

ラベルなし対話スクリプトと Next Sentence Prediction を 用いたテキストグループチャットにおける返信関係の判別 Identification of Reply-to Relation in Textual Group Chat via Unlabeled Dialogue Scripts and Next Sentence Prediction

SHAN Junjie^{1*} 西原 陽子² 韓 毅弘³
Junjie Shan¹ Yoko Nishihara² Yihong Han³

¹ 立命館グローバル・イノベーション研究機構

¹ Ritsumeikan Global Innovation Research Organization, Ritsumeikan University

² 立命館大学情報理工学部

² College of Information Science and Engineering, Ritsumeikan University

³ 立命館大学大学院情報理工学研究科

³ Graduate School of Information Science and Engineering, Ritsumeikan University

Abstract: インスタントメッセージ (IM) が他者とのコミュニケーションにおいて重要な役割を果たすようになったことにより、チャットメッセージを分析することで人々のコミュニケーションを支援することを目的とした研究が増え始めている。チャット内では、メッセージ間の関係 (返信関係など) が明示されていない。グループチャットにおけるメッセージ間の関係性を明らかにすることが重要となる。本稿では、ラベル付けされていないテキストデータから、グループチャットにおける「返信関係」(reply-to) を、Next Sentence Prediction (NSP) の手法を用いて判別する手法を提案する。まずは、学習データの準備として、ラベル付けされていない対話スクリプトから、「reply-to」関係を持つメッセージと持たないメッセージを自動的にサンプリングする方法を提案する。次に、準備した学習データを用いて、二つのチャットメッセージ間の「reply-to」関係を判別するための NSP モデルを三つのセッティングとパラメータで構築する。構築する NSP モデルは、日本語の事前訓練 BERT モデルに基づいている。最後に、学習したモデルを手作業でラベルを付けた実際のテキストグループチャットのメッセージにより評価した。構築した三つのモデル評価した結果、検証セットで最大 88.82%、テストセットで 69.64%の精度を得られた。

1 はじめに

在宅ワークやリモート学習の普及が進行している現在、オンラインコミュニケーションツールを用いたコミュニケーションが活発に行われるようになっており、実況配信サイト、テレビ会議システム、音声会話ルームなど、様々なものが提供されている。その中、インスタントメッセージ (IM) ソフトウェアは、オンライン時代のコミュニケーションにとって不可欠な存在となっている。人々の情報交換が IM に依存するようになり、IM 上での人々のコミュニケーション活動を支援することを目的とした研究も多く見られるようになった。例えば、コミュニケーションメッセージを通して人々の関係を分析し、良好な関係の維持に役立つ研究 [1]

や、IM アプリケーション上の雑談の話題をサポートする研究 [2] などがある。

これらの研究はすべて、詳細的な解析や適用をする前に、各チャットメッセージ間の正しい対応を判別する必要がある。特に「返信関係」の判別が重要となる。チャットソフトウェアのテキストメッセージは、文章や新聞記事などのように、順序付けられて構成されていない場合が多い。チャットメッセージはほとんどの場合において、自由な順序で発言されている。図 1 に示すように、最後のメッセージ「私もです」は最初のメッセージ「土曜日のイベント」に対して返信しており、隣接する 2 番目のメッセージに対する返信ではないことが明らかである。

従来の電子メールなどのオンラインコミュニケーション手段に比べ、テキストのグループチャットには以下のような特徴がある：

*連絡先：立命館グローバル・イノベーション研究機構
〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: shan@fc.ritsumei.ac.jp

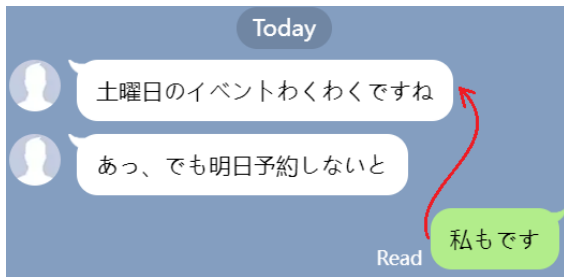


図 1: 自由な順序のチャットメッセージ. 3 番目のメッセージは、隣接の 2 番目メッセージではなく、1 番目のメッセージに対して返信している

- 短文内容で複数回送信
IM のチャットメッセージは、返信速度を上げるために各メッセージの内容は短く書かれることが多い。結果として、IM チャット、特にグループチャットでは、連続しないあるいは順序性のないメッセージが発生しやすくなる。
- ランダム化と断片化
グループチャットの中で、人々は各自のトピックや自分のペースで発言し、チャットのメッセージは不規則な順序や断片的なパターンで流れていく。

これらの課題に取り組むために、グループチャットのメッセージと類似したテキスト上の特徴を持ちながら、順序立てて構成された「対話スクリプト」を学習データとして利用する。

本稿では、BERT モデルの NSP (Next Sentence Prediction) 手法を活用し、グループチャットのテキストメッセージ間の複数の「返信関係」の判別手法を提案する。提案手法では、主に以下の三つの処理を行う：

1. ラベルなしの対話スクリプトのテキストデータから、「返信関係」を持つメッセージペアと持たないメッセージペアを自動的に取得する方法を提案する。
2. 「返信関係」の識別の効果を検証するために、三つの構造で NSP モデルを構築し、訓練とテストを行う。
3. 実際のグループチャットからテキストメッセージの履歴を三つ収集し、各メッセージ間の「返信関係」を人手でラベル付けて提案手法を評価した。

2 関連研究

2018 年に Google が BERT (bidirectional encoder representation from transformers) モデルを提案して以来 [3]、それに関する研究、またそのサブ機能や構成パーツに関する多くの研究が行われている。Next Sentence Prediction (NSP) に関する研究もその一つである。本来の NSP タスクは、BERT モデルの学習を強化するために提案されたものである。NSP タスクは、2 文のペアを受け入れ、そのペアの 2 番目の文が 1 番目の文に対して論理的または意味的に続くかどうかを予測することを学習する。NSP タスクにより、事前訓練した BERT モデルが文ペアの直接入力を受け取る仕組みが提供されたため、自然言語処理の分野で注目されている。2019 年、Shi らは NSP を暗黙の談話関係分類 (implicit discourse relation classification) のタスクに活用し [4]、Liu らは NSP を用いて長文の自動要約を支援する手法を提案した [5]。2021 年に Feng らは、NSP を利用して子供の言語教育における読解問題の難易度を調整することを試みた [6]。Li らは 2022 年にマルチモーダルなエンティティ・リンキングに NSP を利用した [7]。これらの研究は、NSP が文ペアの入力に対する処理能力を持つことを示している。

一方で、テキストチャットの分析に関する研究も幅広く行われている。例えば、チャットメッセージ中の代名詞の共参照解析などを含んでいる「chat disentanglement」はその中の代表的な一つである。このタスクでは、大量の順序が揃っていないチャットメッセージの処理や分析をするために、多数の人手でアノテーションされた特徴を使用する必要がある [8–11]。2017 年、Mehri らは人手でアノテーションされた特徴のあるデータセットを使用して、オンラインのチャットメッセージにおける返信関係の判別を試した [12]。Mehri らは、オンラインのグループチャットから 524 件のメッセージに対して手動で返信関係を表す特徴をアノテーションした。そして、ランダムフォレスト分類器とリカレントニューラルネットワークの 2 つの方法を用いて、返信関係の判別を行った。しかし、このような提案手法は、多くの事前アノテーション作業が必要であるため、チャットメッセージ記録のデータセットが少なくなり、チャットメッセージの内容もより多様なトピックや領域を含みにくくなる。

2019 年に Guo らが最初に人手特徴量を使用しないプレーンテキストのグループチャットメッセージから「返信関係」を判別する手法を発表した [13]。彼らは、グループチャットにおけるメッセージ間の返信関係を入力されたプレーンテキスト情報のみから判別するため、LSTM (Long Short-Term Memory networks) に基づく三つのモデルの提案や検証を行った。その中の一つにエンドツーエンドモデルも含まれる。Guo らが

表 1: 複数の「返信先」のある例。
(新着メッセージ D は、過去メッセージの A、
B、C の全部に対する返信である)

過去メッセージ: A: 次の旅行について検討しよう。 B: 美味しい食べ物のある所がいい C: すてきな景色が見たい
新着メッセージ: D: じゃ、○○○はどうでしょうか？

提案したモデルは、新着のメッセージの「返信先」となる可能性が最も高い一つの過去メッセージを出力する。或いは、新着メッセージが新しいトピックの始まりとして、すべての過去メッセージに返信していない場合には、新着メッセージが自身の「返信先」として判別し、新着メッセージ自身を出力する。Guo らは実際の中国語のテキストグループチャットのデータを用いて提案モデルを検証し、最大 60% に近い精度を得ることができた。しかし、Guo らのモデルは各新着メッセージに対して、「返信関係」のある過去メッセージは一件しか出力できないため、複数の返信先がある場合に対応が難しい。表 1 が示すように、D からの新着メッセージ「じゃ、○○○はどうでしょうか？」の返信先対象は、過去の 3 つのメッセージ全てになると考えられる。この場合、一つのメッセージに対して複数の「返信先」を持つと考えられる、これはグループチャットにおいてよく見られることである。また、Guo らの研究では、同一のデータセットを用いてモデルの訓練と評価を行なっている。検証用のデータセットを新たに用意し、モデルの汎化性能を評価するべきと考えられる。

そこで、NSP を用いて二文のメッセージペアに「返信関係」を持つかどうかを判別する手法を提案する。また、モデルが特定のデータセットに限定されることを避けるため、ラベルなしのデータから自動的にチャットメッセージペアを収集する手法を提案し、多様なトピックのメッセージデータを取得できるようにする。

3 提案手法

3.1 タスク定義

本稿の課題は、テキストのグループチャットにおいて、二つのチャットメッセージが内容的に「返信関係」を持っているかどうかを判別することを目的とする。具体的には、A と B の二つのチャットメッセージが与えられて、メッセージ B がメッセージ A の内容に対する

返信であるかどうかを判別する。内容的にメッセージ B がメッセージ A に返信する確率は、NSP モデルによって式 (1) により計算される：

$$P(\text{reply}(B|A)) = \text{Softmax}(\text{NSP}(A, B)). \quad (1)$$

3.2 対話スクリプトからの「返信関係」メッセージの自動サンプリング

前述したように、一般的なチャットメッセージは、順序と内容のロジックに一貫性がないため、過去の研究で「返信関係」の判別のためのデータセットを準備する時に、各チャットメッセージに対して人手で返信関係を付けることが必要である。本研究では、チャットのテキストメッセージと類似する特徴を持ちながら、かつ順序立てて構成された一つテキストのデータとして映画、ドラマ、舞台劇などの「台本・対話スクリプト」を「返信関係」の学習データとして利用することを提案する。対話スクリプトには以下のような特徴があることを分析・考察した：

- 簡潔で短い。
文章や新聞記事などに書かれる内容と違って、役者の演出や観客への理解を容易にするため、対話スクリプトの台詞は通常に簡明である必要がある。そのため、使用する言葉は十分に考えられ、台詞のテキストは簡潔になっている。そして、対話スクリプトは考えられて作られた文であるため、普通に考えながら喋る話し言葉とも違って、より短くてわかりやすい文面ではっきりと情報を伝えている。これはチャットのテキストメッセージと類似する特徴である。
- 順序的かつ連続的。
チャットメッセージと違い、対話スクリプトは、内容に応じてロジカルな順序で構成する必要がある。そのため、一つの対話シーンにおいて、直後の発話が直前の発話内容に対する「返信」である可能性が高くなる。

以上の理由から、「返信関係」を持つメッセージペアを取得するためのソースデータとして、対話スクリプトを採用する。ポジティブ（「返信関係」あり）とネガティブ（「返信関係」なし）のメッセージペアをサンプリングする時に実行する具体的な方法は以下に示す：

1. 対話スクリプトに含むシーン転換のマークを利用して各対話シーンを分割する。
2. 同じ対話シーンで隣接する二つの対話（メッセージ）はすべて「返信関係」を持つと仮定し、それらを「ポジティブ」のペアとして保存する。

表 2: 収集したデータセットの統計データ.

項目	数値
対話スクリプト件数	5
対話シーンの数	49
総発言数	3,043
一発話の平均長さ (文字)	21.24
ポジティブペアの数	1,698
ネガティブペアの数	3,396

3. 異なる対話シーン或いは異なる対話スクリプトにある二つの発話 (メッセージ) は、「返信関係」を持たないと仮定した. そこで, 異なる台本スクリプトからランダムに二つの発話台詞を選び, 「ネガティブ」のペアとして作成する.

ポジティブペアの信頼性の確保とノイズの削減のために, 本研究では, 話し手が二人しかいない対話スクリプトのみをサンプリングの対象とした. この場合, 話し手が変わるときには, 概ねに前の話し手の台詞に対する「返信」が出てくる.

本稿では, インターネットから日本語の対話スクリプトを五件収集した, 合計 49 の対話シーンがあり, 発話の総数は 3,043 である. 収集したデータセットの統計情報を表 2 に示す. 合計 1,698 組のポジティブの対話ペアが得られた (一部のノイズの入った対話データ (例えば, 沈黙を表す「……。」や表情を説明する「(笑)」などは除去された). 各ポジティブペアに対して二つのネガティブペアを作成し, ネガティブペアの数は 3,396 である.

3.3 NSP モデルの構築

本研究では, 三つの NSP モデルを構築し (Orig-NSP, NSP-FA-1L, NSP-1L), 「返信関係」判別の効果を検証する.

3.3.1 Orig-NSP: オリジナルの NSP モデル

はじめに, 一般的な BERT の構造からオリジナルの NSP モデルを作成し, 検証を行った. 図 2 に示すように, 事前学習した BERT モデルに [A, B] のメッセージペアを入力し, オリジナルの NSP プーリングを使って [A, B] の「返信関係」を判別する.

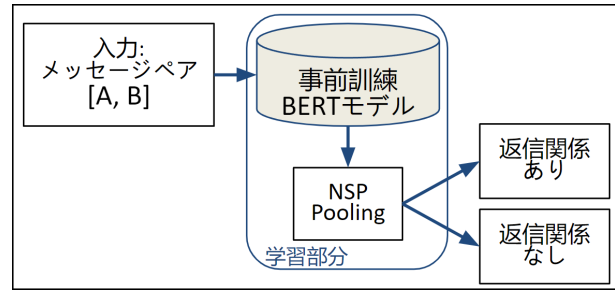


図 2: Orig-NSP: オリジナルの NSP モデル

3.3.2 NSP-FA-1L: NSP モデルでメッセージペアの分散表現を取得

オリジナルの NSP プーリングの後に 128 ユニットを持つ隠れ層を追加し, 隠れ層の出力を使って入力メッセージペアの「返信関係」を判別する. この設定では, 図 3 に示すように, オリジナルの NSP プーリングと事前学習の BERT モデルを凍結し, 追加した隠れ層のみを訓練する. つまり, オリジナルの NSP プーリングからの出力は, 入力メッセージペア [A,B] の分散表現 (embedding) を取得することのみに使用される. この仕組みは既存研究を参考にし導入する [14].

3.3.3 NSP-1L: オリジナルの NSP モデル+一つ隠れ層

「NSP-FA-1L」の構造と同様に, オリジナルの NSP 出力の後に 128 ユニットの隠れ層を追加し, オリジナルの NSP 層と事前学習した BERT モデルも凍結しない. 図 4 に示すように, この設定では, 追加の隠れ層をオリジナルの NSP と BERT モデルとともに, 作成したメッセージペアの訓練データを用いて訓練させる.

訓練の設定を表 3 に, 訓練の結果を表 4 に示す. 訓練データ数は 5,094 件 (ポジティブ 1,698 件+

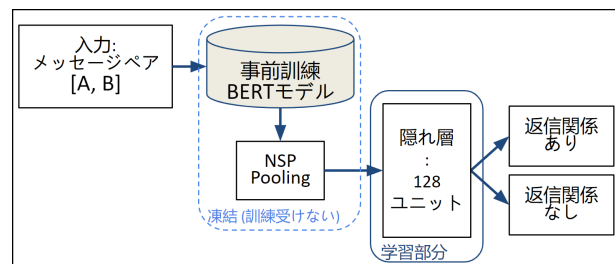


図 3: NSP-FA-1L: NSP モデルでメッセージペアの分散表現を取得.

