経験に基づいた有用で希少なレビュー情報の抽出手法の提案

113171025　織田　翔真（灘本研究室）

あらまし：本研究では，ショッピングサイトのレビューから経験情報を含む有用でかつ少数のレビュー筆者が着目した商品の利点，欠点について記載されたレビュー文を希少なレビューとし抽出する手法を提案する．具体的には経験マイニングを行い抽出されたレビューの中から，クラスタリングによって得られた凝集性の低い単語を含むレビューを抽出する．

1．はじめに

近年，オンラインショッピングサイトの普及により，インターネットを通じて手軽に商品を購入することが可能となっている．このオンラインショッピングサイトでは数多くの商品を扱っているため，ユーザが購入を判断する要因の一つとして，レビューがある．レビューは年間消費を約 1 兆 5200 億円押し上げているという研究報告[1]があり，ユーザの購入判断に大きな影響を与えていると考えられる．レビュー文によっては，商品を利用した経験を含む有用なレビューや，少数のレビューの著者のみが着目した商品の利点，欠点が記載されたレビューがある．これらは有用でありかつ希少であると考えられる．しかしながら，レビューの数は膨大でありユーザがこれら有用であり希少なレビューをすべて探し出すのは困難である．

そこで本研究では，商品を利用した経験を含みかつ，見落としてしまう可能性のある少数意見のレビューを「有用で希少なレビュー」として，膨大なレビューからこの有用で希少なレビューを抽出する手法を提案する．具体的には膨大なレビューから経験マイニング[2]を用いて経験を含む有用なレビューを抽出する．次に抽出されたレビューに対し Repeated Bisection 法[2]を用いてクラスタリングを行い各クラスタ毎の凝集性により希少なレビューを抽出し，提示する．

2．提案手法

2．1．全体の流れ

図 1 に示すように，ユーザが商品名を入力すると Amazon から該当するレビュー文を抽出する．抽出されたレビュー文に対して経験マイニングと希少であるかの処理を行い目的のレビューを抽出する．

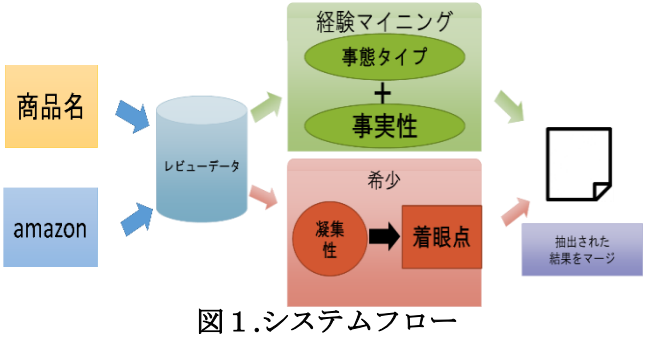

図 1．システムフロー

2．2．経験マイニング

乾ら[3]の研究では，一般性の高い意味的な情報から経験のインスタンスである‘トピック’，‘経験主’，‘事態タイプ’，‘事実性’，‘事態表現’と索引付けする手法で経験マイニングを提案している．本研究の経験マイニングは経験のインスタンスの中でも重要な‘事態タイプ’と‘事実性’を索引付けることで経験マイニングとする．

2．2．1．事態タイプ

事態タイプは経験情報の核となる事態の種類であるポジティブ，ネガティブな出来事，行為である．これらをレビュー文から抽出するために，日本語評価極性辞書[4]を用いてポジティブ/ネガティブ値（P/N）を求める．具体的には以下の式（1）を用いて P/N を算出し，その値がある閾値以上でかつ出来事または行為を表す単語を含むレビューを抽出する．

$$PN = \frac{|P| + |N|}{|A|} \quad \cdots (1)$$

$|P|$はレビュー文中のポジティブの単語出現回数を示し，$|N|$はレビュー文中のポジティブの単語出現回数を，$|A|$はレビュー文中の単語の総数を示す．

2．2．2．事実性

事実性とは事態の事実性に関する情報で本研究では話者態度（モダリティ）を用いて事実性の判断を行う．モダリティ情報の判断には松吉らの日本語機能表現辞書「つつじ」[5]を用いて判断する。レビューの文の総数とレビュー文中に含まれる経験と考えられるモダリティ情報の「過去」「逆説」の出現回数より，経験ではないと考えられる「疑問」「推量」「願望」の出現回数が多い場合，そのレビュー文は経験ではないと考えられる．そこで(2)の条件式を満たすレビュー文を除外する．

