6S-03

楽天市場レビューデータにおけるRを用いたデータマイニングの実践

大楠 拓也† 徐 海燕†
福岡工業大学大学院 工学研究科 情報工学専攻†

1. はじめに

近年、蓄積された大量のデータからビジネスに活用できるデータにするために、データマイニングに関する研究が盛んに行われている。

本研究では、楽天株式会社が公開している楽天データセット[1] [3]内の楽天市場データ内のレビューデータをジャンル ID、性別や年齢別ごとに集計した上で統計解析向けの R 言語を用いて分析し、ジャンル別、性別や年齢別の性質や嗜好情報を導き出す。また、R 言語で形態素解析を行える RMeCab を用いた解析も行う。

2. 楽天市場レビューデータに対する集計

楽天市場データにおけるレビューデータ内には 2010 年 1 月から 2012 年 12 月までのレビューが登録されており、約 6000 万件位存在する。商品ジャンルとしてはルートからまず 30 数個のジャンルに分け、さらに 6 層まで展開される。

次章の分析を行うために本研究ではこの楽天データセットから MySQL を用いて必要となるデータを抽出し、avg 関数や count 関数を用いて集計を行っている。抽出データ列は性別レビュー数や平均ポイント、平均年齢、平均価格がある。また時系列分析のためにレビュー登録日時を各ジャンル、年月別に抽出し集計したデータも抽出している。なお、性別や年齢の欠損値に対しては、除去した上でデータを抽出している。

3. 楽天市場抽出レビューデータに対する分析

本研究では、パソコン・周辺機器/外付けドライブ・ストレージの下に位置するフラッシュメモリと、キッズ・ベビー・マタニテ/バッグ・ランドセルの下に位置するランドセルという 2 つのジャンルに対する分析を行う。

3.1 plot 関数分析

フラッシュメモリにおける抽出データに対して R 言語の plot 関数を用いて分析を行った。得られた結果として記憶容量が多いほどリピート率と平均ポイントが上昇する結果が得られた。この傾向はメモリーカードでも見られた。

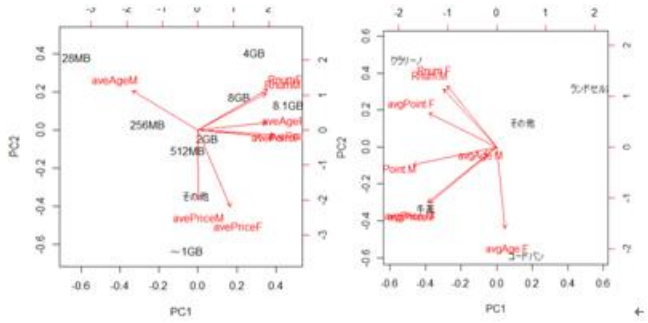
その他の例として記憶容量別に平均価格を算出したものがある。記憶容量が多くなれば平均価格が上がるとは限らないことが分かった。その理由としては 1GB などのフラッシュメモリは漆絵などを施したメモリなど装飾にこだわったものが存在したからである。また男女別の分析結果として、男性は記憶容量が多い方を好む傾向がある。

ランドセルに対する結果はフラッシュメモリと同様に男女共にレビュー数が多いジャンルは平均ポイントも高い傾向がある。

3.2 主成分分析

図 1(a)にフラッシュメモリに対する主成分分析の結果を示しており、８個の集計データが次のように３つの主成分に縮約されていると言える。

- 男性の平均年齢
- 男性・女性の平均価格
- 男性・女性の平均レビュー数と平均ポイント、女性の平均年齢



(a) フラッシュメモリ (b) ランドセル

図 1 主成分分析の結果

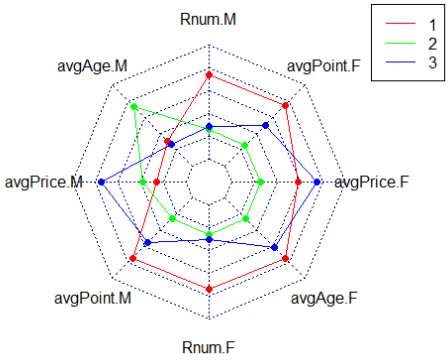
ランドセルに対する主成分分析の結果を図 1(b)に表示している。フラッシュメモリと同様にクラリーノは男女共にレビュー数と平均ポイントが高い位置にあり、逆に高級素材を用いた商品は平均年齢や平均価格が高い位置に配置されている。一方、特徴の少ないその他やランドセルカバーに対する成分がない。

