テキストマイニングによる食レポ口コミからのユーザー嗜好モデリング

システム科学技術学部　経営システム工学科
2年　吉田　妃那
2年　石黒　菫
2年　小島　幸奈
2年　酒井　あかり
指導教員　システム科学技術学部　経営システム工学科
教授　上原　宏

1.本研究の目的

全国の飲食店の情報と共にユーザーの口コミが掲載されているサイト「食べログ」では口コミ評価順や口コミ件数順などのソート機能や料理ジャンルや地域を絞る情報フィルタリング機能が実装されており、ユーザーが求めている店舗を比較的見つけやすくなっている。しかし、食べログのサイトには季節ごとに店舗を検索する機能がない。本研究では既存の視点とは異なる四季という視点から旬にこだわっている飲食店や季節性の高い食べ物（鍋、おでんなど）が楽しめる店舗を提案する方法の有用性を示すことを目的としている。また、洋食と和食で季節を特徴づける単語に差があるのかという検証も本研究の目的である。もし、和食と洋食で季節感を感じさせる単語が一致し、似たような分析結果が出れば、以前、和食の飲食店を利用したユーザーに同じ旬の食材を使った料理や似たような季節性の高い食べ物を提供する洋食の飲食店も提案することができるだろう。

2.方法

python 言語を用いて、「食べログ」からデータを収集、分析した。はじめに、自動的に Web ページを巡回し情報やデータを収集するクローラーと呼ばれるものを用いて「食べログ」から口コミのデータを収集した。本研究では、洋食の店舗、和食の店舗それぞれ 6 店舗から収集している（表 1)。これらの店舗は偏りをなくすために、和食は、割烹料理、京料理、懐石料理から、洋食については、イタリアン、フレンチ、スペイン料理から、といったように多ジャンルから選び、地域についても偏らないように可能な限り異なる地域の店舗を選んだ。そこから、口コミ件数が 200 件以上あり、尚且つ、桜や雪といった四季を感じさせるような単語や食材の名前が含まれているなど口コミの内容が充実している店舗であるという基準を設けてさらに抜粋した。収集した後は形態素解析を行い、juman の意味解析機能を用いて、特定のカテゴリもしくはドメインなどに該当する形態素を抽出した。また、ここで漢字表記やカタカナ表記など表記の仕方は違うものの同じ意味を示す単語や同義語については統一表現に揃えておく。更にそこから、各単語はその周辺の語彙と何らかの関係性を持つという仮定に基づいて単語の特徴をベクトル化することのできる

word2vec というモデルを使い、季節性のある単語、かつ、ベクトルの距離が近い単語のみを抽出した。また、このときのベクトルの成分は tf-idf の値を用いている。tf-idf とは、文書を特徴づける単語を抽出するために単語の重要度を測るものであり、ほかの文章中にはあまり出てこないが、特定の文章中にだけ出現頻度が高い単語ほど値が高くなる。抽出した単語から辞書とコーパスを生成し、kmeans というクラスタリング手法で主成分分析を行い、ベクトルとして二次元平面に可視化し、また、**各店の春夏秋冬がどのような特徴語彙を有していてそれらがどのクラスタとしては所属するかを対照するために、分類されたクラスタ、店舗名、投稿された季節、口コミに含まれる語彙の** tf-idf を csv **ファイルに書き出すようにした。**なお、サンプル間の距離測定方法についてはコサイン距離を適用した。コサイン距離とは、ベクトル空間モデルにおいて、ベクトル同士の成す角度の近さを表現する類似度計算手法であり、文書同士を比較する際に用いられる。

洋食				和食			
店舗	ジャンル	地域	件数	店舗	ジャンル	地域	件数
ア・ニュ ルトゥルヴェ・ヴー	フレンチ	東京都	436	茶茶白雨	京料理	東京都	284
レフェルヴェソンス		東京都	491	草喰 なかひがし		京都府	336
クッチーナ イタリアーナ ガ	イタリアン	愛知県	237	銀座 しまだ	割烹料理	東京都	349
カーザ ヴィニタリア		東京都	354	島之内 一陽		大阪府	201
SALONE 2007	イタリアン・フレンチ	神奈川県	415	虎白	懐石料理	東京都	242
カセント	スペイン料理	兵庫県	299	稚加榮 本店		福岡県	481
計			2232	計			1893