$$|G| - |E| \quad >= \quad |S| \quad \cdots (2)$$

ここで，$|G|$は経験ではないモダリティ情報の出現回数であり，$|E|$は経験であるモダリティ情報の出現回数，$|S|$はレビューの文の総数である．

2．3．希少なレビューの取得

本研究では希少なレビューの抽出は，レビュー文に含まれる名詞の出現回数を用いて Repeated Bisection 法によるクラスタリングを行う．Repeated Bisection 法はハードクラスタリングであるため，クラスタリングの対象となった文はいずれかのクラスタに分類される．そのため，相互に関連性の低い単語のクラスタである，凝集性の弱いクラスタに希少なレビューに含まれるクラスタがあると考える．そこで，各クラスタの凝集性を求める．次に凝集性の弱いクラスタに含まれる文に係り受け解析器 KNP[6]を用いて係り受け解析を行う．そして，日本語評価極性辞書[4]のポジティブ又はネガティブな単語に係り受けされている名詞を抽出し，その名詞をレビューの着眼点とする．その着眼点を含むレビュー文を希少なレビューとして抽出を行う．

クラスタの凝集性は，山本[7]らの提案する式(3)を用いて算出する．算出された凝集性が閾値以下の場合そのクラスタを凝集性の弱いクラスタとする．

$$A_i = \sum_{x \in C_i} \left(\frac{x \cdot c_i}{|x||c_i|}\right)^2 \quad \cdots (3)$$

ここでは，i 番目のクラスタC_iのセントロイドc_iとそのクラスタに含まれるレビュー文xのコサイン類似度をレビュー文ごとに求め，その平方和を凝集性とする．

2.4　経験に基づく希少なレビュー

経験マイニングで得られた結果と希少なレビューを取得する方法で得られた結果の両方で抽出されたレビューを本研究の目的としている経験に基づいた有用で希少なレビューとする．

3．評価実験

3．1．実験内容と結果

経験マイニングの評価実験を行った．Amazon から得たモバイルバッテリーのレビュー文210件を被験者 10 人に読んでもらい，経験に基づくレビューの評価を行った．具体的には，レビュー文の内容が商品に対する経験かどうかを，経験ではないと判断したものを0, 経験の詳しいときは5, 詳しくないときは1とし，6 段階評価を行った．被験者 10 人の結果を平均した値をそのレビューの経験の値とする．

閾値を 2，2.5，3 の 3 種類設定し，各閾値以上となる経験の値のレビューを正解データとし，本研究の経験マイニングの適合率と再現率を算出する．表 1 に結果を示す．

表 1．実験結果

	n=2	n=2.5	n=3
適合率	0.358744	0.278027	0.233184
再現率	0.963855	0.746988	0.626506

3．2．考察

経験マイニングの実験結果より適合率が再現率よりも低い値になっている．これはポジティブ, ネガティブの単語が商品とは関係がない場合も使われるため，経験マイニングで抽出されていると考えられる．

４．まとめと今後の課題

本研究ではレビュー情報から経験に基づく有用で希少なレビュー情報の抽出手法の提案を行い、経験マイニングの評価実験を行った．今後の課題としては，商品に関係がない場合に利用されているポジティブ，ネガティブな単語については処理の対象にしないことが必要であると考えられる．

参考文献

[1] 山口 真一,坂口 洋英,彌永 浩太郎, “インターネット上の情報シェアによる消費喚起効果の実証分析” ,GLOCOM Discussion Paper Series 16-1 2016.

[2]Ying Zhao and George Karypis,” Comparison of agglomerative and partitional doc-ument clustering algorithms.” Technical report, Department of Computer Science,University of Minnesota , MN 55455, 2002.

[3] 乾健太郎, 原一夫, “経験マイニング:Web テキストからの個人の経験の抽出と分類”, 言語処理学会第 14 回年次大会論文集，pp1077- 1080, 2008.

[4] 小林のぞみ,乾健太郎,松本裕治,立石健二,福島俊一.”意見抽出のための評価表現の収集”自然言語処理,Vol.12, No.3, pp.203-222, 2005.

[5] 松吉俊, 佐藤理史, 宇津呂武仁 ”日本語機能表現辞書の編纂” 自然言語処理, Vol. 14, No. 5, pp. 123–146 2007.

[6] 笹野遼平，河原大輔，黒橋禎夫，奥村学，“構文・述語項構造解析システム KNP の解析の流れと特徴”,言語処理学会 第 19 回年次大会 発表論文集 2013.

[7] 山本湧輝，熊本忠彦，灘本明代，” 話題と感情に基づくフォロイー推薦と評価”，第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM2016),2016.