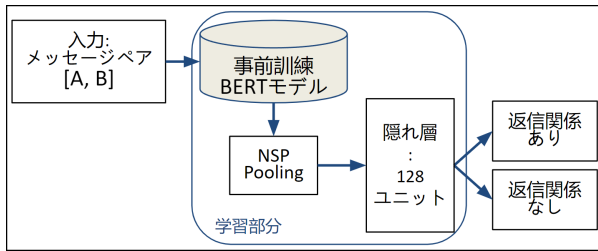


図 4: NSP-1L: オリジナルの NSP モデル+一つ隠れ層.

表 3: 訓練のセッティング.

項目	数値
訓練メッセージペア数	5,094
形態素分割レベル	文字
最大入力長さ	128
Batch サイズ	64
エポック	10
検証セット割合	0.1
学習率	5e-5
最適化アルゴリズム	Adam

ネガティブ 3,396 件) のメッセージペアである。入力するメッセージペアは全部文字レベルに区切り、最大長さは 128 文字 (A と B の合計長さ) である。検証セットのサイズは 10 % である。三つの NSP モデルは、東北大学により公開された事前学習済みの日本語 BERT モデル¹ で初期化した。10 エポックの訓練を行った後、検証セットにおける Orig-NSP, NSP-FA-1L, NSP-1L の精度はそれぞれ 87.06%, 78.24%, 88.82% であった。

4 評価実験

4.1 評価実験手順

実際のグループチャットの記録を用いて、学習したモデルの評価実験を行った。評価実験の手順は以下の通りであった：

1. 三つのグループチャットからチャットメッセージの記録を収集し、各メッセージ間の「返信関係」を手手でアノテーションした。三つのチャットは全て 30 分程度の長さであった。
2. チャット記録中の各メッセージとその前の N 個のメッセージを二つずつペアに組んで、訓練済み

¹<https://github.com/cl-tohoku/bert-japanese/tree/v2.0>

表 4: 訓練結果.

モデル	訓練セットでの損失	訓練セットでの精度	検証セットでの損失	検証セットでの精度
Orig-NSP	0.0389	95.56%	0.6259	87.06%
NSP-FA-1L	0.4974	76.20%	0.4594	78.24%
NSP-1L	0.0807	97.25%	0.3888	88.82%

NSP モデルに入力する。NSP モデルにより各メッセージペアの「返信関係」を予測する。例えば、チャットメッセージの系列が $Seq = [m_{i-N}, \dots, m_{i-2}, m_{i-1}, m_i]$ である場合、与えられたメッセージ m_i に対して、各 $[m_{i-n}, m_i] (n \in [1, N])$ を一つのメッセージペアとして組み合わせる。

3. 人手でアノテーションされたメッセージペアのラベルと NSP モデルの予測結果から、精度を算出する。各メッセージペア $[m_{i-n}, m_i]$ が「返信関係」を持つかどうかについて、事前にチャットの文脈内容によりアノテーションを行った。(m_i が m_{i-n} への返信である場合は 1 を、ではない場合は 0 にする。)

表 5 は、収集された評価実験用のグループチャット記録の統計データである。三つのグループチャットから収集したメッセージの総数は 444 件、各グループのチャットメンバーは 3 名であった。本研究では、手順 2. の N を 5 に設定し、2,052 件のメッセージペアの評価データを取得した (複数「返信先」のペアは 388 件)。 N は、あるメッセージに対して「返信」可能なメッセージの範囲を表す。 N の大きさは、チャットグループのメンバー数と正の相関があると考えられる。チャットグループのメンバーが増えれば増えるほど、二つ「返信関係」のあるメッセージ間の「距離」が長くなる可能性が高くなる。今回収集したグループチャットのメッセージでは、各チャットグループのメンバーが 3 名であり、すべてのメッセージの「返信先」が過去 5 通以内であることが確認された。故に、 N を 5 に設定して訓練済み NSP モデルの評価を行った。

評価結果を表 6 に示す。また、対話スクリプトのデータで訓練を行わないオリジナルの事前学習 BERT モデルの効果も評価した。その結果を「No-Training」として記載する。評価実験の結果により、「NSP-FA-1L」のモデル構築が最も高い精度の 69.64% が得られた。

5 考察

評価実験の結果により、対話スクリプトのデータセットで訓練した三つの NSP モデルは全て、訓練を実行していないオリジナルの事前学習 BERT モデルより精度

表 5: 評価実験のグループチャットの統計データ.

グループ	メッセージ数	メンバー数
グループ 01	237	3
グループ 02	115	3
グループ 03	92	3
合計	444	9

表 6: 評価結果.

モデル	精度
No-Training	49.8%
Orig-NSP	56.63%
NSP-FA-1L	69.64%
NSP-1L	62.77%

が高いことが分かった. この結果から, 対話スクリプトを利用してグループチャットメッセージの「返信関係」を判別する提案手法が有効であることが示された.

提案した三つのモデルは全て, 評価データでの精度が検証セットでの精度より低くなった. これは, 主に訓練データが不足しているためと推測される. 提案手法の試行として, 本研究ではより小規模なデータセットのみを収集してモデル訓練を行った. 学習データには, 1,698 組のポジティブと 3,396 組のネガティブのメッセージペアしか含まれておらず, BERT のような巨大なモデルを十分に学習させるのは困難なものと考えられる. 今後も, より多くの対話スクリプトを収集し, 訓練データの量を増やすことを目指している.

モデル NSP-FA-1L は, 検証セットでの精度は一番低かったが (78.24%), 評価データでは最も高い精度を得た (69.64%). これは, モデル NSP-FA-1L の効果を反映したものではなく, 訓練セットのデータ分布が原因であると考えられる. 訓練セットにはネガティブのペアがポジティブのペアの 2 倍であるので, 隠れ層 1 つだけの訓練では過剰適合になりやすくと考えられる (NSP-FA-1L モデルは元の BERT モデルと NSP の部分を全部凍結している). そのため, NSP-FA-1L モデルは他の二つのモデルに比べて, より多くなネガティブの判別結果が出やすくなる.

NSP-1L モデルの精度は, 検証セットと評価データの両方で Orig-NSP モデルの精度より上回った. このことから, 元の NSP プーリングの後により小さな隠れ層を追加する手法が, モデル訓練の補助に有効であることが示された. BERT モデルにあるオリジナルの 768 個の隠れユニットを持つ NSP プーリング (図 2, 3, 4 にある「NSP Pooling」の部分) のノイズを軽減し, モ

表 7: 正解・不正解の判別結果についての分析.

	一番目メッセージ (A) の平均長さ	二番目メッセージ (B) の平均長さ	入力メッセージペアの平均長さ
正解結果	12.72	14.28	27.0
不正解結果	11.05	9.89	20.95

デルがより特定のタスクに集中できるようになったと考えられる.

検証セットでの精度が最も高い NSP-1L モデルを選択し, それが「返信関係」を正しく判別した/判別しなかったメッセージペアに対して分析した. 表 7 には, 正解と不正解の結果における入力ペアの一番目メッセージ (A) と二番目メッセージ (B) の平均長さ (単位: 文字) をそれぞれ示している. 誤って判別されたペアの二番目メッセージ (B) の平均長さは, 正しく判別されたペアの二番目メッセージより短くなることが確認された (9.89 文字と 14.28 文字). これは, 今回定義したタスクが, 独立した二つのメッセージだけを入力として「返信関係」を判別していて, 文脈情報を含んでいないためと考えた. そのため, 「はい」, 「オッケー」, 「もちろん」, 「賛成」, あるいは「www」などのような, 共通の応答として使える短いメッセージたちは, 前のメッセージに「返信関係」を持つと判別されやすくなると考えられる. この問題を根本的に解決するには, 文脈情報を追加する必要があると考えられ, 今後, この方向でモデルの改善を進めていく予定である.

6 結論

本稿では, ラベル無し対話スクリプトのデータセットと Next Sentence Prediction タスクを用いて, グループチャットのテキストメッセージにおける「返信関係」を判別する手法を提案した. グループチャットにおけるメッセージは, 複数の「返信先」を持つ可能性があると考えられ, それに応じて複数返信先のメッセージを判別できる NSP モデルを活用する手法を提案した. 小規模なデータセットで提案手法を検証した. データセットとしては, 対話スクリプトから作成した訓練データが 5,094 件のメッセージペアと実際のグループチャット記録からの検証データが 444 件である. 三つのセッティングでモデルを構築して検証を行った結果, 検証セットで最大 88.82%, 検証データで 69.64% の精度を得た.

引き続き, 訓練データの収集を続けて, モデルの性能改善を行う予定である.

謝辞

本研究の一部は、立命館グローバル・イノベーション研究機構、第4期拠点形成型 R-GIRO 研究プログラム「心の距離メータ」を用いたフィジカル・サイバー空間における人間関係構築技術の開発」の支援を受けて実施されました。感謝の意を表します。

参考文献

- [1] 西原陽子, 砂山渡, 谷内田正彦: 発話テキストからの人間の仲の良さと上下関係の推定, *電子情報通信学会論文誌*, Vol. 91, No. 1, pp. 78–88 (2008).
- [2] Andalibi., Nazanin., Frank, Bentley., and Katie, Quehl.: Multi-channel topic-based mobile messaging in romantic relationships, *Proceedings of the ACM on Human-Computer Interaction*, 1.CSCW, pp. 1–18 (2017).
- [3] Jacob, Devlin., Ming-Wei, Chang., Kenton, Lee., and Kristina, Toutanova.: Bert: Pre-training of deep bidirectional transformers for language understanding, *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186 (2019).
- [4] Shi, Wei., Vera, Demberg.: Next sentence prediction helps implicit discourse relation classification within and across domains, *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 5790–5796 (2019).
- [5] Liu, Jingyun., Jackie, CK, Cheung., and Annie, Louis.: What comes next? Extractive summarization by next-sentence prediction, arXiv preprint arXiv:1901.03859, (2019).
- [6] J, Feng., J, Mostow.: Towards Difficulty Controllable Selection of Next-Sentence Prediction Questions, *EDM*, (2021).
- [7] L, Li., Q, Wang., B, Zhao., et al.: Pre-Training and Fine-Tuning with Next Sentence Prediction for Multimodal Entity Linking, *Electronics*. Vol. 11, pp. 21–34 (2022).
- [8] Elsner., Micha., and Eugene, Charniak.: Disentangling chat, *Computational Linguistics*, Vol. 36, No. 3, pp. 389–409 (2010).
- [9] Elsner, Micha., Eugene, Charniak.: You talking to me? a corpus and algorithm for conversation disentanglement, *Proceedings of ACL-08: HLT*, pp. 834–842 (2008).
- [10] L, Wang., D, W, Oard.: Context-based message expansion for disentanglement of interleaved text conversations, *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pp. 200–208 (2009).
- [11] J, Kim., W, Lee., et al.: Optimized combinatorial clustering for stochastic processes, *Cluster Computing*, Vol. 20, No. 2, pp. 1135–1148 (2017).
- [12] S, Mehri., G, Carenini.: Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks, *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Vol. 1, pp. 615–623 (2017).
- [13] G, Guo., C, Wang., J, Chen., et al.: Who is answering whom? Finding “Reply-To” relations in group chats with deep bidirectional LSTM networks, *Cluster Computing*, Vol. 22, No. 1, pp. 2089–2100 (2019).
- [14] J, Shan., Y, Nishihara., A, Maeda., and R, Yamanishi.: Question Generation for Reading Comprehension Test Complying with Types of Question, *Journal of Information Science and Engineering*, Vol. 38, No. 3, pp. 571–589 (2022).

新聞記事におけるBERTに基づく意見文 抽出を用いた価値観推定

Human Values Estimation Based on Opinion Extraction Using BERT with News Articles

韓 毅弘^{1*} 西原陽子² SHAN Junjie³
Yihong Han¹ Yoko Nishihara² Junjie Shan³

¹ 立命館大学情報理工学研究科

¹ Graduate School of Information Science and Engineering, Ritsumeikan University

² 立命館大学情報理工学部

² College of Information Science and Engineering, Ritsumeikan University

³ 立命館大学グローバル・イノベーション研究機構

³ Ritsumeikan Global Innovation Research Organization, Ritsumeikan University

Abstract: 本研究は価値観は意見文に反映されると仮定し、意見文抽出に基づく価値観推定手法を提案する。与えられた文書から意見文を抽出し、その意見文が反映する価値観を推定することで、文章全体の価値観分布を示す。提案手法は深層学習の仕組みを利用し、複数の文の連結によるアップサンプリング手法を採用する。毎日新聞の記事コーパスを用いて意見文抽出と価値観推定の実験を行い、意見抽出では92%の精度を得られ、アップサンプリングにより、価値観推定ではベースラインより50%以上の精度向上を確認した。

1 はじめに

Web 2.0 の普及とともに、UGC (User Generated Contents) の時代が幕を開けた。2000 年代以降、インターネットで自分の意見や感想を投稿することは一般的に行われるようになった。SNS (Social Network Service) が発展し、ユーザの興味を研究し、その嗜好に沿ったコンテンツを提供できるよう、あらゆる推薦システムが提案された。これらの技術は確かにユーザにとっての利便性を高め、サービスの体験を向上させた。実際に技術の恩恵を受け、それを支持している人も多いと思われる。

その一方で、「エコーチェンバー現象」という問題が生まれている。エコーチェンバー現象とは、価値観の似た者同士で交流、共感し合うことにより、特定の意見や思想が増幅する現象を指す。ユーザは自分の考えと似ている意見しか見えず、自分と異なる意見の存在を認識できない。その結果、視野が狭められ、社会全体の様子や別の立場の人の訴求が見えなくなる。特に社会的議題に関しては、考え方・価値観の違いで意見が分断され、建設的な議論が行いにくくなることも予想

される。本研究は「エコーチェンバー現象」という、実際に存在している様々な視点からの意見への限定的アクセスを、この問題の核心的問題点として捉える。このための第一歩として、テキストデータから人間の価値観の推定手法を検討する。

本研究はテキストデータから書き手の価値観を推定する手法を提案する。価値観を反映する文書のサンプルとして、新聞の社説を利用する。提案手法は「価値観は意見文の中に潜んでいる」という仮説に基づき、入力文書から意見文を抽出し、抽出された意見文をもとに、文書全体の価値観を推定する。本研究の主な成果は以下の2点になる：

1. 与えられた文書中の意見文を抽出し、各意見文の価値観を推定することで、文書全体の価値観を推定する手法を提案する。
2. 文書中の意見文を組み合わせ、連結することにより、ある価値観の意見文の訓練データ数を増やす。

2 関連研究

意見抽出及び価値観推定に関する既存研究を紹介する。

*連絡先：立命館大学情報理工学研究科
〒525-8577 滋賀県草津市野路東1丁目1-1
E-mail: is0387ps@ed.ritsumei.ac.jp

2.1 意見抽出

意見抽出は、コメント、レビューなどのテキストデータから、人々の観点や意見を探索する自然言語処理技術であり、あるトピック（政策、製品、サービスなど）に対するユーザの態度を把握することを目的とする [1]。意見抽出に関する先行研究の多くは、主観的な評価表現に着目している [2]。意見表現は常に対象、属性、評価の3つの要素で構成されるため、一般的な自然言語処理では、「意見表現」という概念は「評価表現」にと記載されることが多い [3]。「この車のエンジンは良い」という意見文では、「車」が対象で、「エンジン」が属性、「良い」という形容詞が評価となる。また、客観的な記述から評価情報を抽出しようとする研究もある [4]。主観的な表現により記述された評価情報だけではなく、客観的な表現により記述された評価情報も評価表現として扱われる [4]。川田ら [5] は、Web テキストから評価情報を収集・分類し、評価表現のラベル付きコーパスを構築した。中川ら [6] は、川田らの研究をもとに、主観的評価表現と客観的評価表現の抽出手法を提案した。中川らは、評価表現を「当為」「要望」「感情」「批評」「採否」「出来事」という、7つのカテゴリーに分ける評価表現分類法を定義した。

本研究でも同様に意見文抽出を行うが、評価表現の抽出などは行わず、文が意見文であるか非意見文であるかの判断のみを行う。意見文の抽出には、機械学習手法の一種であるニューラルネットワークを用いる。