表 1　分析に用いた口コミの店舗

3.結果

洋食 6 店舗から件の口コミを収集し分析したところ次のような結果になった(図 1)。なお、k-means クラスタ数は 11、主成分平面は第 3 主成分、第 4 主成分を軸として分析を行った。グラフで色分けされている”春” ”夏” ”秋” ”冬”とは、クチコミが投稿された時期のことであり、ここでは 12~2 月を”冬”、3～5 月を”春”、6～8 月を”夏”、9～11 月を”秋”として分類している(グラフ 1、グラフ 2)。グラフ 1 を見ると、クラスタ 2、クラスタ 10 は”春”、クラスタ 8 は”夏”と”秋”、クラスタ 3、クラスタ 4、クラスタ 5 は”秋”と”冬”、クラスタ 6 は”春”と”冬”に投稿された口コミが多いことが分かる。実際に、クラスタの tf-idf の値の高い単語を見ると、クラスタ 2 はアスパラガスや空豆や桜海老、クラスタ 10 はブリや桜や春菊、クラスタ 8 はアユやスイカやキュウリ、クラスタ 3 はセロリ、4 はブリや大根、5 は栗、6 は菜の花やカリフラワーなど口コミが投稿された時期ということ一致している旬の食材の名前が多く挙がっていた。

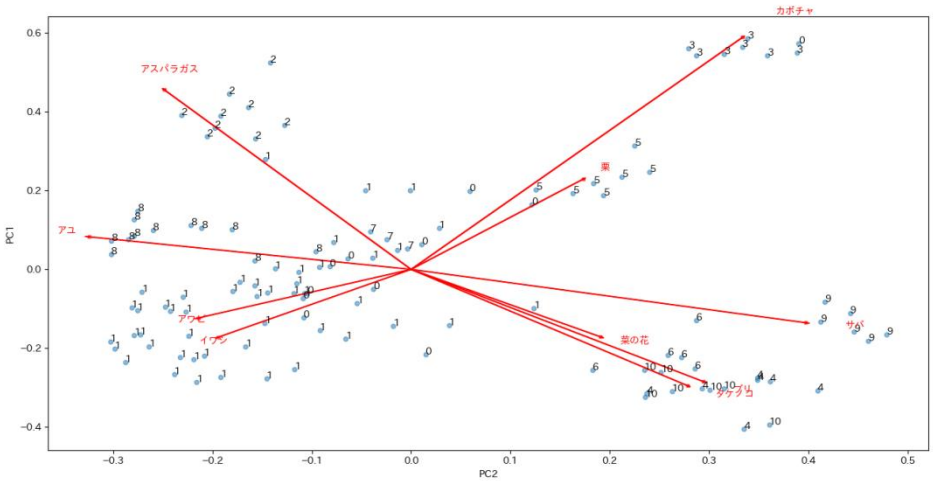
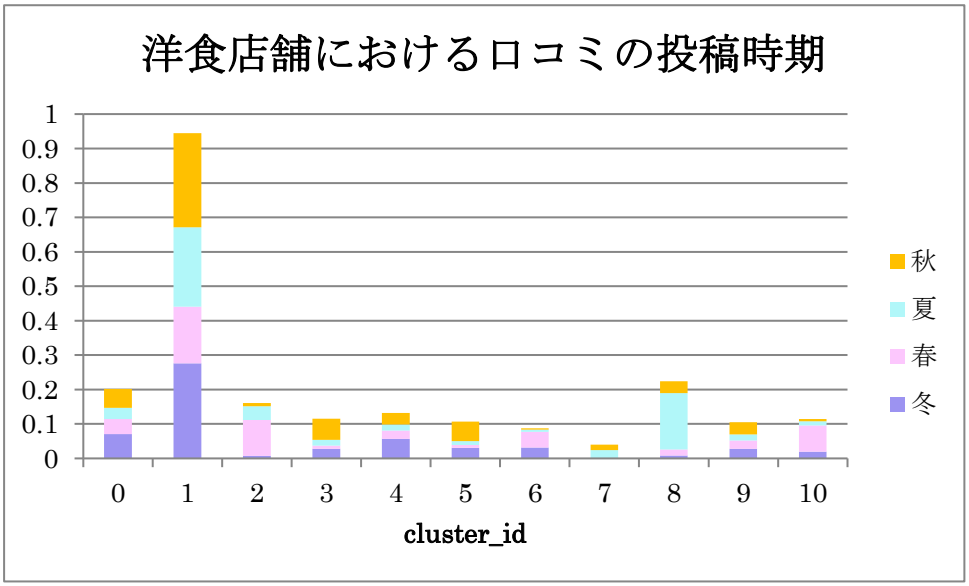


