

商品レビューにおける時系列情報に着目したクラスタ分析と可視化

Cluster analysis and visualization for time-series product reviews

小川 和晃^{*1} 田村 哲嗣^{*1} 速水 悟^{*1}
 Kazuaki Ogawa Satoshi Tamura Satoru Hayamizu

^{*1}岐阜大学 工学部 応用情報学科

Department of Information Science, Faculty of Engineering, Gifu University

Recently, subjective documents on Internet, e.g. product reviews, have attracted attentions. This paper aims at analyzing product reviews focusing on time-series and attribute information. From each review, term document, semantic and attribution features are extracted using LSA(Latent Semantic Analysis), word2vec techniques. We conduct clustering and visualize feature vectors of clusters using SOM(Self-Organizing Maps), extraction of keywords. In addition, using ratio of two conditional probabilities, characteristic clusters are extracted. By subjective evaluation of experiments, we investigate effectiveness of clustering and some interesting clusters from visualization.

1. 序論

近年、インターネットの普及により、日々膨大なテキストデータが蓄積されるようになった。その中でも、個人の感情や意見などの主観的な情報は、我々の意思決定に多大な影響を与えるため、これらを活用する技術が求められている。主観的な情報の一つとして商品レビューが挙げられるが、膨大なレビューの中から網羅的な情報を効率良く得ることは容易ではなく、レビューの整理や要約による傾向の把握が必要である。

主観情報を含むテキストデータを活用する研究は、評判分析や意見マイニングなどの分野として盛んに行われており、[乾 2006] によってまとめられている。また、[打田 2010] はレビューにおける評価内容の変化に着目し、変化する評価情報を時系列で捉えることで、レビュー全体の傾向の把握を支援する手法を提案した。さらに、[高橋 2013] はレビュアーが与える評価値の時系列変化に着目し、変化が見られた時間前後のレビューを分析することで、評価に影響を与えた要因を検出する手法を提案した。このように、時系列情報に着目し、評価情報の変化や相違を踏まえてレビューの傾向を把握することは重要である。しかし、評価の傾向を把握する際は、時系列だけでなく性別や年齢といった属性にも着目する必要がある。

そこで本研究では、Web 上の商品レビューにおける時系列情報と属性情報に着目し、レビュー全体の傾向の分析を目的とする。クラスタリング手法を用いてレビューを分類し、時系列と属性において特徴的なクラスタを抽出することで傾向の把握を目指す。さらに、得られた結果を視覚的に把握するための可視化を試みる。なお、多次元テキストデータの可視化には低次元空間へのマッピングが有効とされているが、[齊藤 2014] は非線形性を持つテキストデータに対し自己組織化マップ (Self-Organizing Maps; SOM) を適用することで、より頑健なレビューの傾向の把握支援を提案している。よって本研究では、SOM を用いた可視化を検討する。

2. 提案手法

以下より、提案手法の詳細を段階ごとに記述する。

2.1 レビューの収集とスクリーニング

本研究では、楽天市場の「みんなのレビュー・口コミ情報」における商品レビューを使用する [楽天データ]。各レビューには性別、ジャンル ID、レビュー投稿日時などの属性情報が付与されている。そこで、ジャンル ID を用いて特定のジャンルのレビューのみを抽出する。また、短いレビューは有用な情報が少ないと仮定し、レビュー文書の名詞、形容詞、動詞の合計語数が一定数以上のレビューのみを扱う。なお、本研究では形態素解析エンジンである MeCab を用いて形態素解析を行い、単語と品詞情報を取得する。

2.2 レビューの特徴量抽出

レビューに対してクラスタリング手法を適用するために、次に示す 4 種類のレビューの特徴量を抽出する。

1. TF-IDF により得られる文書ベクトル
2. word2vec により得られる単語ベクトル
3. 時系列情報にもとづくベクトル
4. 属性情報にもとづくベクトル

(1) TF-IDF により得られる文書ベクトル

文書の特徴量として、高頻出かつ一般性の低い単語を重要単語として抽出する TF-IDF がよく用いられる。しかし、TF-IDF は特徴ベクトルが疎かつ高次元となる傾向がある。特にパターン認識分野の識別などにおいては、高次元であるほど識別の精度が悪化する次元の呪いと呼ばれる問題が発生する。よって本研究では、TF-IDF に標準化処理を施し、次元削減手法の一つである LSA(Latent Semantic Analysis) によって次元数を削減した文書ベクトルを使用する。