2.2 価値観推定

人間の価値観推定は、心理学・社会学・人類学・情報科学など幅広い分野で急速に発展しているトピックであり、主な課題は価値観カテゴリーの定義と、与えられたソースデータからの価値観検出である。Chengら [7] は人間の価値観のメタリストを提案した。彼らは価値観に関する既存研究をレビューし、人間の価値観の定義を作成した。さらに12の価値観リストをレビューし、人間の価値観のメタリストを作成した。石田ら [8] は、人間のアノテーターによる効率的な価値観アノテーションのアプローチを提案した。彼らはアノテーションの作成方法として、アノテーターが協力して各文書に対して最適なアノテーションを作成する方法と、各アノテーターがそれぞれ最大限の文書数をアノテーションする方法の2つを考案した。Kristofferら [9] は第4回 European Social Study (ESS-4) の調査データを分析するための、潜在クラス分析に基づく人間の価値観推定手法を提案した。Kristofferらの研究は、調査結果に隠された人間の価値観要素を探ることに重点を置いている。この推定手法はESS-4のオリジナルデータにどのくらい適合しているかを評価ポイン

トにしている。高山ら [10] は、単語ベースの確率的潜在変数モデルを用いて人間の価値観を検出する手法を提案した。この手法は、学習データの各単語に6種類の価値観を表す6次元ベクトルを付与し、潜在変数モデルを用いて与えられた文章の価値観を推定するものである。

本研究では、単語ベースではなく、文ベースの価値観推定手法を提案する。単語ベースのもの比べ、文ベースの処理は文脈情報を利用できる。また、提案手法は文書ベースのアノテーションを利用することにより、各単語の価値観辞書の作成や、各単語のアノテーション作業は省かれる。これにより、アノテーション作業が軽減され、より効率的なモデル構築につながる。

3 提案手法

提案手法について説明する。提案手法は、与えられた文書に含まれる人間の価値観を推定する。図1は提案手法の処理の流れを示す。

価値観の推定のために、本研究では1つの仮説を置く。ある文書に1つの価値観が反映されているとき、文書内の意見文には同一の価値観が反映されているとする。文書内には意見文と非意見文があり、意見文には書き手の価値観が反映されており、非意見文には価値観は反映されていないとする。非意見文は事実や真実を述べている文を指す。例えば意見文「原子力の安全は、何よりも優先される」があるとき、ここには「Human Welfare」という価値観が反映されており、非意見文「原発への依存の見直しという観点加わった」には価値観は反映されていないとする。

この仮説に基づき、提案手法では、価値観推定の前処理として、入力文書から意見文の抽出を行う。

1. 提案手法は文書が入力として与えられる。文書は意見抽出器に送られる
2. Step.1: 意見抽出器は入力された文書から意見文を抽出する
3. Step.2: 価値観推定器は各意見文の価値観を推定する
4. 各意見文の価値観推定結果の分布を文書全体の価値観推定結果とする

本研究では2つの機械学習モデルを構築し、それぞれ意見文抽出と価値観推定のタスクに当てる。2つのモデルともにBERT 言語モデル [11] と Transformers フレームワーク [12] により訓練された。本研究では、東北大学が公開している pre-trained モデル¹と、石田ら [13]

¹<https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking>

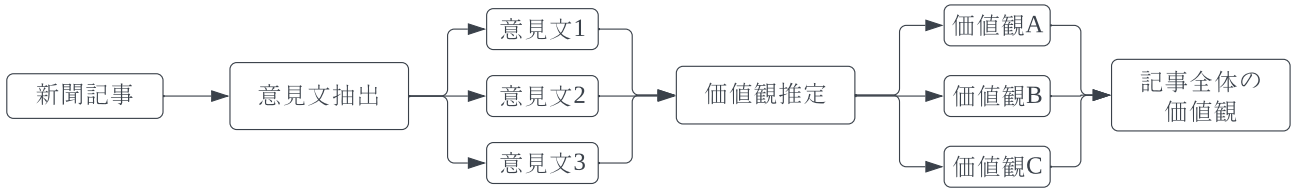


図 1: 提案手法の処理の流れ

が作成した価値観ラベル付き新聞記事データセットが利用されている。新聞記事コーパスは毎日新聞社 2011 年の原子力発電に関する社説記事を利用している。

3.1 意見文抽出

入力された文書から意見文を抽出する。意見文の抽出は機械学習モデルを作成し、モデルで文を意見文と非意見文に分類することにより行う。

機械学習を行うために、訓練用のデータセットを作成する。前述した中川ら [6] が提案した意見文の定義を参考に、毎日新聞の記事データを用いて、意見文と非意見文の訓練用データセットを構築した。30 件余りの記事に含まれた文を手作業でラベルづけし、224 文の意見文と 458 文の非意見文を訓練用データセットとして用意した。各文を BERT の日本語モデルでトークン化し、ベクトルとして表現する。トークン数の上限は 128 とする。ベクトルをニューラルネットワークに入力し、学習を行いモデルを作成する。

3.2 価値観推定

意見文の価値観を推定する。価値観の推定も機械学習モデルを作成し、推定を行う。

文書に 1 つの価値観が反映されているとする。一般に、ある文書に反映される価値観を推定することは複数の意見文を参照することにより達成可能だが、一つの意見文からその文書に反映される価値観を推定することは難しいと考えられる。そこで、機械学習を行う際に、各意見文の価値観を推定するのではなく、同一文書に含まれる意見文の組み合わせに対し価値観を推定するアプローチを取る。これにより 2 つの利点が生まれる。1 つは価値観の推定に用いることができる情報量が増えることである。1 つの文よりも 2 つ以上の文が与えられる方が、学習できる特徴が増える。もう 1 つは、モデルを構築する学習データが増えることである。文書中に n 個の意見文があり、2 文を組み合わせるならば学習できるデータの数 $n(n-1)/2$ となり元の n よりも多くなる。本論文では同一文書に含まれる意見文を組み合わせる学習することにより価値観の推定を行う。

本論文では、組み合わせる文数のことを「拡張レベル」と呼ぶ。文書を D 、文書内の意見文の数を $n(D)$ とする。拡張レベルを L_v とするとき、文書 D 内での意見文の組み合わせ数は $n(D)C_{L_v}$ となる。さらに、訓練データの構築に N 個の文書が使われるならば、最終的に作成される訓練用データセット内のサンプル数 $Dataset$ は式 (1) により与えられる。

$$Dataset = \sum_{i=1}^N (n(D_i)C_{L_v}) \quad (1)$$

4 評価実験

2 つの評価実験の実施方法と実験結果について説明する。1 つ目は意見文抽出の性能を評価する実験、2 つ目の実験は価値観推定の性能について評価する。

4.1 実験 1：意見文抽出

前述した意見文抽出手法の評価実験を行う。

4.1.1 実験準備及び利用するデータ

本研究では、毎日新聞社の 2011 年版の新聞記事から原子力発電に関する記事に価値観がラベリングされたデータセットを使用した [13]。このデータベースは 700 件のニュース記事と 234 件の社説記事の価値観ラベルが含まれる。ニュース記事には 2 人のアナテーターにより複数の価値観ラベルが付与されている。社説記事は 1 つの価値観ラベルが付与されている。社説記事の性質上、作者の評論や意見が多く述べられていることと、シングルラベル分類は比較的安定した分類性能を達成しやすいといった観点から、本論文では主に社説記事を使用した。

意見文抽出の段階では、価値観ラベルには触れず、テキストデータだけを利用している。我々は 234 件の社説記事のうち 23 件をランダムで選び、その 23 件の記事の中にある全部の文を意見文・非意見文で分類し、学習用のデータセットを作成した。学習データセットには意見文 224 件、非意見文 458 件が含まれている。検証

表 1: 学習用・検証用データセットにおける意見文・非意見文の内訳

	意見文	非意見文
学習用データセット	224	458
検証用データセット	165	272

表 2: 訓練パラメータ

パラメータ	値
num labels	2
max length	128
max epochs	100
training set ratio	80%
validation set ratio	10%
test set ratio	10%

用のデータセットにおいては、同じ方法で、学習データとは別のデータセットを作成した。学習用データセットと検証用データセットの作成の際に使用された記事は異なっている。検証データセットには意見文 165 件、非意見文 272 件が含まれている。2つのデータセットの内訳は表 1 に示す。訓練時のパラメータの詳細は表 2 に示す。

モデルの性能は適合率 (式 (2)), 再現率 (式 (3)), F_1 値 (式 (4)) により評価する。

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = \frac{2Precision \times Recall}{Precision + Recall} \quad (4)$$

4.1.2 実験結果

モデル訓練時の validation で得られた精度は最高 96% であった。テストデータで性能評価を行った結果を表 3 に示す。前述した計算法で計算した適合率, 再現率, F_1 値はそれぞれ 83.7%, 87.3%, 85.5% であった。この結果から、提案手法は意見文抽出を十分な精度で行えることがわかった。

4.2 実験 2: 価値観推定

前述した価値観推定手法の評価実験を行う。

表 3: 意見文抽出の評価実験結果

分類結果	意見文	非意見文
アノテーション		
意見文	144	21
非意見文	28	244

表 4: 価値観カテゴリと各価値観の記事数 (降順)

価値観カテゴリ	記事数
Human welfare	57
Effectiveness	56
Importance	48
Power	33
Law and order	17
Wealth	7
Innovation	6
Nature	4
Personal welfare	3
Other	3
合計	234

4.2.1 実験準備及び利用するデータ

実験 2 は実験 1 と同じデータセットを使用した。実験 2 ではデータセットのテキストデータと価値観ラベルの両方を利用する。利用したデータセットには 10 種類の価値観ラベルが存在している、表 4 はその 10 種類の価値観及び各価値観にタグ付けされた記事数を示す。

学習を十分に行うために、記事数上位 5 位の価値観を選び、合計 211 件記事を意見文抽出モデルに与えた。意見文抽出を経て、非意見文は削除され、もとの記事についての価値観ラベルを各意見文に割り当て、意見文と価値観ラベルで構成される訓練データセットを作成した。訓練データには 2,106 文が含まれている、表 5 は各価値観カテゴリの意見文数を表している。

1つの意見文を1つの入力にする(拡張率=1)のパターンと、2つの意見文を1つの入力にする(拡張率=2)パターンをそれぞれ用意し、複数の意見文を連結するという提案手法のアプローチの効果を評価するための実験を行った。訓練時のパラメータは表 6 に示す。

訓練時の入力とは別の、推定時の入力にも意見文の連結による効果を試した。そして、異なる入力長で訓練されたモデルの推定性能を評価するため、2文1入力と1文1入力での価値観推定を行った。テストは合計3つのケースで行われた。

表 5: 訓練データにおける各価値観カテゴリの意見文数

価値観カテゴリ	意見文数
Human welfare	514
Effectiveness	594
Importance	449
Power	375
Law and order	174

表 6: 訓練パラメータ

パラメータ	値
num labels	5
max length	128
max epochs	100
training set ratio	80%
validation set ratio	10%
test set ratio	10%

4.2.2 実験結果

3つのケースの実験結果を表7に示す。

5 考察

5.1 意見文抽出の性能

意見文抽出は学習データが682件であり、データ量は多くなかったが、分類精度は96%であり、有用性が示された。考えられる主な理由は、日本語の意見文には文法的な特徴が現れやすいことである。例えば、分類上「当為」に属する意見文の多くは「○○べき」「○○すべき」という形式で書かれている。「要望」に属する意見文は「○○してほしい」文法が使われている場合が多い。こういった文法的特徴の存在は意見文を発見しやすくしている。そして、今回使用している新聞記事のコーパスも意見文抽出の難易度を下げていること

表 7: 各テストパターンの精度結果

推定時の文数	訓練時の文数	
	1	2
1	40.5%	97%
2	-	99%

が考えられる。普段の話し言葉と違い、新聞記事は言葉使いや文法選びに一定のルールをかけている。こういう表現上の制限が存在しているため、テキスト的な多様性は狭められ、意見表現をより目立たせたと考えている。

5.2 価値観推定における意見文の連結の効果についての検証

表7から、訓練時及び推定時に意見文を連結して入力するという提案手法のアプローチの有効性が示された。訓練時に意見文を連結すると、価値観の推定精度は40.5%から99%に向上した。このことから意見文を組み合わせ、連結させて学習データを増やすことにより、推定精度を高めることができるとわかった。これは複数の意見文を1つの入力にまとめることにより、入力の情報密度が上昇し、ニューラルネットワークにより多い学習材料を提供し、モデルの性能が上がったためと考えている。

2文を連結して学習し、1文の価値観を推定するときも、精度が97%から99%に上昇し、意見文を連結する効果が確認できる。テキストデータから人間の価値観を推定する場合に、2つの文を連結させて学習データ数を増やすことの効果を確認された。

6 終わりに

本研究では、意見文抽出を用いた価値観推定手法を提案した。人間の価値観は意見文に最も反映されると仮定し、意見文から価値観を推定するアプローチを考案した。提案手法は、与えられた文書に対して、テキストから意見文を抽出し、その意見文に含まれる価値観を推定する。文を意見文と非意見文に分類することにより、意見文抽出を行う。価値観の推定では、提案手法は複数の意見文を選択し、それらを連結することで入力テキストの情報量を増やす。各意見文から得られる価値観の分布は、その文書の総合的価値観を表す。

意見文抽出と価値観推定について、2つの評価実験を行った。毎日新聞の社説記事をコーパスデータとして使用した。意見文抽出の精度は92%であった。また、評価実験では、 F_1 スコアが85.5%であった。この結果から、意見文抽出が行えることがわかった。価値観推定の実験では、人間の価値観の推定性能を評価した。訓練時と推定時に複数の文を連結し入力テキストの情報量を増やすことで、精度が向上した。この結果から、提案手法により文の価値観推定が行えることがわかった。

今後の課題として、提案手法をSNSの投稿やネットニュースのコメントに応用し、価値観推定の活用範囲を広めることを計画している。

謝辞

本研究で利用した新聞記事の価値観アノテーションデータは高山泰博先生のチームが作成し、公開したものである。私達はそのデータの詳細について何度もメールで問い合わせしましたが、高山先生は迅速な対応で有益な情報を教えていただきました。高山先生の素晴らしい仕事と親切なサポートに感謝申し上げます。

本研究の一部は立命館グローバル・イノベーション研究機構によって支援されています。

参考文献

- [1] S. Chandrakala and C. S. Sindhu, “Opinion mining and sentiment classification: A survey,” 10 2012.
- [2] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, “Learning subjective language,” *Comput. Linguist.*, vol. 30, no. 3, p. 277 – 308, sep 2004. [Online]. Available: <https://doi.org/10.1162/0891201041850885>
- [3] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, and 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.
- [4] 中川哲治, 宮森恒, 赤峯亨, 乾健太郎, and 黒橋禎夫, “Web 上の客観的記述からの評価情報抽出に関する技術的検討,” *NLP2008 Archives*, 2008.
- [5] 川田拓也, 中川哲治, 森井律子, 宮森恒, 赤峯亨, 乾健太郎, 黒橋禎夫, and 木俣豊, “Web テキストにおける評価情報の整理・分類およびタグ付きコーパスの構築,” *NLP2008 Archives*, 2008.
- [6] 中川哲治, 川田拓也, 乾健太郎, and 黒橋禎夫, “Web 文書からの主観的・客観的評価表現の抽出,” in *電子情報通信学会講演論文集*. 電子情報通信学会, 2008.
- [7] A.-S. Cheng and K. R. Fleischmann, “Developing a meta-inventory of human values,” *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–10, 2010.
- [8] E. Ishita, S. Fukuda, T. Oga, D. W. Oard, K. R. Fleischmann, Y. Tomiura, and A.-S. Cheng, “Toward three-stage automation of annotation for human values,” in *Information in Contemporary Society*, N. G. Taylor, C. Christian-Lamb, M. H. Martin, and B. Nardi, Eds. Cham: Springer International Publishing, 2019, pp. 188–199.
- [9] K. J. Albers, M. Mørup, M. N. Schmidt, and F. K. Glückstad, “Predictive evaluation of human value segmentations,” *The Journal of Mathematical Sociology*, vol. 46, no. 1, pp. 28–55, 2022. [Online]. Available: <https://doi.org/10.1080/0022250X.2020.1811277>
- [10] Y. Takayama, Y. Tomiura, E. Ishita, D. W. Oard, K. R. Fleischmann, and A.-S. Cheng, “A word-scale probabilistic latent variable model for detecting human values,” ser. CIKM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1489 – 1498. [Online]. Available: <https://doi.org/10.1145/2661829.2661966>
- [11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [13] E. Ishita, T. Oga, A.-S. Cheng, K. Fleischmann, Y. Takayama, D. Oard, and Y. Tomiura, “Toward automating detection of human values in the nuclear power debate,” *Proceedings of the Association for Information Science and Technology*, vol. 54, pp. 714–715, 01 2017.

