修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/10/2024 (MM/DD/YYYY)

専攻名（専門分野） Department	経営システム 工学専攻	氏 名 Name	中村　太祐 Daisuke Nakamura	指 導 教 員 Advisor	後藤　正幸　　印 Seal
研究指導名 Research guidance	情報数理応用研究	学籍番号 Student ID number	5222C027-4		
研究題目 Title	クエリ指向要約モデルと LDA を組み合わせたレビュー分析手法に関する研究 A Study on Review Analysis Model Combining Query-Focused Summarization Model and LDA				

1. はじめに

近年，商品についてユーザがオンライン上に投稿した大量のレビューデータは，ビジネスにおいて消費者の意見や評価が蓄積された貴重なデータとなっている．しかし，大量のレビューデータをすべて人手で読み分析することは多大なコストを要するため，レビューの要約 [1] が 1 つの重要なタスクとなっている．レビューデータを要約することで，膨大なユーザの意見を簡潔に把握することができるため，商品の改善や顧客満足度の向上につなげることが可能となる．

現在，テキストデータを要約する手法には，様々なタイプのものが提案されているが，特に分析者が求める観点でレビュー全体を要約する手法として，クエリ指向要約モデル（QFS）が提案されている [2],[3]．QFS とは，要約者が設定したクエリ（要約したい観点）を基に，そのクエリの内容に沿った要約を出力するモデルであり，様々な文書が含まれる文書群に対して，ある特定の観点に沿った情報を要約したい場面で用いられる．QFS を用いたレビュー要約では，特定の情報にフォーカスし，冗長性を削減して要約することができるため，効率的に目的の情報にアクセスすることが可能となる．しかし，QFS はクエリに依存して文章が出力されるモデルであり，設定したクエリ次第では知りたい情報をうまく抽出することができず，商品の改善等に有用な要約を得ることが難しい場合がある．そのため，求める観点において有用な要約文を得るためにどのようなクエリを設定すればよいかは，要約者側で試行錯誤して決定する必要がある．

そこで本研究では，QFS によるレビュー要約に，Latent Dirichlet Allocation（LDA）[4] を用いた単語抽出を導入することによって，商品の改善に有用な要約を得る手法を提案する．具体的には，LDA により，設定したクエリと同じトピックで使用されている単語群を抽出し，それらを新たにクエリとして設定し要約を生成することで，商品改善につながるレビューの要約を得る手法の提案を行う．まず，分析を行いたい商品のレビューデータに対し LDA を適用して，設定したクエリの所属確率の高いトピックを求める．その後，選定されたトピックにおいて生成確率の高い単語を，最初に設定したクエリと関連のある単語として抽出する．最終的に，それらを新たなクエリとして設定し，QFS を用いて要約文を生成することにより，分析対象の商品に関する詳細な情報を抽出することが可能となる．

本研究では，Amazon のレビューデータ [5] を用いた実験を行い，提案手法の有用性を示す．この提案により，要約者がより詳細な情報を得るためにクエリを設定する手間を減らし，より有益な情報の抽出のサポートが可能になることが期待される．

2. 準備

2.1. 要約手法の位置づけ

要約モデルはテキストデータから重要な情報を抽出し，それらを簡潔にまとめることが目的である．要約モデルには，元のテキストから重要な文やフレーズを抽出して要約を生成する抽出型要約モデルや，元のテキストの意味を理解し，新しい文の要約を生成する抽象型要約モデルなど，様々なモデルが存在し，データの性質や，要約の目的によって要約モデルを選択する必要がある．その中でも QFS はクエリを基に要約文を生成するモデルであり，要約者が調べたい特定の問題に関する内容をスムーズに抽出することができる．本研究では QFS による文書要約をレビュー要約に適用したケースに焦点を当てて研究を行っている．

2.2. Query-focused summarization（QFS）

QFS[2] は，ユーザが入力したクエリに関連する情報を文書から効率的に抽出し，要約するための手法である．この手法では，まずユーザが文書から特定の情報を抽出するためのクエリを設定する．このクエリは，要約したい箇所やトピックを指定するものであり，要約の焦点を定める重要なものである．次に，設定されたクエリに関連する文書を検索し，ランク付けを行う．クエリに最も適した文書がトップに来るようにランク付けされ，重要な情報が含まれている可能性の高い文書が特定される．最終的に，ランク付けされた文書から要約を生成する．

QFS は特定の情報にアクセスし，意思決定をサポートするのに役立つため，レビューデータを含む多くの文書データにおいて重要なモデルとなっている．ここで QFS のイメージを図 1 に示す．

