ISS-SP-060

複数のクラスタリング手法の統合利用による商品レビュー閲覧支援システム

ギ　イツリン† 　佐藤　悠大† 　韓　東力††
† 日本大学文理学部情報科学科

1. はじめに

近年ネットショッピングの急激な発展により、読みきれない量のレビューがある。現在それらのショッピングサイトでは、レビューは人手によってレビュー整理を行うのが一般的である。また、文を整理するための手法としても、辞書や学習データを用いる研究[1][2]が多くて、文の単語数が少なく、かつ更新ペースが速い商品レビューを処理するには限界がある。

以上の研究背景から私たちは商品のレビュー文の主な内容を、顧客が短時間で理解できるようにするために、閲覧支援システムを構築していくことを研究目的とした。

2. 提案手法

2.1 データの数値化とストップワードリスト

本研究は Wikipedia 日本語版を用いて Word2vec に学習した結果を利用し、単語をベクトル化する。

また、レビュー内に、一般的に役に立たない等の理由で処理対象外となる単語が含まれているため、対象商品と同カテゴリに含まれている上位30件の商品のレビュー文の集合を用いて tfidf 値を計算する。tfidf 値が低い単語をストップワードとする。レビュー文の値(ベクトル)は中に含まれている単語ベクトルの平均値とする。

単語は Mecab で抽出し、ストップワードに含まれていない名詞のみ利用する。

2.2 クラスタリング

文ベクトルのデータは中心に近づくほど、データの密度が増え、互いの距離が小さくなるので、密度が異なる環境で同一条件の下ではクラスタリングすると、中心部分に大量のレビューが1つのクラスタに集中しすぎるため、クラスタリングがうまくできていない問題が存在する。

以上の問題を解決するために DBSCAN と階層的クラスタリングと統合利用し、DBSCAN のパラメーターの自動調整で実現した階段式クラスタリングを提案する。

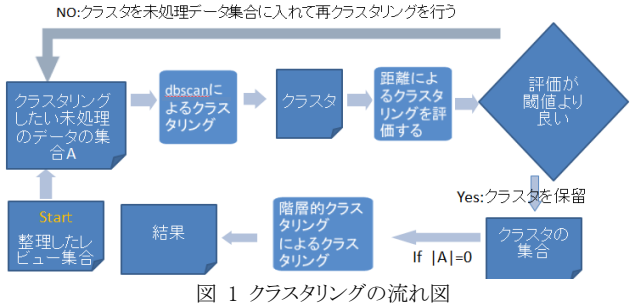

図 1 クラスタリングの流れ図

図 1 のように、2.1 で生成されたレビュー文ベクトルの集合を DBSCAN でクラスタリングする(詳しくは 2.3)。結果に対して式(1)を用いて、クラスタごとに評価する。閾値 a より評価が良いものはそのまま保留し、悪いものはもう一度クラスタリングして評価する。未処理のデータが存在しないとき、DBSCAN によって細切れされたレビューを整理するために、保留された評価の良いクラスタを階層的クラスタリングする。

$$\text{クラスタ i の評価値} = \frac{MAX(|C_i - k_r|)(r \in i)}{MIN((|C_i - k_r|)(r \notin i))} \quad (1)$$

C_i:クラスタ i に含まれている全レビューの値の平均値　k_r:レビューr のベクトル

2.3 DBSCAN の利用

DBSCAN には、2のクラスタの距離の近さで同じクラスタかどうかを判定する探索範囲とよばれているパラメーターがあり、以下のように探索範囲を変化させることで、中心部分から階段式でクラスタリングをする。

ある探索範囲nで集合をクラスタリングし、結果は全部雑音と判断されるとき、クラスタができるまで探索範囲nを拡大する。nは閾値 b よりも大きくなる場合、集合にクラスタが存在しないと判断して集合を雑音として捨てる。

3.実験とまとめ

本研究では、提案手法の有効性を検証するために、k-means と階層的クラスタリングをベースライン 1,2 として、アンケート(実験1)と式(2)(実験2)を利用した比較実験を行った。

$$\text{score1} = \frac{\log_{10}(\frac{\sum_a(-\sum_c p_{a,c}\log p_{ac})}{n} + 10)}{}$$

$$\text{score 2} = -\sum_c N_c \log N_c$$

$$\text{total\_score} = \frac{1}{\text{score1}} * score2 \quad (2)$$

式(2)では、C がクラスタ、n が特徴語数、a が特徴語、$P_{a,c}$が特徴語 a がクラスタ C 内における出現頻度、N_cがクラスタ C に出現した特徴語の数をそれぞれ表している。結果はグラフ1の通り、本手法はより有用性があることを示すことができた。

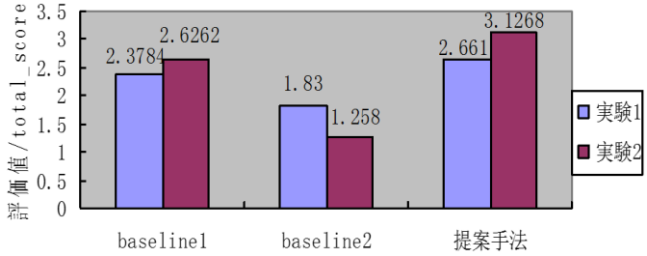

図 2 実験結果

今後は word2vec における未知語の対応などを改良する予定である。

参考文献

[1] 大規模 EC サイトの商品レビュー傾向分析
林 驍,伊東 栄典,廣川 佐千男
情報処理学会研究報告 vol.2015-ICS-181 No.7

[2] 商品推薦のための商品レビューの極性分析に基づく特徴語抽出手法 吉田 朋史 北山 大輔 DEIM Forum 2015 カテーテ