 1 億 2000 万人	実名	有

以上を踏まえて API が存在し, 匿名投稿が可能なぐるなびを使用する. 投稿が匿名であると本音の発言を伺えより精度の高い評価を得られると考える. 続いてぐるなび API について紹介する．ぐるなび API は公式サイトに 10 種類用意されており，幅広い活用が期待される

3.3　抽出語リスト

本研究ではクロス集計表を作成するために抽出語リストを作成する. 本研究ではクロス集計表を作成する際に Excel を使用するので連動可能である「KHcoder」を採用する. 次に簡単な操作方法を説明する.KHcoder を起動し, 読み込みたいテキストを開く.「前処理の実行」をしたのち,「ツール」を選択し,「抽出語リスト」で Excel と連動して抽出語の出現頻度を確認することが出来る.3.2 で抽出した口コミ内には無数のテキストが存在しているが抽出語リストの作成でそれらを品詞別にわけ, 出現頻度の高い順にリストにまとめる.

3.4　クロス集計表の作成

本節では, クロス集計表の重要性について説明する. ぐるなび API より抽出した口コミを 3.5 節の対応分析で活用するためにクロス集計表が必要となる．今回我々が作成するクロス集計表は，4 種類のとんかつレストランのキーワードが何度出現したかを示すものである．3.2 節で品詞別に分けた抽出語リストの名詞部分に着目し，出現頻度が高く，お店を象徴する言葉を選択しクロス集計表を作成する．クロス集計表で着目した言葉以外は 3.5 節の対応分析に反映されない.

対応分析時に重要度の低い言葉までプロットしてしまうと，可視化を困難にしてしまうのでクロス集計表を作成する.

3.5　対応分析

ここでは対応分析について説明する．抽出語を用いて 2 次元のプロットで表示される．集計済みのクロス集計表を用いて，行の要素と列の要素を使い，それらの相関関係が最大になるように数量化する．そしてその行の要素と，列の要素を散布図に表現するものである．本研究ではレストラン名とキーワードを同時に散布図上にプロットできるた

め直感的に相関がわかる特徴がある．

4　ぐるなびを用いた対応分析の検証

4 章では実験について論じていく．4.1 ではぐるなび API を用いたテキスト抽出を行う．4.2 では抽出語リストの作成を行う．4.3 ではクロス集計表の作成を行う．4.4 では前節で作成したクロス集計表を用いて対応分析図を作成し分析する．

4.1　ぐるなび API よりテキスト抽出

本節では，とんかつレストランである「矢場とん」,「浜勝」,「とん Q」,「かつや」という 4 種類のレストランに対する口コミを抽出可能な最大 50 件表示するという条件を加えた上でプログラムを html に保存し, ファイルを開きインターネット上で動かす．ファイルをインターネット上で開くと, 次にアクセスキーを求められる. ぐるなびよりアクセスキーを事前に入手しておく必要がある. これらのレストランはチェーン展開する店のため, 複数の店舗から口コミを得る方が口コミの分析をする上でより効果的であると考えた．そのためぐるなびが提供するレストラン検索 API を用いて複数の店舗 ID を取得した．API のプログラムが書かれた html ファイルにをインターネット上で開くと, アクセスキーを求められる．そのため事前にぐるなびよりアクセスキーを入手しておく必要がある．

図 2 にプログラムの一部を示す．図 2 の 5 行目の name というパラメータでは店舗 ID を取得したい店名を入力する．今回は「矢場とん」と入力する．今回「矢場とん」に対する店舗 ID は 15 件見つかった. その後, 取得した店舗 ID を「応援口コミ API」のプログラムに打ち込む. 図 3 に「応援口コミ API」の一部を示す．

```
var url = 'http://api.gnavi.co.jp/RestSearchAPI/20150630/?callback=?';
var params = {
  keyid: '',
  format: 'json',
  name: '矢場とん',
  hit_per_page: '50'
   };
```

図 2　レストラン検索 API のプログラム例

```
var url = 'http://api.gnavi.co.jp/PhotoSearchAPI/20150630/?callback=?';
var params = {

  keyid: '',
  format: 'json',
  hit_per_page: '50',
  shop_id:
  'fa84619,6955363,5630042,5422555,7218285,6346314,7340967,5656619,6454197,6460485'
  };
```