文書データとクエリを入力とし，クエリそれぞれに対して要約文を生成し，出力するモデルとなっている．

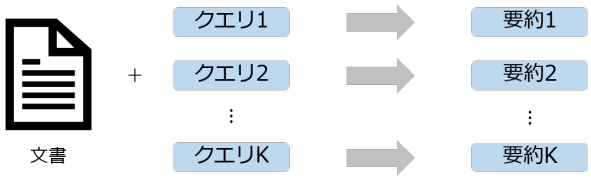

図 1. QFS のイメージ図

2.3. Latent Dirichlet Allocation（LDA）

LDA[4] はトピックモデルの代表的なアプローチの 1 つであり，文書が複数のトピックから生成され，さらにそれぞれのトピックが異なる単語の出現分布を持つと仮定した文書の確率的生成モデルである．具体的には，まず各文書は複数のトピックから構成されているとし，文書のトピック分布からトピックがサンプリングされる．次にそのトピックに基づいて単語が生成される．トピックごとに単語の出現分布が異なっており，その分布から単語がサンプリングされる．このステップを繰り返して，文書内のすべての単語が生成されると仮定している．

トピック分布とは，文書がどのトピックで構成されている確率が高いかを表した分布である．例として，野球について書かれてある文書について考える．この文書に関しては，野球の専門的な単語であったり，スポーツにおいて一般的に使われているような単語が多く含まれていると考えられるため，スポーツというトピックから構成されている確率が最も高いと出力され，次にエンタメ，次に政治といったような確率分布を得ることができる．

また，単語出現分布とは，トピックがどの単語を生成しやすいかを表した分布である．例として，スポーツに関するトピックについて考える．スポーツに関するトピックでは，「勝利」や「野球」，「ワールドカップ」など，スポーツで主に使われるような単語が生成される確率が高いという確率分布を得ることができる．実際に得られる結果としては，出現確率と単語のみであるため，そのトピックが何を表しているトピックなのかは，ユーザ側で判断が必要になる．

文書数は M，トピック数は K，文書 d が n_d 個の単語 $w_{d,i}$ を含むとする．文書 d のトピック分布 $\boldsymbol{\theta}_d = (\theta_{d,1}, \cdots, \theta_{d,K})$ とトピック k の単語の出現確率 $\boldsymbol{\phi}_k = (\phi_{k,1}, \cdots, \phi_{k,V})$ には，以下のような生成モデルを仮定する．

$$\begin{cases} \boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha}) & (d = 1, \cdots, M), \\ \boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta}) & (k = 1, \cdots, K), \end{cases} \tag{1}$$

ここで，$\theta_{d,k}$ は文書 d でトピック k が出現する確率，$\phi_{k,v}$ はトピック k における単語 v の出現確率，V は単語の種類数，$Dir(\cdot)$ はディリクレ分布，$\boldsymbol{\alpha}, \boldsymbol{\beta}$ はディリクレ分布のパラメータである．また各文書 d において単語 $w_{d,i}$ と潜在トピック $z_{d,i}$ は以下のような生成を仮定する．

$$\begin{cases} z_{d,i} \sim Multi\left(\boldsymbol{\theta}_d\right), \\ w_{d,i} \sim Multi\left(\boldsymbol{\phi}_{z_{d,i}}\right), \end{cases} \quad (i = 1, \cdots, n_d), \tag{2}$$

ここで，$Multi(\cdot)$ は多項分布である．これらをまとめたグラフィカルモデルを図 2 に示す．

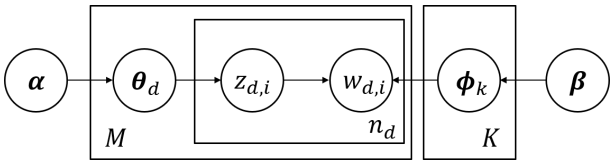

図 2. LDA のグラフィカルモデル

3. 提案手法

3.1. 問題設定

レビュー要約における QFS は，要約者が設定するクエリに基づいて要約文を生成するモデルであるが，クエリ次第でクエリに関する知りたい情報をうまく抽出することができず，商品の改善等に必要な情報を，生成した要約文の中に得ることが難しい場合がある．そのため，適切な要約文を得るためにどのようなクエリを設定すればよいかは，要約者側で試行錯誤して決める必要があるという問題点がある．実際の例として，表 1 はあるプリンターに関する「cost（コスト）」に焦点を当てて，QFS を用いて生成された要約文である．この結果からは，プリンターのコストが高いことは示唆されているが，具体的な要因や側面については詳細には言及されておらず，商品を改善するための情報が得られていないといえる．商品の改善や意思決定のためには，コストが高いとされている要因やその背後にある理由など，さらに詳細な要約結果を得られることが望まれる．そのためには，求める観点において有用な情報が得られるまで，コストに関連のあるクエリを都度設定し直し，要約を生成する必要がある．