(2) word2vec による得られる単語ベクトル

word2vec とは、単語の共起関係にもとづき、単語をその意味的情報を持つベクトルで表現する分散表現の手法の一つである。[Tomas 2013] が提案しており、連続 Bag-of-Words モデルや skip-gram モデルを使用している。本研究では、各レビューにおける TF-IDF 値が最も高い単語を最重要単語とし、word2vec で得られたその単語の単語ベクトルを特徴量に使用する。word2vec は単語の意味的情報を考慮するため、後述するキーワード抽出において特に有効であると期待される。

(3) 時系列情報にもとづくベクトル

時系列情報を月単位で扱うとして 12 次元のベクトルを作成し、レビュー投稿日時に該当する月の要素のみを 1、それ以外

連絡先: 〒 501-1193, 国立大学法人岐阜大学工学部 速水・田村研究室

の要素を 0 とする。この特徴は、後述する特徴的なクラスタの抽出で特に有効であると期待される。

(4) 属性情報にもとづくベクトル

属性情報も同様に、該当する属性のみを 1、それ以外を 0 としたベクトルを用いる。例えば、性別であれば男性と女性の 2 次元ベクトルを特徴とする。この特徴に関しても、特徴的なクラスタの抽出に有効であると考えられる。

2.3 クラスタリング

2.2 節で述べた特徴量を用いて、レビューのクラスタリングを行う。本研究では、非階層的手法として一般的な k-means 法のうち、初期値や外れ値に依存する問題の解消を目的とした k-means++ 法 [David 2007] を用いる。k-means++ 法では、初期値を選択する際に、最初の初期値のみは乱数によって決定する。それ以降の初期値は、最も近いサンプルとの距離を $D(x_i)$ としたとき、それまでの初期値と各サンプルとの距離により得られる確率分布 $\frac{D(x_i)}{\sum_k D(x_k)}$ が最大となるサンプルを初期値として決定する。つまり、初期値同士ができる限り離れるように初期値を選択していく。なお、クラスタリングの距離尺度にはコサイン距離を使用する。

2.4 結果の可視化

(1) クラスタにおけるキーワード抽出

各クラスタを解釈するために、クラスタのキーワードを抽出する。キーワード抽出には、重要単語の抽出によく用いられる TF-IDF を使用する。本研究におけるクラスタのキーワード抽出の流れを以下に示す。

1. クラスタ内の各レビューにおける TF-IDF を算出し、TF-IDF 値上位 5 単語を抽出する。
2. クラスタ内の全レビューで抽出された単語の TF-IDF 値を、各単語ごとに合算する。
3. 2. の合算値の上位単語をクラスタのキーワードとする。

(2) SOM によるクラスタの可視化

k-means++ 法の結果の可視化に SOM を利用する。SOM は入力層と出力層のみで構成されるニューラルネットワークであり、入力データの位置関係を保持したまま、高次元データを 2 次元空間上に写像することが可能である。本研究では、クラスタ同士の関係性の視覚的な把握支援となることが期待される。

2.5 特徴的なクラスタの抽出と可視化

クラスタリング結果に対して、時系列情報と属性情報に着目した分析を行う。各時系列における属性ごとの特徴的なクラスタ群を抽出し、可視化を行うことで分析を進める。

(1) 特徴的なクラスタの判定

本研究では、注目した時系列、属性のレビューが多く、かつそれ以外の時系列、属性のレビューが少ないクラスタほど特徴的であるとする。同時に、注目した時系列、属性のレビューが、他のクラスタよりも多いクラスタほど特徴的であるとする。以下より、注目する時系列を t 、属性を x としたときの、特徴的なクラスタ c を判定する方法を記述する。