図 3　応援口コミ API のプログラム例

図 3 の 6 行目のパラメータ「shop id」では店舗 ID を指定できる．今回はここに「レストラン検索 API」のプログラムから得られた店舗 ID を打ち込む．店舗 ID は最大で 10 件しか入力できないため,「レストラン検索 API」で見つかった 15 件の店舗 ID は 2 つのプログラムに分けて使用することにした．図 3 の 5 行目のパラメータ「hit per page」ではヒット件数を指定できる．ヒット件数とは，1 回のリクエストで得る最大投稿件数のことを意味する．デフォルトは 15 件，上限は 50 件であり，本研究ではなるべく多くのデータを取得したいため，最大の 50 件に設定する．

4.2　抽出後リストの作成

次に前節で抽出したものを txt ファイルで保存をする．KH coder を起動し，[プロジェクト] の [新規] で分析対象のファイルを選択する．続いて [前処理] の [前処理の実行] をし，処理を確認する． [ツール] の [抽出語] の [抽出語リスト] を選択し，エクセルと連動しどの言葉がいくつあるかを確認する．

4.3　クロス集計表の作成

4.1 節ではとんかつに関する口コミのみを抽出した．4.3 節では 4.2 節のリストを使用し抽出できた口コミ内の「味噌」や「ミソ」など, 表記揺れする言葉を一つの語句に統一する．例えば「ミソ」を「味噌」とする. そして最終的にそれぞれのとんかつ屋に対してキーワードであると思われる言葉を Excel を用いてクロス集計表にまとめる．キーワードは名詞で出現頻度の高い言葉とする．表 3 がクロス集計表である．今回は 4 種類のとんかつレストランから, 特徴語を合計 9 個設定し, 出現回数を数えクロス集計表にまとめた．今回キーワードとしてみなした言葉は「味噌」,「ソース」,「ボリューム」,「キャベツ」,「チキン」,「野菜」, 「カレー」,「とん汁」,「味噌汁」である．表 3 内の数値はそれぞれのお店に対しキーワードが何度出現したかを示している．

表 3　レストランとキーワードの座標

	味噌	ソース	ボリューム	キャベツ	チキン	野菜	カレー	とん汁	味噌汁
矢場とん	32	15	14	5	0	10	0	0	1
浜勝	0	7	11	22	10	10	3	8	25
かつや	0	2	1	2	0	3	4	2	1
とん Q	0	1	1	2	0	2	0	2	0

4.4　対応分析の作成と分析結果

今節では 4.3 で作成したクロス集計表を基に対応分析図を作成し，各店舗の分析行っていく．対応分析ではクロス集計表を基において行項目と列項目の相関が最大になるように双方を並び替えることである．本研究では KH coder 付属の R を用いて 4 店舗について対応分析を行う．作図に至るまでの過程を説明する．

まず matrix 関数を用いて行列の要素をベクトルで用意

し，行列に変換する．次に行（Row）と列（Column）を打ち込む．そして library 関数を用いて MASS というパッケージを呼び出す．Corresp 関数を用いて対応分析を行う．ここで nf は求める軸の数を指定する引数である．返される結果は，正準相関係数，行の得点，列の得点である．対応分析では，計算された軸の行・列に対応する値をそれぞれ行，列の得点と呼ぶ. 本研究で作成した対応分析は図 4 のとおりである．