表 1. QFS を用いて生成された要約文の例

クエリ	生成された要約文
cost	cost is very high, so budget should be considered. （コストが非常に高いので，予算を考慮する必要がある．）

3.2. 提案手法の概要

本研究では，抽象的な要約しか得ることができないクエリを設定してしまった場合でも，商品の改善に繋げられるような詳細な要約が得られる手法の開発を目的としている．

そこで本研究では，設定されたクエリに対して，関連性が高く，多様な単語群をレビューから抽出する手法を提案する．そのように抽出された単語群を用いて QFS により要約を生成することで，抽象的な要約しか得られないクエリを設定した場合においても，内容が同じような要約を生成することなく，設定したクエリに関連する詳細な要約を得ることが可能となる．ここで提案手法のイメージ図を図 3 に示す．

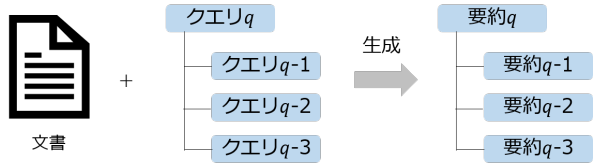

図 3. 提案手法のイメージ図

具体的には，LDA で求まる単語出現確率に着目し，設定したクエリに関連する単語を抽出することを考える．LDA を適用することで，設定したクエリに関連するトピックやトピック内での単語の共起関係を捉えることが可能になる．まず，分析したい商品のレビューデータを用いて LDA を学習する．その後，設定したクエリ q に着目し，各トピックでの単語出現確率 ϕ_k において，クエリ q が最も生成されやすいトピック k_q を求める．k_q は以下のように定義される．

$$k_q = \underset{k}{\operatorname{argmax}} \left(\phi_1^{(q)}, \cdots, \phi_K^{(q)} \right), \tag{3}$$

ここで，$\phi_k^{(q)}$ はトピック k におけるクエリ q の所属確率である．次に，そのトピック k_q で生成されやすい単語群を抽出する．抽出された単語群は，設定されたクエリ q と同じトピックに所属する単語であるため，関連のある単語であると捉えることができる．これにより抽出された単語群を新たなクエリとして設定する．これらの手順により，元々の設定したクエリ q がより具体的かつ，関連性の高い単語群に変換される．最終的に，新しく設定されたクエリを用いて QFS を実行し，要約文を生成する．LDA を用いた単語抽出を行うことにより，設定したクエリの類似語ではなく関連語を抽出できるため，QFS により要約を生成した際に，同じような内容の要約文を生成することを防ぐことができる．

この提案により，要約者がどのようなクエリを設定した場合でも，商品に関する詳細な情報を抽出し，要約文として出力することができる．より具体的な情報が得られることで，商品の改善や戦略の立案においてより効果的な意思決定をサポートできることが期待される．

設定したクエリに対して，関連性が高く，多様な単語群を新しいクエリとして抽出し，レビューデータとそれらのクエリを入力として，クエリそれぞれに対して要約文を生成することで，クエリ設定の手間を省き，商品改善をサポートできる詳細な情報の抽出を可能にすることを目指す．

4. 実験

本研究では，Amazon のレビューデータセット [5] に対して，提案手法を適用し有効性を検証する．

4.1. 分析条件

Amazon のレビューデータセットは商品数が 9.35 億個，それに対するレビュー件数が 82.83 億件の大規模データセットである．レビューはすべて英語で投稿されている．今回対象とするシチュエーションは，ある 1 つの商品に対するレビューの要約を行うものであるため，データセット内の商品を 1 つ抽出し，それに対して投稿されているレビューについて実験を行う．本実験では，表 1 でも扱ったプリンターに対して投稿されているレビューに対して，要約を生成するものとする．レビュー件数は 1,940 件である．LDA の学習を行う前処理として，出現する文書数が 3 件以上かつ全文書数のうち出現する文書数の割合が 50%以下の単語を抽出して学習を行っている．また LDA のトピック数 K は $K = 10$ としている．QFS については，Nema らがいくつか提案した Diversity driven Attention Model のうち，特に精度の良かった SD_2 を用いて要約を行った．比較手法は，Word2Vec による単語抽出と QFS の組み合わせによるレビュー要約とする．Word2Vec による単語抽出は，レビューデータを用いて学習したベクトル表現を用いて，設定したクエリを表すベクトルと二乗誤差が小さいベクトルの上位を抽出する手法である．

4.2. 実験結果

本実験では，「cost」に焦点を当てて要約を行う．プリンターの 1,940 件のレビューにおいて，「cost」という単語が一番生成されやすいトピック k_{cost} における，単語出現確率 $\phi_{k,v}^{(cost)}$ が高い単語を表 2 に示す．

