修士研究について

水上 和秀 (Kazuhide Mizukani)
u355020@st.pu-toyama.ac.jp

富山県立大学 工学部 電子情報工学専攻

October 31, 2023

1.1 本研究の背景

進捗

- 自然言語処理分野で進める

背景

- SNS や Web サイトの利用者が増加し，口コミを参考にして商品を購入する人が増えてきている
- 現代の消費行動の特徴としてインターネット上の口コミや評価を重視し，意志決定や行動に大きな影響を与えている．
- しかし，Web 上に存在するレビューの数は膨大であり，その全てを閲覧し有用な情報を判断した上で利用することは困難である
- 一方，レビュースパムと呼ばれる信頼性の低いレビューの投稿が問題となっており，信頼性を意識したレビューの判断団が必要である．

目的

- レビューサイトの口コミの信頼性，ネガポジ感情値をスコア化し，商品を推薦するシステムを提案する

2 レビューのスコア化について

レビューのサイトの分析

- 感情分析により，特定の商品等に対する一般消費者の意見を定量化することができる
- インターネットにおける商品レビューを感情分析をした場合のネガポジ判定は，商品の評価スコアに比例することが確かめられた (参考論文 1)

レビューの信頼性

- オンラインカスタマーレビューに関する調査では，レビューを参考としていると回答した人が 7 割以上であった．消費行動は口コミにより行動が大きく左右されるため，口コミの信頼性を見極めることが必要である
- レビューの信頼性を表す指標として，**類似性，協調性，集中性，情報性**という 4 つの信頼性指標を定義し，各指標ごとのスコアを求めた．そのスコアにより，ユーザの信頼性に対する意識を促すとともに，有効な判断支援が行えることが確認できた．(参考論文 3)

→レビューの信頼性を考慮した商品のスコア化が必要となる

3 単語のネガポジ感情分析

単語感情極性対応表

- 単語の感情極性を対応させた表
- 感情極性値は，語彙ネットワークを利用して自動的に計算された．(参考論文 2)
- 1 に近いほどその単語が良い印象を持ち，-1 に近いほどその単語が悪い印象を持つ

	word_type_score
0	優れる:すぐれる:動詞:1
1	良い:よい:形容詞:0.999995
2	喜ぶ:よろこぶ:動詞:0.999979
3	褒める:ほめる:動詞:0.999979
4	めでたい:めでたい:形容詞:0.999645
...	...
55120	ない:ない:助動詞:-0.999997
55121	酷い:ひどい:形容詞:-0.999997
55122	病気:びょうき:名詞:-0.999998
55123	死ぬ:しぬ:動詞:-0.999999
55124	悪い:わるい:形容詞:-1

[55125 rows x 1 columns]

4 信頼性判断 1/4

類似性

複製またはそれに近いレビューには多くのスパムが含まれている．そこで，どの程度ほかのレビューと類似しているかを測る指標として**類似性スコア**がある (参考論文 3)

計算式

レビュー r_j を単語ごとに分割したときの単語の集合を X_{r_j} とすると，$jaccord$ 係数を用いて手ビュー r_i と r_j の類似度を以下のように求める．

$$sim(r_i, r_j) = \frac{|X_{r_i} \bigcap X_{r_j}|}{|X_{r_j} \bigcup X_{r_j}|} \tag{1}$$

$|X_{r_i} \bigcap X_{r_j}|$:$X_{r_i}$ と X_{r_j} のどちらにも存在する要素数
$|X_{r_j} \bigcup X_{r_j}|$:$X_{r_i}$ または X_{r_j} に存在する要素数

そして，類似スコアを以下のように求める

$$S\_score(r_i) = 1 - \max_{r_i}(sim(r_i, r_j)|j \neq i, j = 1, 2, ...n) \tag{2}$$

4 信頼性判断 2/4

協調性

商品の評判をあげる (または下げる) ことを目的とした，サクラグループの存在がある．これは同じグループのメンバが同じ商品に対して投稿を行い，協力して評判を変えるものである．そこで，各レビューがサクラグループによって投稿されたものであるか測る指標として**協調性スコア**がある

計算式

ある商品 p_i にレビューを投降したユーザ ID の集合を t_{p_i}，ある投稿者グループが出現した t_{p_i} の数をそのグループの指示度数とする．
各頻出投稿者グループ g_c の指示度数 ($= support(g_c)$) とユーザ ID 数 (size(g_c)) を用いて g_c の協調度を以下のように計算する (参考論文 3)

$$collaborate(g_c) = support(g_c)size(g_c) \tag{3}$$

そして，レビュー r_i の協調性スコアを以下のように求める．

$$C\_score(r_i) = \begin{cases} ln(max_{g_c} \in G_{u_{r_i}}(collaborate(g_c))), & |G_{u_{r_i}}| \neq \emptyset \\ 0 & |G_{u_{r_i}}| = \emptyset \end{cases} \tag{4}$$

4 信頼性判断 3/4

集中性

レビューが時間的に集中して投稿されているものほど，そのレビューがスパムレビューである．そこで，各店のレビューに対して，高い（あるいは低い）評価値のレビューがどの程度集中して投稿されている化を図る指標として**集中性スコア**がある (参考論文 3)

計算式

連続時間間隔が連続して短いレビュー集合 g_b のレビューの数 $size(g_b)$ を用いて以下のように求める

$$T\_score(r_i) = 1 - ln(size(g_b)) \tag{5}$$

どのレビュー集合にも属さないレビューの集中性スコアは 0 とする．

4 信頼性判断 4/4

情報性

文章に情報性であるスパムでない可能性が高い．また，情報性のある文章は名詞が多く使われていることがわかっている．つまり，レビュー本文中に特徴的な名詞の割合が少ないほどスパムらしいといえる．そこで，どの程度情報性のある文章であるか測る指標として**情報性スコア**がある (参考論文 3)

計算式

n をレビュー r_i と同じジャンルに属するレビューの数，K_i を r_i に出現する名詞集合とする．また，$term_j \in K_i$ とする．
$df(term_j)$ は r_j と同じジャンルのレビュー集合において $term_j$ を含んだレビューの数とする．

以上を用いて情報性スコアを以下のように求める

$$I\_SCORE(r_i) = ln(1 + \sum_{j=1}^{|K_i|} ln(\frac{n}{df(term_j)})) \tag{6}$$

分析の流れ

流れ

1. ある商品のレビューをスクレイピングする
2. クリーニング処理を行い，不要なテキストデータを除去
3. レビューの類似性スコア，信頼性スコア，集中性スコア，情報性スコアを計算し，レビューの信頼性スコアを計算する
4. 形態素分析で文章を単語ごとに分解する
5. 分解した単語と単語感情極性対応表と比較し，レビューのネガポジ度合いをスコア化する
6. 求めたネガポジ度合いのスコアと信頼性スコアによりレビューの評価スコアを計算
7. レビューの評価スコアの平均値により商品のスコアを求める
8. 商品のスコアをもとにしたユーザにとって最適な商品の推薦

まとめ

中間発表までにできそうなこと

- レビューの信頼性とネガポジ感情を考慮した商品のスコア化

方向性 (予定)

- レビューのスコアをもとにした商品の推薦方法の模索 (遺伝的アルゴリズム?)
- もっと理論について調べる