ARデバイスを用いたプレゼンテーションのシナリオ構成ツール

Scenario composition tool for presentations using AR devices

植田 惇也^{1*} 今井 克暢¹
Atsuya Ueta¹ Katsunobu Imai¹

¹ 広島大学 大学院 先進理工系科学研究科

¹ Graduate School of Advanced Science and Engineering, Hiroshima University

Abstract: With the outbreak of the new coronavirus, the format of presentations has changed from face-to-face to online. We have proposed a presentation tool using AR glasses as a presentation support tool for such occasions. This tool has a scenario selection function and can change slides during the presentation according to the remaining presentation time. In this study, based on this tool, we have made it possible to intuitively use related slides in the presentation by displaying them in the AR space.

1 はじめに

近年の新型コロナウイルスの流行以降、プレゼンテーションの形態は大きく変化した。対面のみでの発表は減り、オンラインの場でリアルタイムでの発表や、収録済のビデオによる発表が増えている。このようなプレゼンテーションの場では、音声や画面の共有などの機材の設定、発表時間や発表内容に関する事など、発表者が考えることが多い。こういった場でのプレゼンテーションを支援する方法としてAR(Augmented Reality)を用いることが有効的であると考えられる。発表時間、スライドの構成、話す内容といった発表をする際に必要な情報を常に表示しておくことができれば発表者の負担は減らすことができる。

これらの情報をPCのディスプレイに表示しようとすると、表示サイズが小さくなってしまふ。また、ディスプレイを増やす手段をとっても、一目ですべての情報が得られないといったデメリットがある。このような問題はARを用いることで解消できる。ARでは空間に必要な情報を表示することで、情報が一目でわかる。さらに、ARにはPCのディスプレイでは表現しづらい3次元空間で表示できるという利点もある。

ARに着目したプレゼンテーションツールの研究は、視聴者に対して3Dの物体を見せることを目的としたARの利用がほとんどで、発表者を支援するという目的でARを利用しているものはほとんどない。プレゼンテーションスライドの作成を支援するツールの研究はある[2]が、発表者がプレゼンテーションを行う際に支援するという点では次のような研究がされている。

われわれの所属する研究室ではMixed Reality (MR)

ツールであるHoloLensを用いたプレゼンテーションツールの研究を行っている[1]。このツールはHoloLensを用いることで、AR空間に発表スライド、操作パネル、発表シナリオを配置し、あらかじめ録っておいた音声や動画を流してプレゼンテーションを行うことができる。また、シナリオをあらかじめ何パターンか用意して空間に表示しておき、残りの発表時間に応じてスライドを選択することで発表時間の調整を行うことができる。

この機能は便利だが、実際の発表の場では発表スライドとは別に、関連する予備のスライドも必要になる。このとき、スライドの枚数が増えると質疑応答の時に即座に対応するスライドを探すことが難しくなることが考えられる。このような事態を避けるために、本研究では発表スライドを構造化し、空間に表示することで、直観的にスライド同士の関連性を理解することができるツールの作成を目指す。

2 事前知識

ARやVR、MRはビジュアルライゼーションのための技術であり、プレゼンツールとしての特徴を持つてはいるものの、発表者側を支援する研究は少なく、視聴者側が視聴するためのものがほとんどである。本研究では発表者がHoloLensを装着し、視聴者側は発表内容をPC、スマートフォン上等のスクリーンで視聴することを前提としている。

*連絡先: m215670@hiroshima-u.ac.jp

2.1 HoloLens

HoloLens は Microsoft が開発したヘッドマウントディスプレイ型の MR デバイスである。HoloLens は現実環境をセンサーで認識し、仮想環境のオブジェクトに反映する。HoloLens2 には精度の高いハンドトラッキング、アイトラッキング機能が搭載されており、プログラムを快適に操作できる [3]。



図 1: HoloLens2

3 先行研究

先行研究では HoloLens2 を用いた非同期発表が可能なプレゼンテーションツールを作成している。このツールは Unity 上でプレゼンテーションツールを構成し、スライドの作成等はすべて Mathematica 上で行っている。Mathematica には UnityLink と呼ばれる Unity を制御できるライブラリが備わっており、Mathematica 内から Unity のオブジェクトやコンポーネントにアクセスできる。これに加えて、Unity Editor 上のデバッグ再生を HoloLens で行う Holographic Remoting という機能を用いることで、Mathematica でのリアルタイムな内容変更を発表中に HoloLens に反映させるシステムを実現している [4]。

スライドセットは Mathematica のグラフオブジェクトで表現され、グラフの各ノードには、スライドと複数の音声トラック、動画クリップ、3D オブジェクト等を割り当てられる。

プレゼンテーションツールは Mathematica 側で操作する機能と、HoloLens 側で操作する機能に分類される。

Mathematica 側では発表構成や収録に関する以下の機能を搭載している。

- スライドの登録・変更
- 発表に用いる 3D オブジェクトの構成と転送
- 発表音声の収録・修正・変更

- 発表シナリオの設定

HoloLens 側では発表に関する以下の機能を搭載している。

- スライドの再生・停止・一時停止
- スライド選択
- 発表シナリオの選択
- 3D オブジェクトの表示

発表シナリオの選択は、発表を時間内に収めるために、スライドを省略する場合に用いる。予め発表スライドのシナリオを何パターンか用意しておき、各シナリオに必要な時間をリアルタイムで概算し、残りの発表時間を見てシナリオを変更できる機能である。この機能は発表シナリオを有向グラフ(シナリオグラフと定義されている。)として Mathematica に設定し、シナリオの経路を計算して Unity に送出することで実現している。

図 2 は実際に Mathematica にシナリオグラフを設定し、全ての発表経路を表している図である。1 から 7 のスライドに至るまでの全経路を算出している。

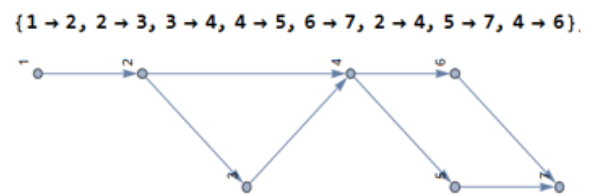


図 2: シナリオ経路の探索

また、図 3 はシナリオグラフを実際に HoloLens 上で表示している様子である。Mathematica 上で作成したシナリオグラフが表示されており、5 枚目のスライドから分岐していることがわかる。

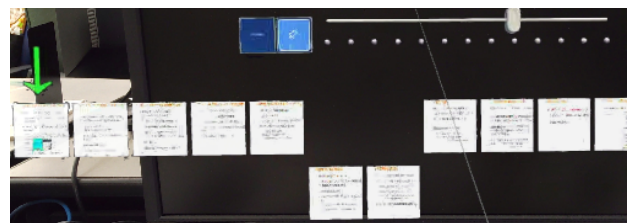


図 3: スライドグラフの表示

図 4 はメインメニューであり、各ボタンは以下の役割を持つ。

Start 現在のスライドの最初から音声を再生する

Pause 音声を一時停止し、再び押すと途中から再開する

Reset 音声を停止し、経過時間をリセットする

SlideSelect 現在のスライドからのシナリオを選択する



図 4: メインメニュー

また、手のひらには表示メニュー（図5）が配置されており、メインメニューやスライドのサムネイル、設定した 3D オブジェクトの表示を切り替えることができる。さらに、発表者の視点と連動した映像や任意の位置のカメラからの映像を PC スクリーン上に流すことも可能となっている。

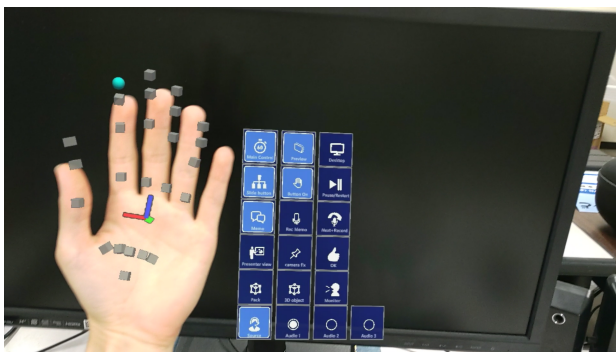


図 5: 表示メニュー

3.1 動作例

図 6 はプレゼンテーションツールの動作例である。UnityEditor 上にあるメインメニューやスライドグラフ等の各種情報が HoloLens を通じて空間に表示されており、PC 画面上には視聴者と共有している画面が映っている。緑の矢印は現在説明中のスライドを指しており、赤く強調されている部分は選択された経路を表している。

また、スライドグラフの側には各スライドに紐づけられた音声の長さを基に、そのスライドまでの経過予

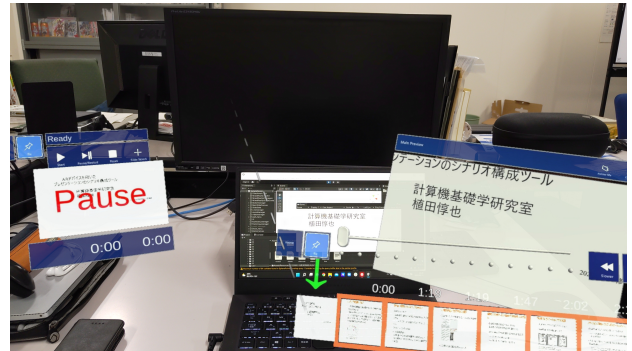


図 6: 動作例

測時間が表示されている。発表者はこの表示と残り時間から判断してシナリオを選択できる。

図 7 は Mathematica から送ったグラフを表示した例である。このように Mathematica で作成したグラフを簡単に表示することができる。

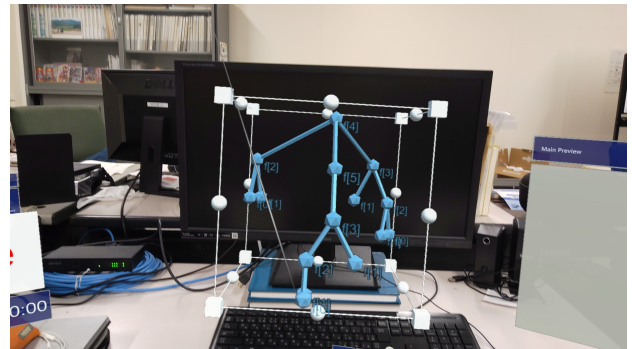


図 7: グラフの表示

4 シナリオ構成の支援

先行研究のプレゼンテーションツールは発表時間に応じた動的なスライド経路の選択を可能にしているが、実際の発表の場においては、以下に挙げるように状況に応じて使用するスライドを変更する必要がある。

- 質問対応時に発表には利用しない別のスライドを用いて説明したほうが良い場合
- 発表会場において想定していた視聴者とはことなる視聴者に対して急遽発表シナリオを変えねばならない場合
- 時間設定のないディスカッションのような場で視聴者からのリアクションによって動的にシナリオを変更していく場合

このような場合には、前節のシナリオグラフで対応することは困難である。

そこで、われわれは本来発表に利用する予定はないが、関連するスライドを用意し、状況に応じて適宜利用できるようにした。

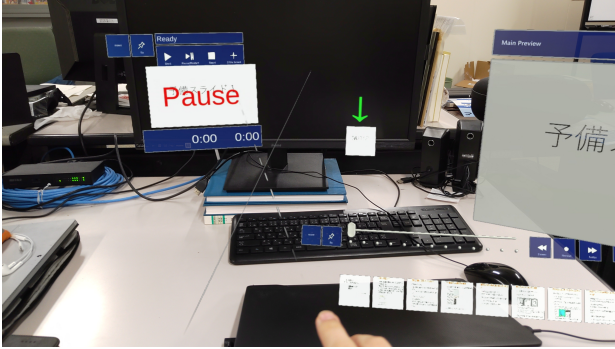


図 8: 関連スライドの表示

図 8 は発表スライドグラフとは別に上部に関連スライドを表示している。このようにシナリオグラフとは別に関連するスライドを用意しておき、状況に応じて画面に表示する。これらのスライドは下にあるスライド同様にボタン操作で表示することができる。

また、この状態ではこのスライドがどの発表スライドと関連しているかがわかりづらいため、スライド同士を結び付けることでどのスライドと関係があるのかをわかるようにした。こうすることで、発表スライドだけでは説明が不十分な箇所の質問が来た時でも即座に予備のスライドを表示することができる。スライド同士を線で結ぶ方法として、スライドグラフの機能を利用した。作成したシナリオ経路はピンク色の線で結ばれて表示される。予備スライドに対するシナリオ経路を追加し、表示位置を調整することによって、関連するスライド同士を線で結ぶことを可能にした。



図 9: 関連するスライドの紐づけ

図 9 は Unity 内にあるオブジェクトの画像で、発表中に空間に表示されるものである。発表スライド 10 枚と関連スライド 3 枚が表示されている。発表スライド

の 5,6,7 番目のスライドとそれぞれ関連するスライドがピンクの線で結ばれていることが確認できる。

5 まとめ

HoloLens を用いたプレゼンテーションツールを使用し、発表スライドとは別に関連スライドを表示することで、発表状況に応じて柔軟に関連スライドを表示できるようにした。また、関連スライドと発表スライドを線で結んで AR 空間に表示することで直観的に関連するスライド同士をわかるようにした。

今後の課題は、シナリオグラフを用いない方法で関連するスライド同士を結ぶことだ。現在のシナリオグラフを用いた方法では、発表中には使用しない余計なシナリオを作成することになってしまう。シナリオが増えてしまうと、シナリオ選択を行う際に悪影響が出る可能性がある。そのため、シナリオグラフとは別の方法でスライド同士を結ぶ方法を考えたい。

参考文献

- [1] 森本章弘, 今井克暢, Mixed Reality ヘッドマウントディスプレイを用いた非同期プレゼンテーションツールの作成, 電気・情報関連学会中国支部連合大会, 2021.
- [2] 三橋郁, 丸山智章, 竹下浩, AR 技術を用いたプレゼンテーション作成手法の開発, 電気学会研究会, 2021.
- [3] Microsoft HoloLens2,
<https://www.microsoft.com/ja-jp/hololens/>
- [4] 中原 良真, コン ギルタク, 今井 克暢, “AR デバイスによるセル・オートマトン遷移の可視化,” 日本応用数理学会 2020 年会, pp.1-2, 2020.
- [5] 今井克暢, 認知機能低下を支援するという視点から見たプレゼンテーションツール, 電気・情報関連学会北陸支部連合大会, pp.1-2, 2022.

灯油タンク内の液面高の計測における レーザセンサデバイスの故障検知

Failure detection of a laser sensor device in the measurement of liquid level height in a kerosene tank

三浦颯太^{1*} 横山想一郎² 山下倫央² 川村秀憲² 多田満朗³
Sota Miura¹, Soichiro Yokoyama², Tomohisa Yamashita², Hidenori Kawamura²,
and Mitsuo Tada³

¹ 北海道大学 大学院情報科学院

¹ Graduate School of Information Science and Technology, Hokkaido University

² 北海道大学 大学院情報科学研究院

² Faculty of Information Science and Technology, Hokkaido University

³ ゼロスペック株式会社

³ ZeroSpec Inc.

Abstract: In recent years, IoT laser sensors have been developed and installed on household kerosene tanks to measure the amount of remaining kerosene in order to improve the efficiency of kerosene delivery. For efficient operation of the laser sensor, this study proposes a method for estimating the refueling time of kerosene tanks using laser sensor measurements, a method for detecting failures of the laser sensor itself, and a method for detecting abnormal fluctuations in the measured values. Experiments using actual laser sensor measurements confirm that the three proposed methods are effective and accurate in practical use.