表 2. トピック k_{cost} における単語出現確率が高い単語

単語	出現確率
cartridge	0.052
paper	0.024
price	0.018

表 2 より，所属確率の高い単語として，「cartridge（カートリッジ）」や「paper（紙）」といった単語が抽出できた．これらの単語は，プリンターの運用において消耗品として重要であり，交換や補充にかかるコストが影響を与えている可能性がある．次に所属確率が高い単語として，「price（価格）」という単語が抽出された．これはコストと同じような意味で使われている単語であると解釈ができる．

次に提案手法と Word2Vec により抽出された単語群を新たなクエリとして，QFS を用いて要約文を生成し，比較した結果を表 3 に示す．

表 3 の実験結果から，Word2Vec ＋ QFS では同ような内容の要約文が生成されていることがわかる．一方提案手法では，コストに関するレビューの詳細な要約が生成され

表 3. 新たなクエリとそれに対する要約文

提案手法		Word2Vec+QFS	
クエリ	生成された要約文	クエリ	生成された要約文
cartridge	frequent replacement of ink cartridges is required, and the cost is high.（インクカートリッジを頻繁に交換する必要があり，コストも高い.）	price	price is not commensurate with quality.（価格と品質が見合っていない.）
paper	paper jams occur frequently, often leading to waste.（紙詰まりが頻繁に発生し，よく無駄になる.）	expenses	the purchase expenses are higher compared to other printers.（他のプリンターと比べて購入費用が高い.）
price	price is not commensurate with quality.（価格と品質が見合っていない.）	costs	due to the high costs, I cannot recommend this printer.（コストがかかるため，このプリンターはお勧めできない.）

ており，商品の改善に役立つ結果が得られているといえる．具体的に見ていくと，「cartridge」や「paper」に対しては，このプリンターではインクカートリッジの交換回数が多いことや，紙詰まりによる紙の無駄使いが多いという要約が得られており，これらがコストが高いとレビューされている原因であると考えられる．また，「price」に対しては，価格と品質が見合っていないという要約が得られており，価格の見直しも検討に入れることで，コストが高いという評価に対するユーザの許容範囲を広げる可能性があると考えられる．これらの点について改善を行うことで，コストの面に関して，より顧客満足度の高い商品をユーザに提供できると考えられる．

5. 考察

表 3 の結果から，Word2Vec で得られる単語は，「cost」と同じような単語であったり，「cost」と活用形が変わっただけの単語が抽出されたりしていることがわかる．そのため，それらの単語をクエリとして QFS により要約を生成すると，同じような内容の要約が生成されてしまい，コストに関するより詳細な情報を得ることができない．一方，提案手法では LDA を用いた単語抽出方法であるため，類似語ではないかつ，関連のある単語を抽出することが可能であることがわかる．そのため，それらの単語をクエリとして QFS により要約を生成すると，商品の改善に必要な情報をうまく抽出することができないクエリに対しても詳細な情報を得ることが可能である．

今回はコストに絞って実験を行ったが，他に要約を生成したい観点に対しても同様の手順でレビューの要約を行うことにより，適切な要約を得るためにクエリを試行錯誤して設定する手間を減らし，商品の改善や顧客満足度の高い商品を提供することにつながる結果の獲得が可能になるといえる．

6. 結論と今後の課題

本研究では，商品の改善や顧客満足度の高い商品の提供につながるような結果を得ることを目的とした，QFS におけるクエリのさらに詳細な情報を抽出する手法を提案した．具体的には，LDA を用いることで，設定したクエリと共起関係にある単語群を抽出し，それらを新たなクエリとして QFS を用いて要約文を生成した．実データを用いた実験では，LDA を用いた単語群の抽出において，設定したクエリに関連する単語や，似たような意味で使われる単語を抽出することが可能であることを示した．また，抽出された単語を新たにクエリとして設定して，QFS により要約を生成することにより，設定したクエリに関するより詳細な情報が得られることが明らかとなった．今後の課題として，他サイトでのレビュー分析の有用性の確認や，定量的な提案手法の評価が挙げられる．

参考文献

[1] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, 2004.

[2] Yllias Chali and Sadid A. Hasan. Query-focused multi-document summarization: Automatic data annotations and supervised learning approaches. *Natural Language Engineering*, Vol. 18, No. 1, pp. 109–145, 2012.

[3] Raghavan H. Cardie C. Wang, L. and V Castelli. Query-focused opinion summarization for user-generated content. *arXiv preprint arXiv:1606.05702*, 2016.

[4] Andrew Ng Blei, David and Michael Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, pp. 993–1022, 2003.

[5] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th international conference on world wide web*, pp. 507–517, 2016.