1. 注目する時系列 t かつ属性 x の条件の下でクラスタ c が生成される条件付き確率 $P(c|t, x)$ を求める。この確率は以下の式 (1) に変形できる。

$$P(c|t, x) = \frac{P(c, t, x)}{P(t, x)} \quad (1)$$

一方で、注目した時系列 t 以外の時系列を $\bar{t}$ 、また注目した属性 x 以外の属性を $\bar{x}$ としたとき、時系列 $\bar{t}$ かつ属

性 $\bar{x}$ の条件の下でクラスタ c が生成される条件付き確率 $P(c|\bar{t}, \bar{x})$ を求める。

2. 次に、クラスタ c の条件の下である時系列 t かつ属性 x のレビューが存在する条件付き確率 $P(t, x|c)$ と、注目したクラスタ c 以外のクラスタ $\bar{c}$ の条件の下で時系列 t かつ属性 x のレビューが存在する条件付き確率 $P(t, x|\bar{c})$ を、同様の手順で求める。
3. 1. で求めた確率の比と 2. で求めた確率の比を各々算出し、それらを用いた次の式 (2) を指標値とする。

$$rank(c) = \log \frac{P(c|t, x)}{P(c|\bar{t}, \bar{x})} + \log \frac{P(t, x|c)}{P(t, x|\bar{c})} \quad (2)$$

式 (2) の第 1 項目はクラスタ c における注目した時系列 t と属性 x の割合、第 2 項目は注目した時系列 t と属性 x におけるクラスタ c の割合にもとづく値である。これらの値が高いほど特徴的であると判定する。

(2) 特徴的なクラスタ群の可視化

2.4 節の (1) と (2) より、本節の (1) で抽出した特徴的なクラスタ群の可視化を試みる。k これにより、時系列情報と属性情報ごとの相違の視覚的な把握支援が期待される。なお、より多くの情報を視覚的に把握するために、次の処理を加える。

- 各クラスタにおける最上位キーワードを提示する。
- 特徴的なクラスタは、属性情報ごとに色分けを行う。
- 本節の (2) で得られた指標値と、表示するキーワードの大きさを対応付ける。

3. 実験

本研究における手法の妥当性を検証する実験を行った。

3.1 実験条件

本実験では、2 章で述べた手法を実際の商品レビューに適用した。クラスタリングの実験条件を表 1 に示す。word2vec は、ジャンルを問わない 1436083 件のレビューを使用し、単語ベクトルの次元数を 100、文脈窓の大きさを 5 単語とした。SOM は、マップの大きさを 20×20 、学習回数を 1000、学習率を 0.005 に設定した。なお、提案手法における 2.2 節の条件としては、実験的に LSA を適用した TF-IDF のベクトルを 50 次元、word2vec により作成した単語ベクトルを 100 次元、時系列情報のベクトルを 12 次元、属性情報のベクトルを 2 次元の合計 164 次元とし、クラスタ数を 200 に設定した。本実験では属性情報の例として、最も単純である性別に着目した。

3.2 実験結果

「デジタルカメラ」ジャンルのレビューに本研究の手法を適用した結果を述べる。最初に、クラスタリング結果に対して SOM を用いた可視化結果を図 1 に示す。図 1 中では、各クラ

表 1: クラスタリングの実験条件

対象ジャンル	デジタルカメラ
データ件数	10564 件
対象期間	2010 年から 2012 年
クラスタリング手法	k-means++ 法
距離尺度	cos 距離
クラスタ数	200