1 はじめに

1.1 研究背景

灯油配送は主に寒冷地で行われている業務である。寒冷地の住宅には冬場の暖房やロードヒーティングに使用する灯油を貯蔵するタンクが設置されており、定期的に配送を受けることで燃料を切らさずに暖房などを使用し続けることができる。通常、灯油配送のタイミングや配送量は配送業者側が灯油タンクを空にしないように決めていく。

灯油配送業を取り巻く問題としては、近年人手不足によって、一人当たりの労働負荷が上昇していることが挙げられている。この労働負荷、および、労働時間の増加がさらなる呼び水となり、離職率の増加や、灯油配送遅れによる灯油切れが生じてしまう危険性がある。灯油切れにより、冬場に暖房が使用できず室温が下がることで、居住者が生活を行う上での支障を来す可能性がある。また、ロードヒーティングが行えな

くなることで道路が凍結するため、発生する事故の増加も考えられる。さらに、一度灯油を切らしてしまうと、灯油用のホース内に空気が詰まってしまう、それを取り除くために30分程度の時間を要するため、灯油を切らさないための配送計画の作成が重要となる [1]

。配送計画を作成するうえで、タンク内の灯油の在庫管理が重要である。従来は顧客の過去の消費傾向からタンク内の残量が無くなるように配送頻度を決定し周期的に配送していた。また、機械学習を使い過去の消費傾向を学習して灯油タンクの現在の残量推定を行う研究も行われているが、新規の顧客や、消費傾向が変化する顧客に対する残量推定は難しい [2]。そこで近年、リアルタイムで灯油残量の計測が可能なIoTレーザセンサの開発及び設置が進んでいる。しかし、センサは各家庭の灯油タンクに設置しなければならないため、コストの制約から安価なものを使用しなければならない。また、通常屋外にある灯油タンクにセンサを設置するため、夏は炎天下、冬は極寒に晒されるなど計測環境が悪い。このような要因からセンサが故障したり、異常値を計測してしまうなどの事例が発生するため、計測値を配送計画に使用するには信頼度が低

*連絡先：北海道大学 大学院情報科学院
〒060-0814 北海道札幌市北区北14条西9丁目
E-mail: miurasota@ist.hokudai.ac.jp

いのが現状である。

1.2 研究目的

本論文では、灯油配送計画のための信頼性の高い残量把握における、故障センサへの適切な対応を支援するシステムを開発することを目的し、システム内で使用する異常検知手法を提案する。具体的には、センサの計測値と給油時刻データから計算可能な特徴量を用いた、分類器ベースの異常検知手法の検証を行う。特徴量には、一定期間内で給油時刻推定を行った予測給油時刻と、実際の給油時刻から計算できるF1スコアを含める。したがって、まず初めに計測値から給油時刻を推定する手法を提案し、検証を行う。ここで、正常な計測値から正しく給油時刻を推定できることの確認を行い、その後給油時刻推定のF1スコアとその他の計算可能な特徴量を用いた異常検知手法の検証を行う。また、発生期間の長さが異なる異常をそれぞれ検知するために、異常検知手法の検証は扱う時系列の長さを変えて、二種類の枠組みの検証を行う。一つ目は、給油の消費が多い冬季を含む7ヵ月間の計測値を一つの時系列として扱い、正常か異常かの分類を行う。以下、一つ目の枠組みを異常変動検知(1シーズン)とする。二つ目は、計測値を給油時刻から次の給油時刻までで区切った期間を一つの時系列として扱い、正常か異常かの分類を行う。以下、二つ目の枠組みを異常変動検知(給油間期間)とする。

図1に考案したシステムの全体図を示す。異常変動検知(1シーズン)はセンサが故障しているかの判定に使用し、異常変動検知(給油間期間)は、より短い期間でセンサが不調であるかの判定に使用する。異常の判定にこの二つの枠組みを使用する理由は、発生期間が短い異常と長い異常で別々の対応をするためである。発生期間が長い異常はセンサが故障していると判断し交換をしなければならない。発生期間が短い異常は、レーザの射出口を清掃することにより再び正常な計測が可能になる場合があるため、射出口を清掃し様子を見る必要がある。

1.3 本稿の構成

以下に本論文の構成を示す。第2章では時系列データを扱った異常検知に関連した研究の解説を行う。第3章では、本論文で扱うレーザセンサデバイスの概要と、扱うデータの性質についての解説を行い、第4章では本論文で提案する給油時刻推定と異常変動検知(1シーズン)、異常変動検知(給油間期間)についての解説を行う。第5章の5.1節で提案手法の有効性を検証するための実験についての解説を行い、5.2節で実験

の結果及び考察を行う。最後に、第6章にて本論文のまとめを述べる。

2 関連研究

実世界の重要な問題として時系列異常検知は広く研究されており、それらの多くは教師ありと教師なしに分類される。教師あり時系列異常検知は時系列分類問題(TSC: Time Series Classification)の一つとして研究されており、TSCの中で分類クラスが異常と正常の二値分類である問題を異常検知としている。

TSCの先行研究としては、Maletzkeらの研究[3]が存在する。この研究では5つの医療時系列データから、平均や分散などの統計的な特徴量やエントロピー、フラクタル次元などの複雑さを表す特徴量を抽出し決定木モデルを用いて、てんかん患者の推定を行った。さらに、井上らの研究[4]では家庭の電力時系列から統計的な特徴量を抽出し、単純ベイズ分類器で個別家電機器の動作状態の推定を行った。これらの研究では、時系列から固定長の特徴量を計算し、分類器で教師あり学習をしている。また、茂木らの研究[5]では、振動センサの計測値から流量実験設備に組み付けられた調節弁の異常検知を行っている。この研究では計測値から得られた時系列をそのまま入力として、CNN-LSTMを用いて教師あり学習をしている。

教師なし異常検知は異常の判定基準により、密度推定法[6, 7]、クラスタリングベース[8, 9]、再構成ベース[10, 11]、自己回帰ベース[12]などに分類できる。また、近年様々な問題で力を発揮しているトランスフォーマーを用いた異常検知手法もJiehuiら[13]によって提案されており、時系列異常検知ベンチマークにおいて最先端の結果を出している。

教師なし異常検知が多く研究されている背景として、実世界の問題では異常データが正常データに比べて極端に少なく教師あり学習が困難であることが挙げられる。本論文で扱う問題では異常データが十分あるため、教師なし学習よりも有効であると考えられる教師あり学習を採用する。

3 レーザセンサデバイス

本論文で扱うレーザセンサデバイスは灯油タンクの蓋に設置され、タイムオブフライト方式で油面との距離を計測する。タイムオブフライト方式とは、光を対象に向けて発射し、照射時間と受光時間の時間差から距離を計測する方式である[14]。計測した距離はLPWAネットワークによりクラウドに保存されるため、インターネットが繋がる環境であればどこでも灯油タンク内の残量を把握することができるが、1.1節で述べた通

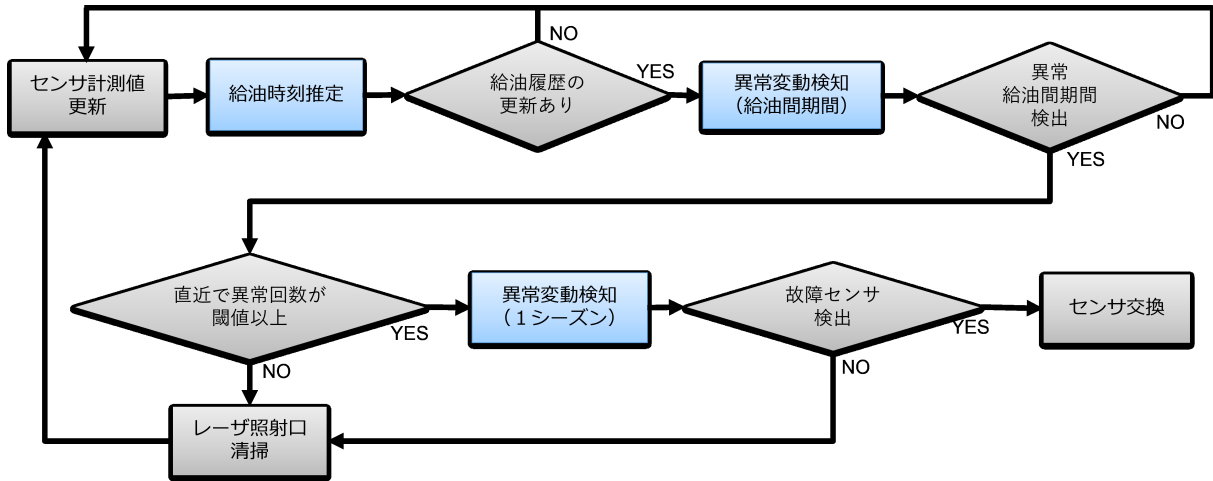


図 1: 異常センサ検出システム.

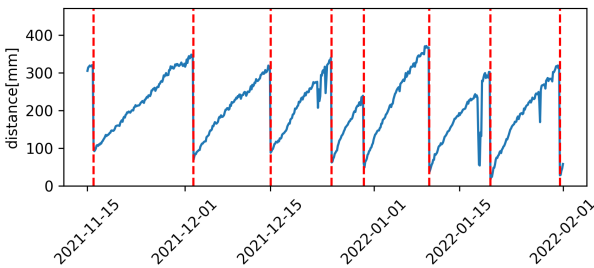


図 2: センサ計測値.

り、コストの面から高精度なセンサは使うことが出来ないため、ある程度の誤差は想定している。また、消費電力低減のために距離計測は3時間毎に行われるため、センサから得られるデータは3時間間隔の時系列データである。

本論文で扱うデータの例として、あるセンサにおける計測値の時系列および給油履歴を図2に示す。図に示す期間は2021年11月15日から2022年2月1日までとなっており、縦軸が油面との距離、横軸が時間である。赤い点線はタンクに給油した時刻を表している。図2より、センサ計測値は時間経過とともに徐々に油面との距離が大きくなり、給油が行われると油面との距離が急激に小さくなる。次に、センサにみられる異常の例を図3に示す。図3(a)はある一定期間に計測値が変動していない。図3(b)は連続でスパイクが立っている。図3(c)は計測値が減少傾向にある。いずれも図2とは異なる挙動を示している。

また、提案する手法には、センサ計測値の他にもタンクへ給油した時刻と量のデータが必要であるが、本論文ではそのような給油履歴が取得可能な状況を想定して実験を行う。

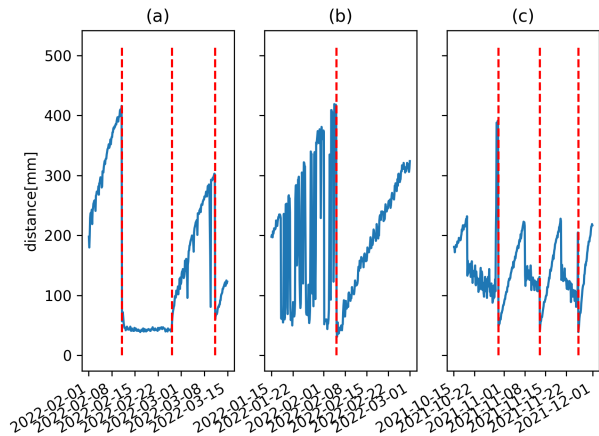


図 3: 異常センサ例.

4 提案手法

本章では、提案する給油時刻推定と異常変動検知手法について述べる。使用するデータはセンサ計測値と、各灯油タンクへの給油時刻と給油量が記録された給油履歴である。計測値 \mathbf{X} は時刻 t の計測値を $x_t \in \mathbb{R}$ とし、集合 $\{x_1, x_2, \dots, x_N\}$ で表す。また、給油時刻 \mathbf{T}_s は時刻の集合 $\{t_{s,1}, t_{s,2}, \dots, t_{s,K}\}$ で示され、 $t_{s,k} \in \mathbb{N}$ は k 回目の給油の給油時刻を表している。

4.1 給油時刻推定

センサ計測値から給油時刻を推定するために、給油時刻付近での計測値の挙動について考える。第3章で述べた通り、センサ計測値は時間経過とともに徐々に油面との距離が大きくなり、給油が行われると油面との距離が急激に小さくなる。つまり、給油時刻を推定するには前ステップとの差分が大きい箇所に注目すれば

良い。しかし、図2の2022年1月15日付近にみられるように、計測値にはスパイクが立つことがしばしばある。このようなスパイクも前ステップとの差分が大きくなるため、給油時刻を推定するためには、差分だけでなくその後の時系列も考慮しなければならない。

このことを踏まえて、以下の手順で給油時刻推定を行う。

1. 隣り合う2点間の差分の絶対値を計算し、差分が閾値以上である時刻を給油時刻候補として挙げる。
2. 過去の給油時刻から、給油時刻直後数ステップの計測値をサンプリングして平均と分散を求める。
3. ホテリング理論により、求めた平均と分散から給油時刻候補直後数ステップの計測値に対する異常度を計算し、異常度が閾値以下である給油時刻候補を給油時刻と推定する。

ここでは、ある灯油タンクの l 回目までの給油時刻が分かっている、それ以降の給油時刻を推定する問題を考える。まず、以下の式で前ステップとの計測値の差分を計算する。

$$\Delta x_t = x_t - x_{t-1} \quad (1)$$

ここで Δx_t は絶対値ではなく、相対値で計算する。これは、給油時刻では計測値が負の方向に大きく変化することが分かっているためである。 Δx_t が負の方向に大きい時刻の集合を給油時刻候補 $\mathbf{T}_{sc} = \{t_{sc,k'}\}_{k'=1}^{K'}$ とする。また、このときの閾値を $\gamma_{\Delta x_t}$ とする。次に過去の計測値と給油履歴から、給油直後 n ステップの計測値をサンプリングして、平均値 μ と分散 σ^2 を以下の式で求める。

$$\mu = \frac{1}{nl} \sum_{k=1}^l \sum_{t=t_{s,k}}^{t_{s,k}+n-1} x_t \quad (2)$$

$$\sigma^2 = \frac{1}{nl} \sum_{k=1}^l \sum_{t=t_{s,k}}^{t_{s,k}+n-1} (x_t - \mu)^2 \quad (3)$$

n は十分小さく正規分布に従う測定誤差を除き計測値は変動しないため、サンプリングした計測値は正規分布に従うと仮定できる。したがってホテリング理論より、給油直後の計測値の異常度 a が計算できる。ここで、給油時刻推定には \mathbf{T}_{sc} 直後の数ステップを考慮する必要があるため、異常度 a の計算には \mathbf{T}_{sc} 直後の n ステップの平均を用いる。 k' 番目の給油時刻候補の異常度 $a_{k'}$ は以下の式で計算できる。

$$a_{k'} = \frac{(\bar{x}_{t_{sc,k'}} - \mu)^2}{\sigma^2} \quad (4)$$

$$\bar{x}_{t_{sc,k'}} = \frac{1}{n} \sum_{t=t_{sc,k'}}^{t_{sc,k'}+n-1} x_t \quad (5)$$

(4)式で求めた異常度は正規分布からの距離を表しており、異常度が大きいほど過去の給油時刻直後の計測値の傾向から外れていることが分かる。したがって、異常度 $a_{k'}$ が小さい給油時刻候補 $t_{sc,k'}$ を給油時刻として推定し、このときの閾値を γ_a とする。また、(4)式で計算される異常度はカイ二乗分布に従うため、閾値 γ_a は確率を表すパラメータ α により一意に決まる。

$$\alpha = 1 - \int_0^{\gamma_a} \chi^2(a|1,1) da \quad (6)$$

4.2 異常変動検知 (1 シーズン)

提案手法では、図3のような計測値をとる異常センサを検知することを目的として、計測値と給油履歴から計算できる固定長の特徴量を抽出し、単純な分類器で教師あり学習を行う。教師ラベルは正常と異常の2種類で、7ヵ月間の時系列に対してアノテーションを行う。抽出する特徴量と計算式を以下に示す。

1回当たりの平均給油量： \bar{V}

$$\bar{V} = \frac{1}{K} \sum_{k=1}^K V_k \quad (7)$$

ここで、 K は7ヵ月間で行われた給油回数、 V_k は k 回目の給油の給油量を表している。

平均給油間隔： $\Delta \bar{t}_{s,k}$

$$\Delta \bar{t}_{s,k} = \frac{1}{K-1} \sum_{k=1}^{K-1} (t_{s,k+1} - t_{s,k}) \quad (8)$$