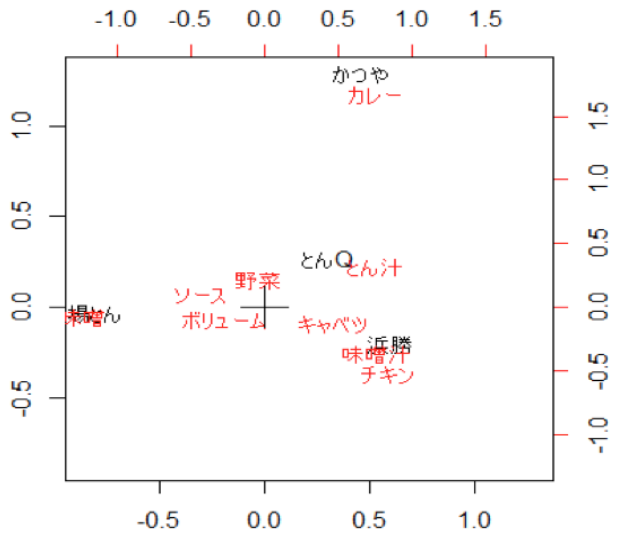

図 4　とんかつレストランに関する対応分析

図 4 では店舗名 (矢場とん, 浜勝, かつや, とん Q) の成分が左側 (y 軸) と下側 (x 軸) の目盛り，キーワード (味噌, ボリューム, ソース, カレー, とん汁, 味噌汁, キャベツ, チキン, 野菜) の成分が右側 (y 軸) と上側 (x 軸) の目盛りで表されている. それぞれの x 軸が第 1 主成分分析，y 軸が第 2 主成分分析である．各店舗，キーワードがプロットされている位置は，x=第 1 主成分分析×第 1 主成分分析の正準相関係数，y=第 2 主成分分析×第 2 主成分分析の正準相関係数である．例えば矢場とんの座標は (-1.2381289 × 0.7015185，-0.06902368 × 0.3880429）で表される．

4.5　対応分析考察

実験結果よりわかったことを考察していく． 今回は「味噌」,「ボリューム」,「ソース」,「カレー」,「とん汁」,「味噌汁」の 6 つのキーワードに着目する．まず「味噌」については明らかに「矢場とん」との距離が約 0.372 と最も近いことがわかる．「ボリューム」に関しては「矢場とん」,「浜勝」,「とん Q」の 3 つのレストランで距離の誤差が約 0.239 以内となり，この 3 つのレストランでは差はほとんどなかった．「ソース」に関して矢場とんが約 0.455 と最も近い．「カレー」に関しては丼系のかつやが約 0.475 と最も近い．「とん汁」に関しては 0.454 ととん Q が最も近く，地域ごとに具材を変えていることが大きいと考えられる．最後に「味噌汁」に関しては浜勝が 0.234 と最も近く，味噌汁がおかわり自由であることが要因と考えられる．

今回の結果でも注目したのが矢場とんと言えば「味噌」を連想しがちだが，今回の分析結果から「ソース」に関しても関係性が強いことが判明した．次のページの図 10 にとんかつレストランとキーワードの座標と図 11 にキーワードとの距離を示す．

表 4　レストランとキーワードの座標

レストラン	X 軸	Y 軸
矢場とん	-0.869	-0.268
浜勝	0.597	-0.203
かつや	0.474	1.289
とん Q	0.301	0.272
キーワード	X 軸	Y 軸
味噌	-1.238	-0.069
ソース	-0.433	0.106
ボリューム	-0.254	-0.100
キャベツ	0.476	-0.122
チキン	0.852	-0.522
野菜	-0.039	0.218
カレー	0.751	1.675
とん汁	0.752	0.322
味噌汁	0.768	-0.363

表 5　レストランとキーワードの二点間の距離

	味噌	ボリューム	ソース	カレー	とん汁	味噌汁
矢場とん	0.372	0.619	0.455	2.349	1.658	1.67
浜勝	1.84	0.858	1.076	1.883	0.547	0.234
かつや	2.186	1.568	1.491	0.475	1.006	1.678
とん Q	1.577	0.668	0.753	1.473	0.454	0.788

5　むすび

本研究ではグルメサイト（ぐるなび）からぐるなび API を用いて口コミを抽出した．得られたテキストを KHcoder を用いて抽出語リストを作成し，それを基に対応分析に必要なクロス集計表を作成した．

対応分析は R を用いて作成し，各点の座標とその距離を求めることができた．それを基に店舗の特徴や相関の強さを数値的に，視覚的に確認することに成功した．

課題としては本研究で抽出した口コミは多いものとは言えず，もっとより多くの口コミを収集出来れば高精度な分析が可能であると考えられる．

参考文献

[1] 那須川哲哉, 吉田一星, 西山莉沙, 吉川克正, 伊川洋平, 大野正樹, 村上明子, “大量のつぶやきから日本酒のおいしいお店を発掘する”, 言語処理学会第 21 回年次発表論文集, pp.820-823, 2015.

[2] Hatena Blog テキストマイニングでご当地ラーメンを特徴ごとに分類してご紹介. http://yeoman.hatenablog.com/ (Dec,2016,Access).

[3] ぐるなび WEB サービス for Developers. http://api.gnavi.co.jp/api/ (Dec,2016,Access).