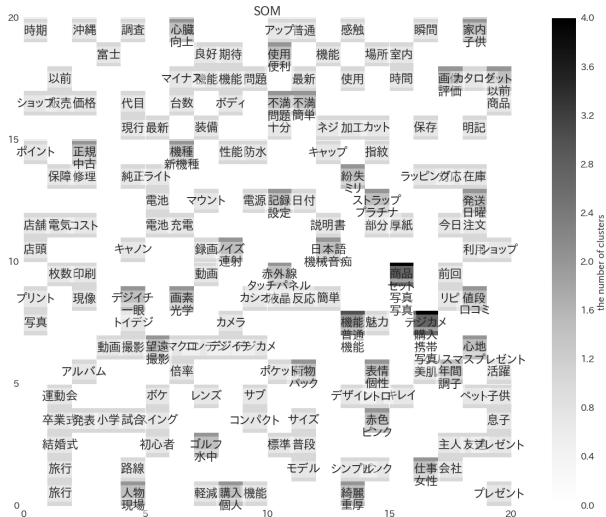


図 1: SOM の可視化結果



図 3: 2012 年 10 月の可視化結果

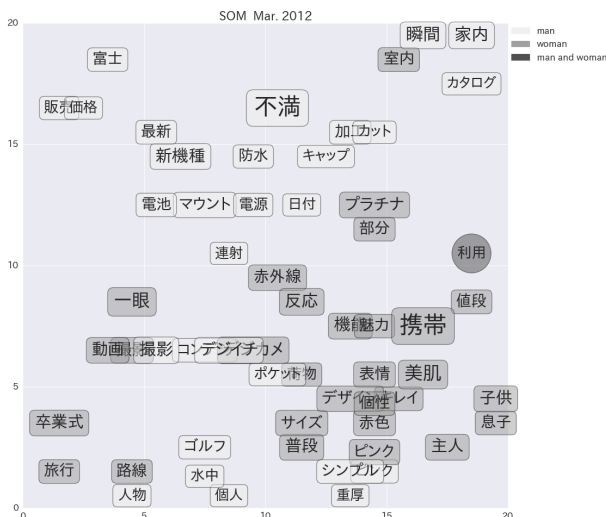


図 2: 2012 年 3 月の可視化結果

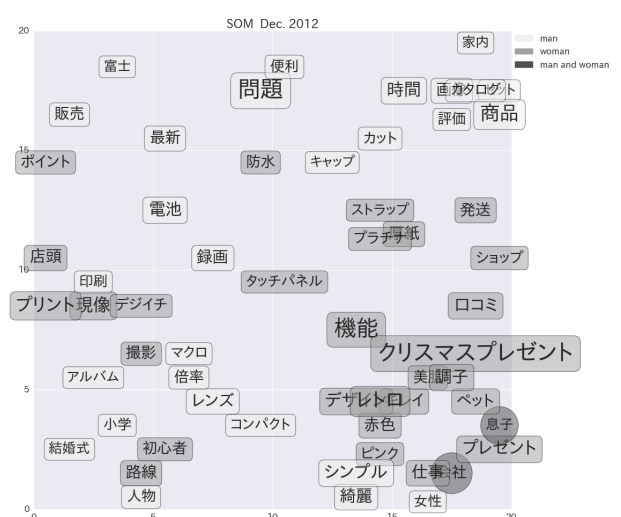


図 4: 2012 年 12 月の可視化結果

スタのキーワード最上位とクラスタ間の位置関係を表している。続いて、各時系列における性別ごとの特徴的なクラスタ群を抽出した。それらのクラスタを可視化対象とした SOM の結果の一例を図 2 から図 4 に示す。

3.3 結果の考察

「デジタルカメラ」ジャンルのレビューに対する実験結果の考察を行う。

(1) 可視化結果の考察

図 1 における SOM の結果を見ると、類似したキーワードが近くに位置しており、クラスタ間の関係性を把握できることが分かる。よって、SOM による位置関係の保持が作用したといえ、クラスタリング結果の把握における有効性が確認された。

次に、図 2 から図 4 に示された特徴的なクラスタ群の可視化結果を考察する。図 2 の 2012 年 3 月の結果を見ると、男性では「新機種」や「最新」といった単語が見られる。これは新年度をきっかけとした製品の機種などが影響していると推測される。また、女性では「卒業式」や「卒業」の単語が確認できるが、3 月から 4 月は卒業・入学シーズンであるため、卒業式でのカメラの使用に関するレビューが多く存在すると考えられる。図 3 の 2012 年 10 月の結果では、女性において