図 1　洋食の店舗における主成分分析



グラフ 1　洋食店舗における口コミの投稿時期

同様に、和食 6 店舗から件の口コミを収集し分析したところ次のような結果になった（図 2)。なお、k-means クラスタ数及び主成分は洋食と同一にして分析を行った。最終的に分析に使われる口コミ件数は 2363 件。グラフ 2 を見ると、クラスタ 7、クラスタ 8 は”春”と”冬”、クラスタ 6 は”夏”、クラスタ 1 は”秋”、クラスタ 9 は”秋”と”冬”に投稿された口コミが多いことが分かる。実際に、クラスタの tf-idf の値の高い単語を見ると、クラスタ 7 は菜の花やミカン、クラスタ 8 はハマグリや菜の花や桜、クラスタ 6 はトウモロコシやアユ、クラスタ 1 はさんま、クラスタ 9 は栗やホウレンソウなど口コミが投稿された時期と一致している旬の食材の名前が多く挙がっていた。

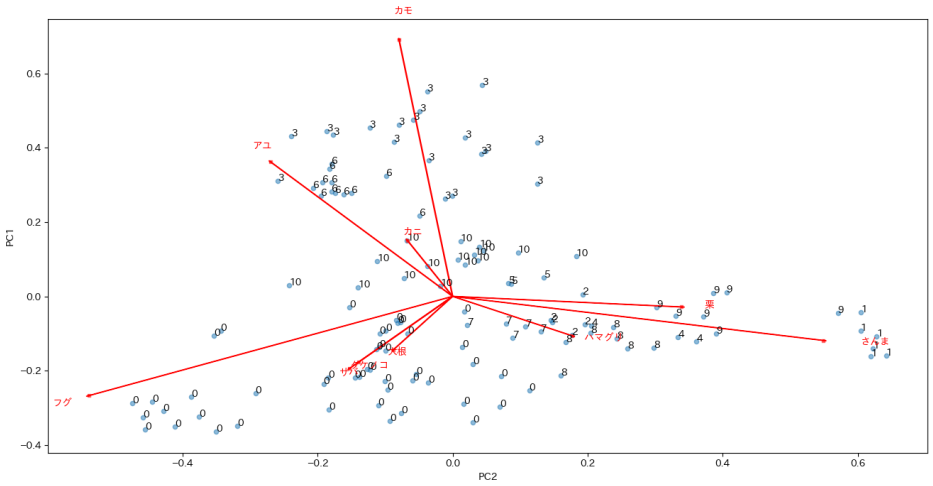
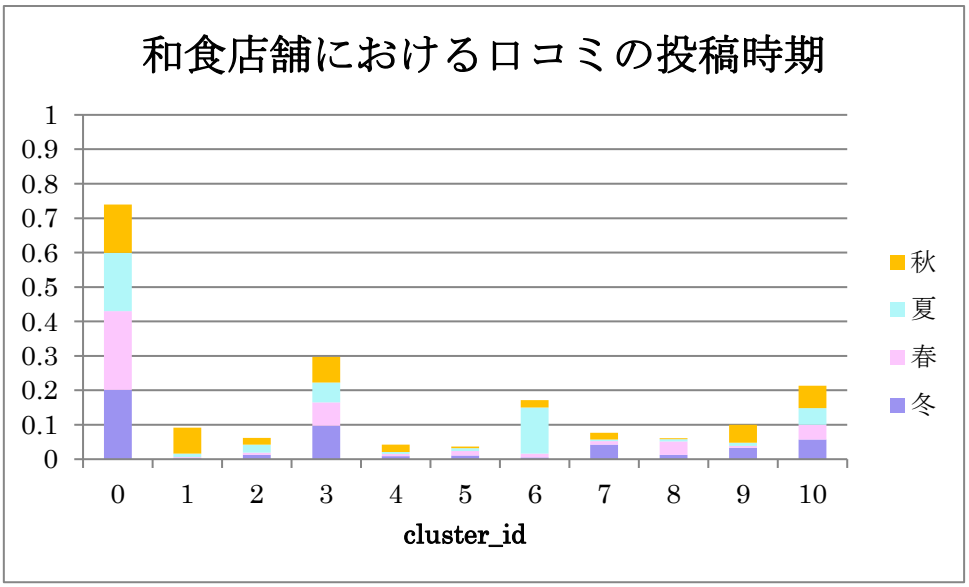


図２　和食の店舗における主成分分析



グラフ２　和食店舗における口コミの投稿時期

4.考察

クラスタを特徴づける語彙の季節性と口コミが投稿された季節が一致しているときそのクラスタに多く当てはまる店舗は季節性の高い料理がある、旬の食材にこだわっている店舗であるといえるだろう。その一方で、口コミが投稿された時期がある季節に偏っていないことから、季節という基準では上手く分類されなかったクラスタがある。近年、栽培環境や生産技術の向上で通年食べられる食べ物が増えてきたことが影響している可能性がある。和食と洋食では分析結果を比較すると、季節感を感じさせる単語にはそれほど差はなかったものの主成分分析のグラフについては一致したとは言い切れないため、和食の飲食店を利用したことのあるユーザーに似たような洋食も提案することは難しいということが明らかになった。