1ヶ月当たりの平均給油回数： \bar{K}

$$\bar{K} = \frac{240K}{N} \quad (9)$$

ここで、分母は時系列のステップ数、分子の240は1ヶ月当たりのステップ数を表している。異なる長さの時系列にも応用可能にするため、1ヶ月で正規化する。

給油時刻推定の F1 スコア

4.1節で提案した給油時刻推定が正常センサに対して性能が良く、異常センサに対して性能が悪いと考えられるため、給油時刻推定で得られたF1スコアを特徴量とする。まず、時系列に対して4.1節で提案した給油時刻推定を行う。次に、給油時刻推定を各ステップが給油時刻か給油時刻以外か

の二値分類問題として扱い、F1スコアによって評価を行う。二値分類問題において、給油時刻以外を給油時刻とラベル付けされた場合の数を偽陽性 (FP) と表す。偽陰性 (FN) は、給油時刻が給油時刻以外であるとラベル付けされた数を表す。真陽性 (TP) および真陰性 (TN) は、給油時刻と給油時刻以外が正しく識別された数を表す。これらを用いて、精度、再現率、F1スコアは以下のように評価される。

$$precision = \frac{TP}{TP + FP} \quad (10)$$

$$recall = \frac{TP}{TP + FN} \quad (11)$$

$$F1score = 2 \times \frac{precision \times recall}{precision + recall} \quad (12)$$

ノイズの強さ: d

計測値に対するノイズの強さを定量的に数値化するためにローパスフィルタを用いる。ローパスフィルタは、ノイズなどの特定の閾値よりも高い周波数信号を減衰させて遮断し、低域周波数のみを信号として通過させることができる。提案手法では、元の計測値 \mathbf{X} とローパスフィルタ適用後の計測値 \mathbf{X}_{lpf} との差分を時系列の長さで正規化した値をノイズの強さとして定義する。

$$d = \frac{1}{N} \sum_{t=1}^N |x_{lpf,t} - x_t| \quad (13)$$

閾値以上の Δx_t の数 (1ヶ月当たり): l_{so}

$$l_{so} = \frac{240}{N-1} \sum_{t=2}^N f_{so}(\Delta x_t) \quad (14)$$

$$f_{so}(\Delta x_t) = \begin{cases} 1 & \text{if } \Delta x_t > \gamma_{f_{so}}, \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

ここで f は Δx が閾値以上であるかの判定を行っていて、 γ_f は閾値である。この特徴量を計算することにより、計測値の正の方向への急激な変化を抽出することができる。

閾値以下の Δx_t の数 (1ヶ月当たり): l_{sw}

$$l_{sw} = \frac{240}{N-1} \sum_{t=2}^N f_{sw}(\Delta x_t) \quad (16)$$

$$f_{sw}(\Delta x) = \begin{cases} 1 & \text{if } \Delta x_t < \gamma_{f_{sw}}, \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

ここで g は Δx が閾値以下であるかの判定を行っていて、 γ_g は閾値である。この特徴量を計算することにより、計測値の負の方向への急激な変化を抽出することができる。

閾値以上の $\Delta x_{lpf,t}$ の数 (1ヶ月当たり): $l_{lpf,so}$

$$l_{lpf,so} = \frac{240}{N-1} \sum_{t=2}^N f_{so}(\Delta x_{lpf,t}) \quad (18)$$

$\hat{l}_{lpf,so}$ は \hat{l}_{so} と比べて、ローパスフィルタ適用後の時系列を用いて特徴抽出をしているため、より大域的な変動を抽出することができる。

Δt 間の計測値の傾きの平均: $\mu_{x'_t}$

計測値 \mathbf{X} における Δt 間の傾きを次式で定義する。

$$x'_t = \frac{x_{t+\Delta t} - x_t}{\Delta t} \quad (19)$$

正常なセンサの計測値の傾きは、センサ間である程度近いと考えられるため、計測値の傾きの平均を特徴量とする。また、給油回数によるばらつきを防ぐため、 Δt 間に給油時刻が含まれる場合は平均の計算に含めない。

$$\mu_{x'_t} = \frac{1}{N - K\Delta t} \sum_{t=1}^{N-\Delta t} g_{\mu}(x'_t) \quad (20)$$

$$g_{\mu}(x'_t) = \begin{cases} 0 & \text{if } \exists t_{s,k} \in \mathbb{N}, t \leq t_{s,k} \leq t + \Delta t, \\ x'_t & \text{otherwise} \end{cases} \quad (21)$$

Δt 間の計測値の傾きの分散: $\sigma_{x'_t}^2$

ノイズや計測値の減少により異常なセンサの計測値の傾きは分散が大きくなると考えられるため、 x'_t の分散を特徴量とする。 $\mu_{x'_t}$ と同様に、給油回数によるばらつきを防ぐため、 Δt 間に給油時刻が含まれる場合は分散の計算に含めない。

$$\sigma_{x'_t}^2 = \frac{1}{N - K\Delta t} \sum_{t=1}^{N-\Delta t} (g_{\sigma^2}(x'_t) - \mu_{x'_t})^2 \quad (22)$$

$$g_{\sigma^2}(x'_t) = \begin{cases} \mu_{x'_t} & \text{if } \exists t_{s,k} \in \mathbb{N}, t \leq t_{s,k} \leq t + \Delta t, \\ x'_t & \text{otherwise} \end{cases} \quad (23)$$

閾値以下の $|x'_t|$ の数 (1ヶ月当たり): $l_{x'_t}$

図3の(a)のセンサは、正常センサと比べて $|x'_t|$ が0に近い値をとる期間が存在する。このような特徴を表現するために、閾値以下である $|x'_t|$ の数をカウントして、1ヵ月で正規化した値を特徴量とする。

$$l_{x'_t} = \frac{240}{N - \Delta t} \sum_{t=1}^{N-\Delta t} h(x'_t) \quad (24)$$

$$h(x'_t) = \begin{cases} 1 & \text{if } |x'_t| \leq \gamma_{|x'_t|}, \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

ここで h は $|x'_t|$ が閾値以下であるかの判定を行っており、 $\gamma_{|x'_t|}$ は閾値である。

4.3 異常変動検知 (給油間期間)

時系列の長さに対して異常が発生している期間が短い場合、異常部位が正常部位に埋もれて特徴抽出が困難である可能性がある。したがって本節では、計測値を給油時刻から次回給油時刻までで区切った期間を一つの時系列として扱う異常検知手法を提案する。提案手法では異常変動検知 (1 シーズン) と同様に時系列から固定長の特徴量を抽出して、単純な分類器で教師あり学習を行う。教師ラベルは正常と異常の2種類で、給油時刻から次回給油時刻までの計測値に対してアノテーションを行う。抽出する特徴量を表1に示す。計算方法は4.2節と同じである。

表 1: 特徴量 (給油間期間)

記号	特徴量
N	時系列のステップ数
d	ノイズの強さ
l_{so}	閾値以上の Δx_t の数 (1ヶ月当たり)
l_{sw}	閾値以下の Δx_t の数 (1ヶ月当たり)
$\mu_{x'_t}$	Δt 間の計測値の傾きの平均
$\sigma_{x'_t}$	Δt 間の計測値の傾きの分散
$l'_{x'_t}$	閾値以下の $ x'_t $ の数 (1ヶ月当たり)

5 実験

本論文で提案した給油時刻推定と異常変動検知の有効性の検証を行う。扱うデータはセンサ計測値と、各灯油タンクへの給油時刻と給油量が記録された給油履歴である。

5.1 実験設定

5.1.1 給油時刻推定

正常データに対して給油時刻を正しく推定できることを確認するために、異常が少ないセンサを10個サンプリングして実験する。また推定における最適なパラメータ探索のため、 $\gamma_{\Delta x_t}$ と α を変化させて評価値を計算する。給油直後の計測値のサンプリング数 n は全センサで5と設定する。評価方法については、給油時刻を陽性、給油時刻以外を陰性として(12)式で計算できる F1 スコアを用いて比較を行う。

表 2: 給油時刻推定の F1 スコア

α	$\gamma_{\Delta x_t}$			
	50	100	150	200
10^{-1}	0.77	0.78	0.77	0.74
10^{-3}	0.91	0.92	0.91	0.84
10^{-5}	0.93	0.94	0.93	0.90
10^{-7}	0.91	0.91	0.92	0.90
10^{-9}	0.91	0.91	0.92	0.90

5.1.2 異常変動検知

各特徴量の計算に用いるパラメータは、検証用データに対して単一の特徴量に閾値を設けて分類を行い、 AUC_{PR} が最大となるものを使用する。検証用データは様々な異常が含まれるようにサンプリングしたセンサ10個分の時系列を使う。分類器の比較には、ロジスティック回帰 (LR)、ランダムフォレスト (RF)、サポートベクターマシン (SVM) を用いる。実験に使うデータ数は、異常変動検知 (1 シーズン) がセンサ200個分の時系列で、異常変動検知 (給油間期間) が給油間期間1000個分の時系列である。扱うデータの期間は2021年10月1日から2022年5月1日の7ヶ月とし、間訓練データとテストデータの割合は、異常変動検知 (1 シーズン) では5:5、異常変動検知 (給油間期間) では7:3とする。異常ラベルを陽性、正常ラベルを陰性として(10)~(12)式で計算できる精度、再現率、F1スコアと、PR曲線の下面積である AUC_{PR} で評価し、分類器の比較を行う。また、提案手法との比較のため、テストデータに対して単一の特徴量に閾値を設けて分類 (SFM: Single Feature Model) を行い評価する。SFMの実験は全ての特徴量に対して行い、最も AUC_{PR} が高かった特徴量の実験結果のみを掲載する。

5.2 実験結果

5.2.1 給油時刻推定

給油時刻推定の実験結果を表2に示す。表2より、検証したパラメータの大半が90%以上の精度を示した。また、 $\alpha = 10^{-5}$ 、 $\gamma_{\Delta x_t} = 100$ のとき最も良い精度を示し、F1スコアが0.94となった。したがって以降は、このパラメータを使って実験を行う。

5.2.2 異常変動検知 (1 シーズン)

異常変動検知 (1 シーズン) の実験結果を表3に示す。精度、再現率、F1スコアはF1スコアが一番高くなるような閾値設定での値を載せている。 AUC_{PR} はSVMが最も高く、精度、再現率、F1スコアはLRが

表 3: 異常変動検知 (1 シーズン) の実験結果

	AUC_{PR}	precision	recall	F1score
SFM	0.52	0.58	0.52	0.55
LR	0.84	0.80	0.85	0.82
RF	0.73	0.67	0.73	0.70
SVM	0.86	0.77	0.82	0.79

表 5: 異常変動検知 (給油間期間) の実験結果

	AUC_{PR}	precision	recall	F1score
SFM	0.55	0.58	0.56	0.57
LR	0.88	0.85	0.81	0.83
RF	0.82	0.77	0.84	0.80
SVM	0.88	0.84	0.81	0.82

表 4: 異常変動検知 (1 シーズン) 混同行列

true	predict	
	anomaly	normal
anomaly	27	6
normal	8	59

表 6: 異常変動検知 (給油間期間) 混同行列

true	predict	
	anomaly	normal
anomaly	47	11
normal	8	233

最も高いことが分かる。また、 AUC_{PR} が最も高かった SVM の F1 スコアが最も高い閾値設定での混同行列を表 4 に示す。表 4 より、異常センサの 82% を検知可能であり、異常センサであると予測したセンサのうち 77% が正しく分類できていることが読み取れる。これは SFM と比較して大幅に精度が向上しており、且つ、提案手法の目的である故障センサ検出システムとして、労働者の負担を十分減らせる結果であるため、提案手法の有効性を示すことができた。

5.2.3 異常変動検知 (給油間期間)

異常変動検知 (1 シーズン) の実験結果を表 5 に示す。精度、再現率、F1 スコアは F1 スコアが最も高くなるような閾値設定での値を載せている。全ての評価指標において LR と SVM が同水準で良い性能を発揮している。 AUC_{PR} が最も高かった LR の F1 スコアが最も高い閾値設定での混同行列を表 6 に示す。表 6 より、異常な給油間期間の 81% を検知可能であり、異常な給油間期間であると予測した時系列のうち 85% が正しく分類できていることが読み取れる。異常変動検知 (1 シーズン) と同様に、SFM と比較して大幅に精度が向上している。また、表 3 の結果と比較しても同水準の性能を発揮していることから、時系列を短くしても提案手法が有効であることを示すことができた。さらに、異常変動検知 (1 シーズン) との比較のため、異常変動検知 (1 シーズン) では偽陰性 (FN) となっていたセンサの分類結果を可視化したものを図 4 に示す。赤い点線がタンクに給油した時刻を表し、黄色で色付けされた期間が提案手法により異常であると推定された給油間期間である。異常変動検知 (1 シーズン) では時系列の長さに対して期間が短いため検知できなかった異常が、時系列を給油間期間で区切る異常変動検知 (給油間期間) では検知できることが読み取れるため、提案手法の有効性を示すことができた。一方で、図 5

に偽陰性の例を示す。図 5 では、1 月末から 2 月末までの給油間期間の一部で異常が見られるが、提案手法では異常を検知できていない。これは、異常変動検知 (1 シーズン) のときと同様に異常が発生している期間が扱う時系列と比べて極端に短いため、十分に特徴抽出できずに偽陰性になったと考えられる。

6 まとめ

本論文では、センサの計測値と給油時刻データから灯油タンクへの給油時刻を推定する手法と、給油時刻推定の F1 スコアとその他計算可能な特徴量を用いた分類器ベースの異常検知手法を提案した。実データに対して実験を行い、一般的な異常検知問題で扱われる評価指標で評価した結果、どの手法も現実の問題に有効であることを示すことができた。

本論文で提案した異常変動検知では、異常値が計測された具体的な時刻が分からないため、今後は時系列のどこの時刻で異常値が計測されたか分かるような手法を提案する予定である。

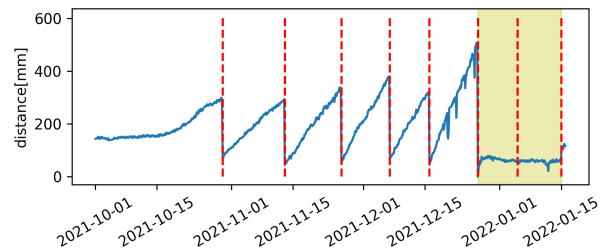


図 4: 分類結果の可視化。

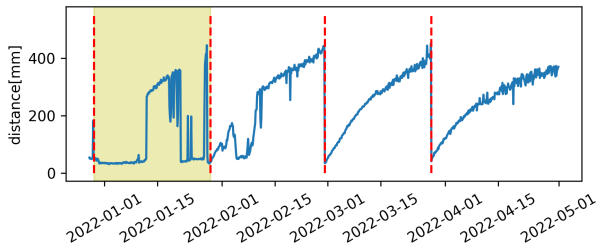


図 5: 偽陰性の例.

謝辞

本研究は、ノーステック財団「研究開発助成事業」の支援を受けて実施されたものです。ここに感謝の意を表します。

参考文献

- [1] 大江弘峻, 横山想一郎, 山下倫央, 川村秀憲, 多田満朗. タブーサーチを用いた灯油配送計画の最適化. 第 206 回知能システム研究発表会, 2022.
- [2] 劉兆邦, 横山想一郎, 山下倫央, 川村秀憲, 多田満朗. 灯油配送計画決定に向けた多層パーセプトロンを用いた家庭内灯油消費量の推定. 第 21 回複雑系マイクロシンポジウム, 2022.
- [3] André Gustavo Maletzke, Carlos Andres Ferrero, Chris Mayara Tibes, Everton Alvares Cherman, and Willian Zalewski. Medical time series classification using global and local feature extraction strategies. *Journal of Health Informatics*, Vol. 9, No. 3, p. 73–80, 2017.
- [4] 井上洋思, 石山文彦, 渡辺敏雄, 大山孝. O-028 家庭の電力時系列から得た統計的特徴に基づく家電機器の動作状態推定方法 (o 分野: 情報システム, 一般論文. 情報科学技術フォーラム講演論文集, Vol. 13, No. 4, pp. 379–380, 2014.
- [5] 茂木貴弘, 中澤友哉, 田原鉄也. 1d cnn-lstm による調節弁内部の異常検知. 人工知能学会全国大会論文集, Vol. JSAI2018, No. 3, pp. 3Pin144–3Pin144, 2018.
- [6] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, Vol. 29, No. 2, p. 93–104, 2000.
- [7] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.
- [8] Lifeng Shen, Zhuocong Li, and James Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. In *Advances in Neural Information Processing Systems*, 2020.
- [9] Youjin Shin, Sangyup Lee, Shahroz Tariq, Myeong Shin Lee, Okchul Jung, Daewon Chung, and Simon S. Woo. Itad: Integrative tensor-based anomaly detection system for reducing false positives of satellite systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020.
- [10] Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.
- [11] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, Vol. 54, pp. 30–44, 2019.
- [12] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.
- [13] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.
- [14] 今城勝治, 吉永秀雄, 亀山俊平. 位相差方式 tof センサと車両・車軸検知応用. 計測と制御, Vol. 59, No. 5, pp. 321–325, 2020.