表 2: 男性に特徴的なクラスタ群の一例

期間	クラスタのキーワード
3 月	新機種 機種 被写体 ドット コンデジ カセット 録音 今頃 工事写真 ライティング シンプル デザイン 携帯 格好 機能 カッコイイ マッチ レトロ カッコ 最新 後継 新型 進歩 搭載 旧型 一昔 発売 モデルチェンジ 進化
10 月	ゴルフ バイク アウトドア 散歩 競技 ランニング 体育館 観戦 行楽 登山 電池 乾電池 単三 予備 寿命 充電 バッテリー 充電機 消耗 機種 マクロ 接写 夜景 野鳥 撮影 顕微鏡 手ブレ レンズ 機能 パノラマ
12 月	時間 経過 以内 充電 撮影 半日 開始 一週間 連続 週間 録画 起動 動作 削除 再生 メモリ ソフトウェア 認識 パソコン 画像 倍率 ピント アングル 望遠 無限遠 不満 カメラ レンズ 必要 コンパクト デジカメ

表 3: 女性に特徴的なクラスタ群の一例

期間	クラスタのキーワード
3 月	美肌 メイク 成果 オンチ 小顔 効果 化粧 クマシミ 把握 卒業式 入園 入学 式 入学 子供 卒園 写真 小学校 カメラ 七五三 ピンク ホワイト ピンク色 黄色 ブラック レッド ブラウン 茶色 タッチパネル ビビッド
10 月	運動会 行事 花火大会 幼稚園 子供 遠足 音楽会 花火 保育園 一眼レフ 荷物 収納 整理 旅行 一眼レフ タクチン パンパン レジ袋 母親 容量 簡単 便利 調節 片手 スライド 機能 可能 操作 最初 ワンタッチ
12 月	クリスマスプレゼント 息子 デジカメ ピンク クリスマス 愛用 値段 子供 電池 年齢 プリント 写真 注文 大量 用紙 アルバム 自宅 インク サイズ 出方 ペット 黒猫 愛犬 愛猫 ワンコ 動物 子猫 子犬 ネコ 撮影

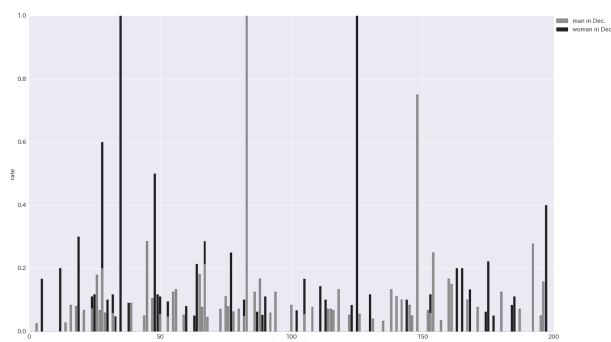


図 5: 2012 年 12 月の各クラスタにおけるレビュー分布

「運動会」の単語が見られる。これは、この時期の開催が多い運動会を始めとする行事などが要因であると推測される。図 4 の 2012 年 12 月の結果を見ると、男性では「レンズ」や「倍率」といったカメラのピントなどに関する単語が見られ、女性では「クリスマスプレゼント」、「プレゼント」などの単語が特徴的である。「クリスマスプレゼント」は、12 月がクリスマスのシーズンであることが大きく影響していると推測される。また、「印刷」や「現像」、「プリント」などの単語が特徴として確認できるが、この時期は年賀状に向けた動きが予想されるため、それらが多少なりとも関係する可能性が考えられる。

以上より、各時系列における性別ごとの傾向を視覚的に把握できる可能性を示した。また、特定の時系列におけるイベントが検出できたため、それらの結果の妥当性を確認した。しかし、本実験で示したキーワードは必ずしも適切ではないため、より確実な分析に向けてキーワードの提示方法などを改善する必要がある。

(2) クラスタのキーワード群の考察

表 2 と表 3 より、各クラスタにおけるキーワード抽出の結果を考察する。表 2 の 3 月では、「最新」、「新型」などがあるクラスタはカメラの新機種に関係すると推測される。10 月では、「ゴルフ」や「アウトドア」などのクラスタは屋外での使用に関係すると推測される。また、「電池」、「充電」といった単語より、充電に関係するクラスタも予想できる。表 3 では、3 月に「ピンク」や「ホワイト」などのカメラの色に関係するクラスタが読み取れる。また、3 月の「卒業式」や「入学式」、10 月の「運動会」や「音楽会」が見られるクラスタは学校行事での使用に関係すると考えられる。12 月では、「クリスマスプレゼント」や「息子」などのクリスマスの贈り物などに関係するクラスタが推測される。また、「ペット」や「愛犬」などのペットの撮影に関するクラスタも予想される。以上より、単語の意味や使われ方が類似したキーワードの集合が見られた。よって、word2vec による単語の意味的情報の有効性が確認された。