Practice of Data Mining with R in Rakuten Ichiba Review Data

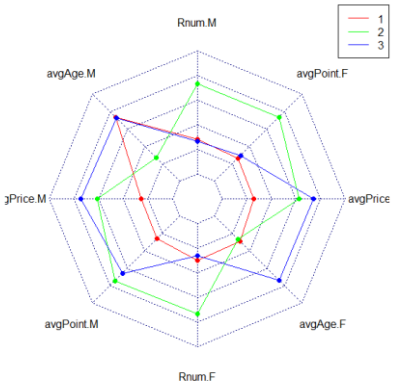
† Ohgusu Takuya・Xu Haiyan , Fukuoka Institute of Technology

3.3 クラスタリング分析

R 言語で k-means によるクラスタリングを行った結果、128MB のみの 2 グループ目、4GB 以上の 1 グループ目、それ以外の 3 グループ目に分かれた。図 2(a)にフラッシュメモリに対するクラスタリングの結果を表示したレーダーチャート図を示している。この図より、グループ 2 は 128MB 男性レビュー者の年齢層の高さを示していることが分かる。グループ 1 は男女共にレビュー数や平均ポイント高いかつ女性の平均年齢が高い結果である。さらに、グループ 1 と 2 の中間に属する記憶容量を持つグループ 3 は、前述した 1GB を含むため男女共に平均価格が高い傾向にある。



(a)フラッシュメモリ



(b)ランドセル

図２　レーダーチャート

図 2(b)にランドセルに対するレーダーチャート図を示している。コードバンと牛革いうジャンルは、男性・女性の平均価格が高く、かつ女性の平均年齢が高いグループに、クラリーノは、男性・女性のレビュー数も平均ポイントも多いグループに、ランドセルカバーとその他は、男性の平均年齢が高いグループに分けられていることが分かる。

3.4 時系列分析

図 3 にランドセルデータ[2]を時系列で表わしたものを表示している。図より入学を控えた 1 月頃が多い。さらにお盆を過ぎた 8 月あたりからレビュー数が増大していることが分かる。理由は、お盆とお正月に帰省した孫のために祖父母が購入するからだと考えられる。しかしレビューという特性上、購入して数日後しか投稿しないため、多少誤差が生じることがある。一方フラッシュメモリでは、あまり年月に差はない。

図３　ランドセルの時系列データ

3.5 形態素解析

3.1 節の楽天市場レビューデータ内のフラッシュメモリに関する商品名とレビュー内容に対して RMeCab を用いて形態素解析を行った。今回は世代別に着目して解析した。

世代別商品名に対する名詞の解析としては、全世代において”無料”や”高速”が商品購入において重要視される傾向が得られた。また 40 代以上は”キャップレス”を好む傾向がある。

4. まとめ

本研究では楽天市場データ内のレビューデータの分析や解析を行ってきた。さらに得られた結果に対する理由づけも商品情報などを通して行うように努めている。今回は主にフラッシュメモリなどの記憶媒体とランドセルの分析を行った。前者に対する分析結果として、レビュー数が多いジャンルは平均ポイントも高くなる傾向にあることが分かった。後者に対する分析結果として、男性と女性の平均年齢は異なる傾向にあることが分かった。異なるジャンルや商品類に対する分析を行っていくことが今後の課題である。

参考文献

[1]高橋・天笠・北川,『レビューデータにおける評価の時系列変化的変化に着目したイベント抽出』，DEIM Forum 2012
[2]『日経情報ストラテジー』(2014 年 12 月号)
[3]楽天データセット:
http://rit.rakuten.co.jp/opendataj.html