顧客満足度アンケートデータに基づく 経営課題優先度可視化ツール CSIMG

CSIMG: a Tool for Visualizing Management Issue Priorities Based on Customer Satisfaction Survey Data

新福一貴^{1*} 笹嶋宗彦¹
Kazuki Shinfuku¹ Munehiko Sasajima¹

¹ 兵庫県立大学
¹ University of Hyogo

Abstract: Improving CS(Customer Satisfaction) is one of the most important issues for companies, but it is difficult to determine which evaluation axis improvements will lead to higher customer satisfaction. Although CS portfolio analysis and IPA have been studied as visualization methods to support the selection of issues on CS, visualization of the relationship with competing companies is still insufficient. In this paper, we propose CSIMG, a tool to visualize the position of customer satisfaction of a company within its industry, and show that it has the potential to output evaluation results that are similar to those of human experts.

1 はじめに

顧客満足度とは企業が提供するサービスや商品に対する顧客の満足度を表す指標であり、言うまでもなく、顧客満足度を向上させることは企業にとって最も重要な課題の一つである。しかし、最も投資対効果の高い顧客満足度向上方法を発見することは容易ではない。なぜなら、サービスや商品は、それ自体の品質や価格だけではなく、提供のされ方や提供される店の雰囲気、利用した時に得られる満足感など、多くの軸から評価されるものであり、どの評価軸における課題を解決することが顧客満足度向上につながるのかを見極めることは、一般に困難だからである。

評価軸には、「この企業に対する満足度」のような企業に対する総合的な満足度の軸と、総合的な満足度を判断する根拠となる、個別のサービスや商品に関する満足度の軸の2つがある。本研究では、前者を総合満足度、後者を個別項目と呼ぶ。例えば、カフェ業種に所属するある企業が店舗を展開するチェーン店Aの総合満足度は、「店Aについてのあなたの満足度は何点ですか?」のような、サービスや商品を提供する主体全体についての質問で評価することができる。他方、個別項目を評価するための質問としては、「店Aの商品について、味の満足度は何点ですか?」や、「店Aの接客態度に満足していますか?」などがある。

業種によって異なるが、一般に満足度を評価する軸は多岐にわたるため、総合満足度と個別項目を調査するアンケートを行ったとしても、どの個別項目を改善すれば良いのかを判断することは一般に困難である。カフェを例にすると、店の総合満足度に影響する要素としては、提供する商品の味や量や価格、メニュー種類の豊富さの他に、店員の接客態度やトラブル時の対応、店内の雰囲気や清潔感、客層、最近では、リモートワークに適した電源やネットワーク環境があるか、さらには店舗の立地が良いか、など、個別項目には非常に多くの軸が含まれる。アンケートの結果として得られるデータは、各個別項目の評価値だけであり、どの軸について優先的に取り組むべきかを判断する情報は、明示的になっていない。

総合満足度と個別項目の関係を可視化することで、総合満足度を改善する可能性の高い軸の選択を支援するための可視化手法については、従来から研究が行われている。総合満足度とそれに対する個別項目の重要度を相関関係より算出し、可視化するCS(Customer Satisfaction)ポートフォリオ分析や、IPA (Importance-Performance Analysis) と呼ばれる方法がある。Martilla&James(1977) [5]が開発したIPAと、松本ら(2004) [3]、岡本ら(2016) [1]、相良ら(2006) [2]がそれぞれ開発したCSポートフォリオ分析は、総合満足度とそれに対する個別項目の重要度を相関関係より算出し、可視化する点で本質的には同じである。

しかし、既存研究は、優先すべき課題を判断する際

*連絡先：兵庫県立大学社会情報学部社会情報科学科
〒651-2197 兵庫県神戸市西区学園西町8丁目2-1
E-mail:sasajima@sis.u-hyogo.ac.jp

に重要である、個別項目の評価軸について、複数の競合他社との関係性を可視化していない。既存研究は、各個別項目の重要度と満足度を可視化することで、解決すべき個別項目を明らかにする [5] が、他社との関係性が分からなければ、既に競合他社に対して良い評価を得ているような課題の改善に投資してしまったり、多くの企業が努力しても個別項目の評価値が互いに変わらないような、改善の余地が無い課題に投資してしまったりする誤った経営判断をしてしまうことになりかねない。Dolinsky(1991) [6] が提案する IPA は特定の企業に対する満足度だけでなく、その企業と同じ業種の他社との満足度の差を可視化することで、競合他社との関係性を導入している。これは既存の IPA では考慮できていない他社との比較ができる点で有用であると考えられるが、Dolinsky の方法は、1つの個別項目に対して1つの企業との比較しか可視化することができない。複数の企業との関連性が可視化できれば、同業の企業全体の中での自社の位置や、自社の差別化要因や、伸びしろの有無を判断することができる。

そこで本論文では、顧客満足度アンケートのデータを入力として、自社企業に対する顧客満足度が、同じ業種の中でどのような位置にあるかを可視化するツール CSIMG(Customer Satisfaction IMportance Graph) を提案する。CSIMG は、従来の研究が可視化してきた、自社サービスや商品の総合満足度と個別項目の関係だけでなく、各個別項目の評価値が、同業種の他社と合わせてどの程度ばらついているか、言い換えれば、改善の余地があるか否かを可視化することで、経営者が、どの個別項目に優先して取り組むべきかを判断することを支援する。また、CSIMG を評価するために、カフェ業、ドラッグストア業、来店型保険ショップ業、自動車保険業の4業種を対象として、株式会社 oricon ME によって行われたアンケートを入力として、CSIMG の出力と、人間の経営アドバイスの専門家である中小企業診断士の判断とを比較した。その結果、4つの業種のうち、カフェ業など、アンケート回答者である一般消費者が、サービスを自分で複数社利用し比較できるような業種については、CSIMG が既存の CS ポートフォリオ分析の手法と比較して、より人間の専門家に近い評価結果を出力することができた。

2 先行研究

この章では IPA と CS ポートフォリオ分析、及び、その拡張についての先行研究について説明する。

IPA は Martilla&James(1977) [5] が提案した手法で、サービスに関してどれくらい期待していたかを表す重要度 (Importance) と、どれくらい満足したかを表す満足度 (Performance) をリッカート尺度のアンケートで測定した値を入力に、2次元グラフを可視化し、課題

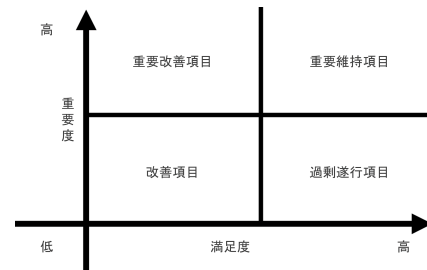


図 1: Martilla&James らの考案した IPA

の優先順を出力するものである。Martilla&James の論文では、重要度と満足度のどちらについても、サービスの利用者に4段階リッカート尺度で直接質問する形で調査している。調査の際、重要度はこの項目に対してどの程度重視していたか、という趣旨の質問であり、満足度は個別項目に対してどれくらい満足したか、という趣旨の質問で計測している。図1に示す。

CS ポートフォリオ分析は、1999年には日本で顧客満足度を用いた経営状況の可視化手法として香川県農業試験場研究報告にて大西らに農業技術の評価手法として利用されている [10]。CS ポートフォリオ分析は IPA と同じく、重要度と満足度を可視化し、経営上の課題点を発見するための手法であるが、一般に、IPA では縦軸を重要度、横軸を満足度としてプロットするのに対して、CS ポートフォリオ分析は縦軸を満足度、横軸を重要度としてプロットするという違いがある。両者の違いを図3に示す。

IPA [5] は発表してから現在まで様々な議論があり、多様化している。これら多様化した IPA は主に4つの視点から分類することができる。

1つ目は、重要度についての評価値を、直接測定と間接測定のどちらで得ているかという視点である。直接測定とは重要度について、回答者に直接質問して調査する方法であり、間接測定とは、総合的な満足度に対する影響度を相関や回帰分析を用いて統計的に算出する方法である。

2つ目は、重要度の間接測定について、何の手法を用いているかという視点である。主な手法としては、相関係数を用いるアプローチと、線形回帰分析の係数を用いるアプローチがあるが、[7] は、重要度と満足度に関係があることを根拠に、相関分析が適切であるとしている。

3つ目は、可視化する際の軸とデータをプロットする際の中心点をそれぞれどこに取り、課題と判断される領域をどのように定義するかという視点である。

Martilla&James(1977) [5] は、アンケートに用いたリッカート尺度の中央の値を中心とする方法を提案した。しかしこの方法は、各質問項目に対する実際の回

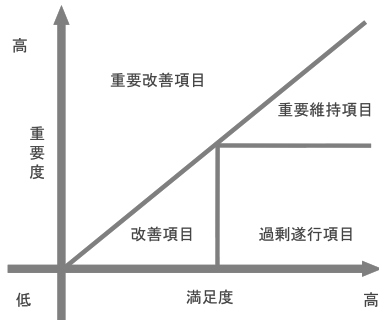


図 2: rial らの提案した IPA

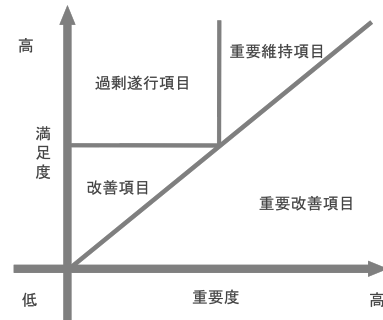


図 3: rial らの IPA を反映した CS ポートフォリオ分析

答の平均値がリッカート尺度の中央値と異なり、かつ、互いに近い値をとっていた場合、全ての項目が、軸の中心を外れた特定の象限に集中してしまう。この場合、どの項目が課題なのかが分かりづらくなるため、可視化方法として適切ではないという問題点があった。これに対し、Rial ら [9] は、回答から得られた重要度と満足度のそれぞれの平均値をもとに中心点を決定するという方法を提案した。図 2 に示す。

さらに、Martilla&James(1977) [5] は、設定した軸が分割する 4 つの象限を用いて、課題として認識すべき領域を定義したが、Rial ら [9] の方法を用いた場合、中心がずれるため、絶対値で見ると改善すべき個別項目が、改善の必要がないと判断される象限にプロットされてしまうことが起き得る。これを解決するため、Rial ら [9] は満足度と重要度が等しい直線(以後、等評価線と記載)を引き、等評価線からの距離で課題点の優先順位を算出する方法を提案した。

4 つ目は、満足度について、競合企業との比較を支援する方法についてである。Deng ら (2008) [7] は、満足度について、対象組織の満足度を分子とし、競合企業の最大の満足度を分母とする満足度比を算出し、その値を満足度として IPA に組み込んでいる。これにより、それぞれの個別項目について、その項目の評価値について最大の値をもつ企業との消費者目線での評価の違いを可視化することができている。一方で、この方法は項目の最大値をもつ企業との比較のみで可視化しているため、同種の企業が 3 つ以上ある場合、その全体の中での自社の立ち位置を表現できていない。

以上述べた通り、IPA については、4 つの視点から分類することができるが、「はじめに」で述べた通り、顧客満足度の可視化においては、4 つ目に挙げた論点である競合要素を取り入れることが必要である。本節で取り上げた過去研究は、競合要素について十分に取り入れられていない。そこで、本研究では、競合他社との位置関係を可視化することによって、従来の IPA が行ってきた課題の優先度を明らかにすることに加えて、同業種の中における自社の位置を可視化する点が新規

表 1: 元となるアンケートデータ (* n, m は整数値)

顧客 ID	最近利用した企業名	総合満足度	個別項目 1	...	個別項目 n
0	企業 A	10	8	...	7
1	企業 A	9	10	...	6
...
m	企業 C	7	9	...	10

であると考えている。

3 CSIMG の提案

本稿で提案する CSIMG(Customer Satisfaction Importance Graph) は、CS ポートフォリオを拡張した手法であり、自社を含めた同業種の複数企業を対象に取ったアンケート結果を入力すると、他企業との競合要素を考慮した自社の経営課題を優先順位をつけて出力する。

3.1 CSIMG のアルゴリズムと入力データ

CSIMG の入力は、B2C の形態で一般に行われる事業における、ある 1 つの業種に所属する複数の店舗について、それぞれの利用顧客からとった、満足度に関するアンケートデータである。各アンケートデータは、対象の店が提供するサービス全体と、そのサービスを構成する個別項目について、リッカート尺度で計測された満足度アンケートデータである。元となるアンケートデータの形式を例として表 1 に示す。表 1 のデータを可視化して、課題優先順位を出力するために、アンケートデータから、以下の 3 つの情報を抽出する。

■情報 1. 個別項目の平均値: 各個別項目の平均値を算出する。分析対象企業における個別項目の大小は、企業の課題発見のための基本的な情報であると考え。

■情報 2. 総合項目と個別項目との相関関係: 分析企業と同じ業種の企業群における個別項目と総合項目との

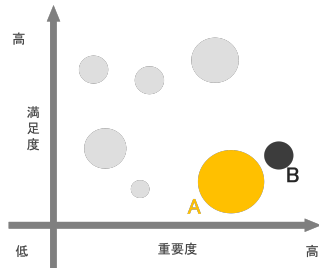


図 4: CSIMG の概念図

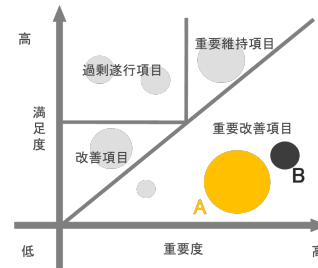


図 5: CSIMG の各象限の意味を表した図

偏相関係数を算出する。

■情報 3. 個別項目の企業間でのばらつき: 個別項目の平均値を企業ごとに算出したものの標準偏差を企業間のばらつきを表す指標として算出する。

これら3つの情報を抽出したうえでアンケートデータを表2のデータ形式に加工し、さらに、これら3つの情報を図4のように可視化する。図中の各点が個別項目に対応しており、x軸、y軸、点の大きさはそれぞれ次を表している。

■ x 軸: 分析対象の企業の業種における、各項目に対応する個別項目と総合満足度との偏相関係数の偏差値

■ y 軸: 分析対象企業の個別項目の平均値の偏差値

■ 点の大きさ: 各項目に対応する個別項目の、分析対象の企業と同じ業種の企業における標準偏差を最大値を1, 最小値は0に近い値になるようラプラススムージングをして正規化した値

$$NormalizedStd_x = \frac{Std_x - \min(Std_x) + 0.01}{\max(Std_x) - \min(Std_x) + 0.01} \quad (1)$$

* Std_x : 個別項目 x の分析対象の企業と同じ業種の企業における標準偏差, $\min(Std_x)$: Std_x の最小値, $\max(Std_x)$: Std_x の最大値, $NormalizedStd_x$: 0.01でラプラススムージング(加算スムージング)した正規化後の Std_x

以上のように可視化することで、個別項目について、総合満足度に与える影響が大きい、重要なものほど、x軸方向の値が大きくなるまた、消費者から、相対的に不満に思われている個別項目ほど、y軸方向の値が小さくなる。さらに、点の大きさが大きいほど、その個別項目の満足度評価が企業によって大きく異なっていることを表す。