しかし、表 3 の 12 月における、「プリント」や「写真」が見られるクラスタ内のレビューを見ると、カメラに関連したデジカメプリントのレビューの混在が見られた。また、文字数の制限だけでは防ぐことができないスパムレビューの存在も確認された。よって、商品レビューの選定が課題の一つとして挙げられる。その一方で、特定のジャンルに絞るだけでなく、複数のジャンルに渡る関係性を見ることも重要であると考えられる。

(3) クラスタ内のレビュー分布の考察

図 5 に、クラスタ内のレビューにおける時系列と性別の分布の一例を示す。これより、各クラスタ内のレビューの時系列と性別に偏りが確認され、時系列と性別を考慮したクラスタリングとなっていることが推測できる。

4. 結論

本研究では、商品レビューを複数のクラスタに分類した後、クラスタのキーワード抽出と SOM による関係性の可視化を行うことでレビューの傾向の把握を検討した。また、時系列情報と属性情報に着目した特徴的なクラスタの抽出と可視化により、時系列と属性ごとの傾向の相違を分析した。キーワードと SOM の可視化結果をもとに主観的評価を行ったところ、特定の時系列に特徴的と考えられるイベントの検出が見られ、妥当性の確認ができた。

クラスタリングでは、意味的に類似したキーワード群が得られたため、word2vec による単語の意味的情報が有効に作用したと考えられる。また、時系列や性別に特徴的だと想定される結果が得られたことから、時系列情報と属性情報を考慮した特徴量の追加や特徴的なクラスタの判定によって、レビューの傾向を抽出できることを確認した。

5. 今後の課題

本実験では、k-means++法における特徴量や次元削減手法、次元数、クラスタ数、さらには SOM の学習回数や学習率などを十分に検討できなかった。よって、これらの調整や最適化が求められる。また、他のジャンルや性別以外の属性情報を用いた実験を行う必要がある。そして、クラスタリングの初期依存問題を完全には解決していないため、より確実性の高い分析結果を追求すべきである。一方で、時系列や属性に対する着目方法や SOM の可視化方法の改善、SOM 以外の可視化手法の適用なども検討したいと考えている。最後に、本実験では主観的評価の占める割合が高いため、より客観的な評価を検討することで、さらなる妥当性の検証を進める必要がある。

参考文献

- [乾 2006] 乾 孝司, 奥村 学, “テキストを対象とした評価情報の分析に関する研究動向”, 自然言語処理, Vol.13, No.3, pp.201-241, 2006
- [打田 2010] 打田 裕樹, 吉川 大弘, 古橋 武, 平尾 英司, 井口 浩人, “Web ユーザーレビューにおける評価情報の時系列変化の可視化”, 知能と情報 (日本知能情報ファジィ学会誌), Vol.22, No.3, pp.377-389, 2010
- [高橋 2013] 高橋 毅, 天笠 俊之, 北川 博之, “評価値の時系列変化に着目したユーザーレビューの分析”, DEIM Forum, D-10, 2013
- [齊藤 2014] 齊藤 史哲, “Web カスタマーレビュー文の理解支援を目的とした自己組織化マップによる評価分布の可視化法”, J Jpn Ind Manage Assoc Vol.65, No.3, pp.180-190, 2014
- [楽天データ] “国立情報学研究所 情報学研究データリポジトリ”, <http://www.nii.ac.jp/dsc/idr/>
- [Tomas 2013] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space”, ICLR, 2013
- [David 2007] David Arthur, Sergei Vassilvitskii, “k-means++: The Advantages of Careful Seeding”, ACM-SIAM, pp.1027-1035, 2007