4V-03

ユーザの評価を用いたニュース記事における コメントのクラスタリング

大関陸† 安村禎明†
芝浦工業大学大学院 電気電子情報工学専攻†

1. はじめに

総務省情報通信政策研究所が行った令和2年度情報通信メディアの利用時間と情報行動に関する調査報告書[1]では，令和2年に初めて，全年代平均で平日の「インターネット利用」の平均利用時間が「テレビ(リアルタイム)視聴」の平均利用時間を超過したとの報告があった．日常生活において，インターネットの利用が増えたことにより，インターネットのニュース記事には日々多くのコメントが寄せられている．1つのニュース記事に寄せられるコメントは膨大なものもあり，数千件のコメントが寄せられることも多い．数千件のコメント全体を把握するには多くの時間がかかる．また，コメント欄は評価が高いコメントを優先して表示する場合が多いため，偏った見方になる可能性がある．これらの問題を解決するためには，コメント全体を短時間で把握できるような形式で提示できる機能が必要である．

従来，インターネットのニュース記事におけるコメントを扱った研究として以下のようなものがある．但馬による「多値ラベルによるニュース記事のコメントに対する感情推定」[2]では，コメントに対する感情推定の研究がされている．この研究ではコメントに対して「喜び」や「悲しみ」などの感情を推定する．しかし，この研究ではコメントに対する感情ラベルの付与を人が手作業で行っており，その作業に多くのコストがかかっている．また，感情のラベル付けは主観的な判断になってしまうので，人によってラベルが変わる恐れがある．

そこで本稿では，人が付与したラベルを用いることなく、コメントをクラスタリングする手法を提案する．しかしながら，何の基準もなくクラスタリングしたとしてもユーザにとって有用なクラスタリングになるとは限らない．このため，本研究ではコメントのユーザ評価を基準にクラスタを構成することを目指す．

Clustering of comments in news articles using user ratings
† Riku Ozeki, Yoshiaki Yasumura, Shibaura Institute of Technology

2. ユーザの評価を用いたニュース記事におけるコメントのクラスタリング

本研究で使用するデータセットは，インターネットのニュース記事と，その記事に対するコメント，各コメントに対する高評価と低評価の数で構成されている．

本研究で扱うデータは日本語の文章データである．このデータをニューラルネットワークに渡す前にデータの前処理を行う．前処理は文章データのベクトル化を行う．データのベクトル化には日本語 Wikipedia で事前学習された BERT を使用した．学習に用いたモデルは，隠れ層のない単純なニューラルネットワークである．

本研究でのクラスタリングにおいて，コメントをクラスタとして分ける際に何らかの基準を用いることでユーザにとって有用なクラスタリングができると考えた．そこで，コメントに対するユーザの評価数に着目した．各コメントには高評価の数と低評価の数のデータがあり，これらはコメントを形成する大事な要素である．本研究では，このユーザの評価から高評価数と低評価数の比を取り，それをもとに疑似的なラベル付けを自動的に行う．このラベルが最終的なクラスタリングの理想とすることで，ユーザに有用な情報を提示することを目指す．このようなクラスタリングを実現するために，実際にはラベルがついていないデータに対してクラスタリングする手法を提案する（図1）．

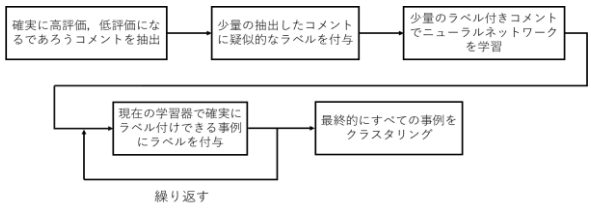

図1 提案手法の概要

本手法では，まずコメントの内容に基づいて確実に高評価，または低評価になるであろうコ

メントを少数抽出し，疑似的なラベルを付与する．次にこの疑似的なラベルからニューラルネットワークを用いて学習し，その学習器で確実にラベルがつけられるような事例にラベルを付与する。これを繰り返し行うことで，最終的にすべての事例をクラスタリングする．

3. 実験

3.1 実験設定

実験データとして，インターネットのニュース記事データセット 50 件の中からコメント数が多いものを用いる．選んだデータの詳細を表 1 に示す．コメントのクラスタをユーザからの評価から，低評価より高評価の方が多い，高評価より低評価の方が多い，どちらも同じ程度の 3 つのクラスタに分けることを基準とした．

表 1　ニュース記事データのコメントの内訳

ラベル	コメント数
高評価 ＞ 低評価	235
高評価 ≒ 低評価	3482
高評価 ＜ 低評価	44
合計	3761

今回は，ラベルの情報を用いない K-means 法でのクラスタリング，図 1 の手順で行う提案手法，ラベル付けしたデータでの教師あり学習を行い，それぞれの実験結果を比較する．

3.2 実験結果

ラベルの情報を用いない K-means 法でのクラスタリング時の特徴量数を変更して精度を測定した．特徴量数は学習済みのネットワークの重みを参照し厳選した．ラベル情報無しクラスタリングの実験結果を表 2 に示す．

表 2　ラベル情報無しクラスタリングの実験結果

特徴量数	精度
768(元の特徴量数)	40.5%
384	41.1%
96	41.2%
24	41.1%

提案手法でのラベル付けしたデータ数を変更して学習し精度を測定した．提案手法の実験結果を表 3 に示す．

表 3　提案手法の実験結果

ラベル付けデータ数	精度
60	55.0%
120	66.0%
200	65.0%

教師あり学習での精度は 90.4%であった．実験結果の混同行列を表 4 に示す．

表 4　教師あり学習での混同行列

	ラベル 1	ラベル 2	ラベル 3
ラベル 1	188	47	0
ラベル 2	41	3438	3
ラベル 3	3	3	38

表 2 から，特徴量数を変えても精度が変化しなかったことから，ラベル情報無しではクラスタリングは難しいことが分かる．また，追加実験として，主成分分析や VAE(Variational Autoencoder)での特徴量数の削減を行い，クラスタリングを行ったが，精度の向上は見られなかった．

表 3 から，ラベル付けしたデータ数が 60 の時に 55.0%の精度であった．ラベル情報無しのクラスタリング時の精度と比べると精度が上がっていることが分かる．ラベル付けしたデータ数が 120,200 の時は約 65%の精度が出ているが，これはデータの偏りによるものだと考えられる．

表 4 から計算した macro-F1 は 89.5%であり，予測できていると考える．教師あり学習で予測できているということは，教師なし学習でも精度を出せる可能性があると言える．

4. 終わりに

本研究ではコメントのユーザ評価を基準にクラスタを構成することを目指した．提案手法の精度を教師あり学習での精度に近づけさせるためには，ネットワークの見直しや，ラベルごとのデータ数の調整をうまく行うことが必要であると考える．また，コメントデータに対する前処理を増やし，ベクトル化をより良いものにすることで精度の向上が見込めるのではないかと考える．

参考文献

[1]総務省情報通信政策研究所．2022.8 .令和２年度情報通信メディアの利用時間と情報行動に関する調査報告書

[2]但馬康宏．2013.9.19. 多値ラベルによるニュース記事のコメントに対する感情推定．情報処理学会研究報告 研究報告数理モデル化と問題解決（MPS） P1-P6