この3種類の情報を可視化することにより、企業として、どの項目を改善すれば良いかを発見することが容易となる。例えば、図4の個別項目AとBは、両者ともX軸方向の値が、他の個別項目と比較して大きいため、総合満足度との偏相関係数が比較的高い項目であり、総合満足度を向上させる上で重要であることが分かる。また、Y軸方向の値は、他の個別項目と比較して小さく、これらの個別項目に対して、アンケート

回答者が不満を持っており、この点からも改善すべき項目であることも分かる。

さらに、AとBを比べると、Aの方が点の大きい。AとBは、X、Y軸上でプロットされた位置から、いずれも改善すべき個別項目であることが分かるが、Aの項目に対する個別項目評価値の企業間のばらつきの方が大きいことから、Aの項目の方が、改善の余地があり、改善への投資対効果が高いと判断できる。項目Aのように、個別項目について、満足度のばらつきが大きいということは、同じ業種の中で、その項目について、顧客を大きく満足させている企業と、不満を与えている企業があることを意味する。自社が、その項目について不満を与えているならば、大きく満足させるように改善する余地があることを意味する。逆に、同業他社が複数あるにも関わらず、項目Bのように、満足度のばらつきが小さいということは、項目B改善のために投資や努力をしたとしても、改善の余地がAに比べて小さい、すなわち、改善のための投資に対する効果が小さくなってしまいう可能性が高いことを意味する。従来のIPAとCSポートフォリオ分析では、点の大きさにあたる部分がなく、企業への評価のばらつきを可視化していない。

次に、CSIMGを利用した課題の抽出と、順序付けの方法について図5に示す。満足度と重要度が等しい値を結んだ線を等評価線とする。等評価線を境に、右下を重要改善項目、左上を過剰遂行項目、右上を重要維持項目、左下を改善項目と名付ける。このうち、重要改善項目にプロットされた個別項目のみを課題項目であるとして抽出する。

課題の順序付けでは、はじめに、抽出後データの3種類の情報のそれぞれを偏差値化した値をもとに、次の条件を満たす基準点を算出する。

■満足度が最も小さい(最大値0)

■重要度が最も大きい(最小値100)

■ばらつきが最も大きい(最大値100)

次に、基準点からユークリッド距離に近いものから順に課題解決の優先度が高いと順序付けをする。以上は、Rialら[9]の優先順出力の手順を、企業間のばら

表 2: 加工した後のデータ (部分)

企業名	個別項目 1~n の平均値	総合満足度と個別項目 1~n の偏相関係数	企業ごとの個別項目の 平均値の標準偏差
企業 A	8.2, ... , 8.5	0.5, ..., 0.4	0.68
企業 B	7.4, ... , 9.5	0.3, ..., 0.5	0.68
...
企業 h	5.4 ... , 8.3	0.6, ..., 0.2	0.68

*個別項目の平均値や偏相関係数は企業ごとに、企業ごとの個別項目の平均値の標準偏差は業種ごとに値を算出

表 3: 株式会社 oricon データセットの概要

ジャンル	業種	年度	回答数
カフェ	カフェ	2020	8,427
保険	自動車保険	2016	7,150
保険	自動車保険	2017	5,471
...
保険	来店型 保険ショップ	2018	2,159
ドラッグ ストア	ドラッグストア	2020	8,290
14 ジャンル	98 業種	計 210	約 750,000 件

つきの視点から拡張したものである。これにより、従来の方法に、他企業に対する満足度のばらつきを加えた課題優先順を出力できる。

以上が、満足度アンケートデータを入力とし、課題の優先度と各課題の他社との関係性を可視化して、個別項目の解決すべき優先順を出力する CSIMG のアルゴリズムである。

3.2 受理可能な入力データについて

本評価実験において、入力データセットは、株式会社 oricon の顧客満足度データセット (以降、オリコンデータセットと表記) を利用する。株式会社 oricon は、2016 年から 2020 年までの 5 年間に収集した、14 ジャンル 98 業種を対象に行った 210 セットのアンケート調査の個票を収集しており、調査項目は、利用経験、満足度、サービスを利用する際の重視点、他者への推奨意向、再利用意向などが含まれ、アンケートの個別項目の構成は、対象となる業種ごとに異なる。回答者数は延べ約 75 万名で、性別や年齢など回答者の属性も付随している。表 3 に概要を示す。

この中から対象の企業についてどれくらい満足しているかを表す総合満足度と、総合満足度を構成する個別のサービスについての満足度を調査した個別項目を用いる。満足度は全て 10 段階リッカート尺度で測定されており、同じ業種に所属する複数社を対象に集められている。個別項目については、例えばカフェ業で

は、「店内・店外の雰囲気、居心地の良さ (照明, BGM, インテリアなども含め)」や「商品の見た目のよさ」、「店内の広さ (座席間のゆとり, 移動のしやすさなども含め)」などの個別項目を 10 段階 (10 点: 非常に満足 ~ 1 点: 非常に不満) で集めたデータとなっている。

後述する CSIMG の評価には、oricon データセットのデータ 210 セットのうち、個別項目が重要改善項目の象限に多くプロットされた、2020 年カフェ業、2020 年ドラッグストア業、2018 年来店型保険ショップ業、2018 年自動車保険業の 4 業種のデータセットを用いる。

4 CSIMG の評価

CSIMG を評価するにあたって、CSIMG と既存方法を、4 つの業種についてアンケートデータを入力として課題優先順を出力し、比較して評価する。既存方法は、可視化については相良ら [2] の CS ポートフォリオ分析、課題優先順位の抽出アルゴリズムは Rial らの方法をベースラインとする。可視化方法について、CS ポートフォリオ分析を選択した理由は、IPA を可視化すると、縦軸と横軸に対応するものが CSIMG と逆であるため、比較しにくいと考えたからである。

また、同じ作業を人間の専門家である中小企業診断士に行わせて、CSIMG とベースラインによる課題優先順位付けと、中小企業診断士の優先順位付け結果とを比較し、どちらの手法の方が専門家の判断と近いかを評価する。ベースラインは、表 2 の個別項目の平均値と、総合満足度と個別項目の偏相関係数を用いて可視化し、課題優先順抽出については、Rial ら [9] の提案した満足度と重要度が一致する線を示す等評価線からの距離で課題優先順を出力する方法を用いる。Rial らは重要度について、直接測定を用いているが、今回はアンケートデータに該当する質問が無いため、間接測定の値を用いる。等評価線からの距離を図示したものを図 6 に示す。

課題優先順抽出の提案手法とベースラインを比較した有用性評価について、最適な方法は、まったく同じ条件及び環境の企業を二つ用意し、ベースラインで算出した課題順に取り組む企業と、提案手法で算出した

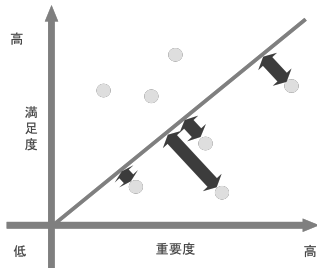


図 6: CS ポートフォリオ分析の課題優先順位抽出

課題順に取り組む企業のその後の業績を評価することである。しかし、このような対照実験は現実世界では難しく、同じ条件や環境の企業を用意できたととしても、その後の推移には他の様々な要因が絡むため、手法による影響の評価は困難と考えられる。

そこで、マーケティングや分析に関する知識をより深く持ち、中小企業の経営課題に対応するための診断・助言を行う専門家である中小企業診断士が選ぶ課題順を正解データであると考えて、提案手法とベースラインの方法を評価することにした。

中小企業診断士から正解データを取得する際には、今回の満足度データのうち、要約統計量として表 2 で示したデータの他、各個別項目及び総合満足度の中央値を見せた。それらデータと本人の持つ専門的な知識を合わせて、どの個別項目がその業種における平均的な課題であり、また、優先順位はどうかを回答させた。

中小企業診断士から収集した正解データをもとに、まず最初に、中小企業診断士の回答とベースライン及び CSIMG が課題抽出した項目を、カフェ業、ドラッグストア業、自動車保険業、来店型保険ショップ業の 4 業種について比較する。次に、CSIMG やベースラインで出力した項目に関する課題順位とその項目に対する専門家の正解の課題順位との差を取得し、高順位における誤差により大きなペナルティを与えるために、差を順位で割った値の合計値をランク考慮残差和と定義して、この値が小さいほうがより専門家の判断に近いとした。また、業種どうしの人間の専門家の正解との誤差を比較するため、ランク考慮残差和を、それぞれの業種ごとに異なる課題抽出数で割ることで、項目当たりのランク考慮残差和を算出する。

4.1 評価実験

はじめに、カフェ業、ドラッグストア業、自動車保険業、来店型保険ショップ業の 4 業種について、oricon アンケートデータを入力とし、ベースラインによって可視化した結果を図 7 に示す。

CS ポートフォリオ分析による可視化については、軸と等評価線を図示した。等評価線よりも右下にプロッ

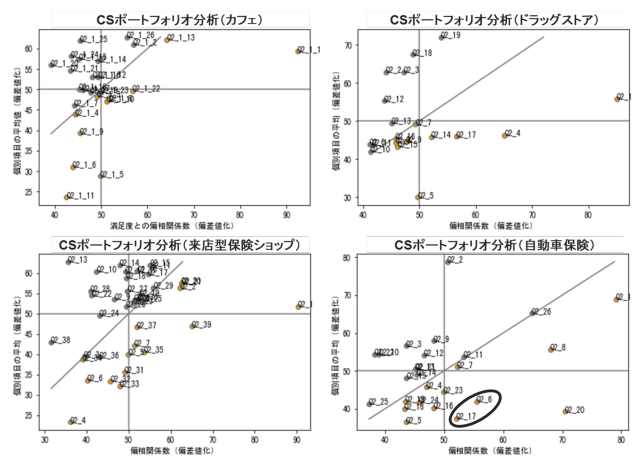


図 7: 4 業種のアンケートデータの CS ポートフォリオ分析による可視化

トされた項目が、企業が優先して取り組むべき項目となる。

次に、同じ 4 業種に対する CSIMG の手法による可視化結果を図 8 に示す。

図 7 と図 8 を比較すると、個別項目に円の大きさ、すなわち、その項目への満足度の、同業種間でのばらつきという情報が加わることで、企業が優先して取り組むべき課題がより明らかになっている。例えば、自動車保険の Q2.6(補償内容に対する保険料) と、Q2.17(事故・故障担当者の交渉力) は、CS ポートフォリオ分析で見ると、等評価線からの距離がほぼ同じである(図 7 中に表記)。しかし、CSIMG で見ると、Q2.17(事故・故障担当者の交渉力)の方が点の大きさが大きく、すなわち、企業ごとのばらつきが大きく、Q2.6(補償内容に対する保険料)の方は、小さいことがわかる。つまり、CSIMG を利用すれば、Q2.6(補償内容に対する保険料)よりも Q2.17(事故・故障担当者の交渉力)の項目の方が、企業間でばらつきが大きく、特に自社のサービスに対する満足度が低かった場合には、改善による満足度の向上がより期待できるため、優先して取り組むべきであることが、一目で分かる(図 8 中に表記)。

CSIMG も課題として認識する項目は CS ポートフォリオ分析と同じく、等評価線の右下の点であるとした。

また、専門家からは、各業種の課題を抽出して優先順位をつけた理由と、それぞれのアンケートデータについてのコメントを得た。

カフェ業種については、一般に多くの人が、無理なく複数のカフェ店を利用できるので、異なる店舗を比較したうえでアンケートに回答できると考えられる。そのため、アンケートの回答数値を信頼することができ、統計量と業種のビジネスモデルから、あまり苦労せずに課題を抽出できたとのことであった。次に、ドラッ

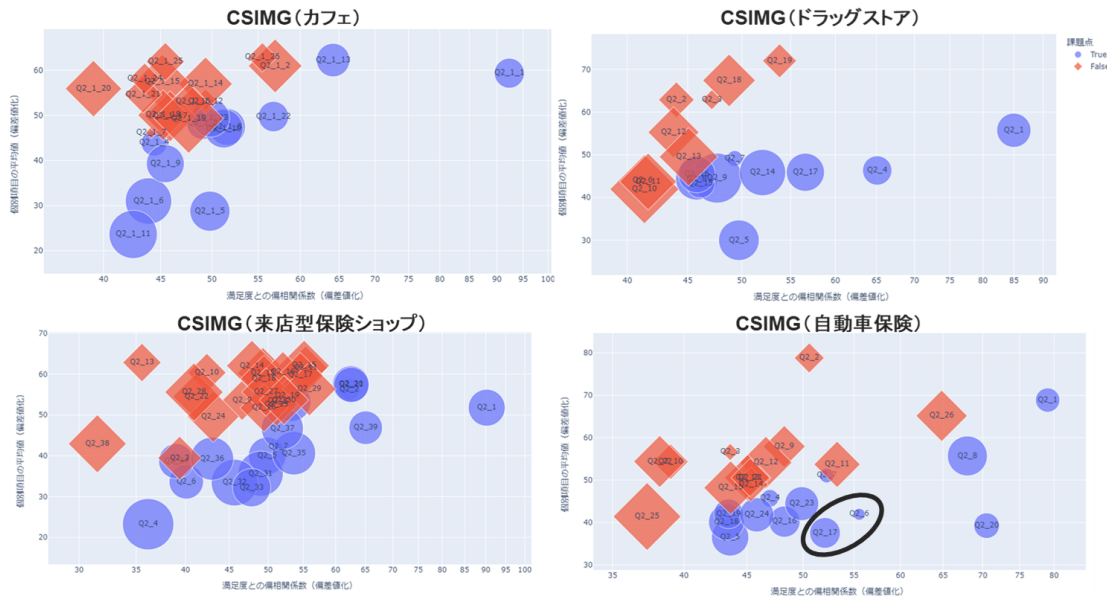


図 8: 4 業種アンケートデータの CSIMG による可視化

表 4: CS ポートフォリオ分析及び CSIMG の課題優先順位結果と専門家による課題優先順位結果との比較 (項目当たりの順位考慮残差和による誤差評価)

業種名称	CSIMG	CS ポートフォリオ分析 (ベースライン)
カフェ	<u>1.382</u>	1.703
ドラッグストア	<u>1.488</u>	1.605
保険ショップ	2.711	<u>2.598</u>
自動車保険	2.367	<u>2.085</u>

グストアについては、地域によって、ドラッグストアの機能が全く異なり、例えば都市部の店は狭い店舗で経営しているので、限られた種類の品ぞろえしかできないが、地方では店舗が広く、スーパーマーケットの機能を備えたものも出現しており、アンケートの「品揃えの豊富さ」の意味も異なる。同じ満足度の点数が表すものが回答者によって異なるので、点数の信頼性がカフェに劣り、課題抽出と順位付けが難しかったとのことであった。

自動車保険と保険ショップについては、事故に遭ったことがある人とならない人で、回答の数値の意味が大きく異なると考えられる。また、そもそも、複数の保険を比較することが出来にくい商品(個人が、複数社と同じ種類の保険契約を結ぶことはあまり無いと考えられる)であるため、ドラッグストアの場合よりも回答数値の信頼度が低く、課題抽出は更に困難であったとの

ことであった。

次にベースラインと CSIMG の課題優先順位抽出結果をもとに、4 章で述べた項目当たりの順位考慮残差和による誤差評価を、専門家の正解データと比較した結果を表 4 に示す。ベースラインと CSIMG を比較した結果、カフェとドラッグストアについては CSIMG の誤差が小さく、来店型保険ショップと自動車保険に関してはベースラインの誤差が小さかった。また、業種間比較した際、カフェとドラッグストアの誤差が小さく、来店型保険ショップと自動車保険の誤差が大きいことがわかり、専門家がコメントで指摘した難易度順に誤差も大きくなっていることがわかる。ここから、専門家がコメントで指摘したもののうち、比較的信頼できると述べた業種で、CSIMG の優先順位抽出が誤差が小さく、専門家の判断に近いことがわかった。

5 議論とまとめ

本稿では、満足度アンケートデータを入力にした既存手法の CS ポートフォリオ分析に企業ごとのばらつきを加えた手法を提案し、その可視化と課題優先順の出力について比較した。その結果、CSIMG は、競合関係を考慮した可視化をすることで、従来の手法よりも複数の企業との立ち位置を考慮した解釈をすることができ、また、より有用な可視化ができる可能性を示すことができた。課題優先順位抽出については、専門家がある程度自信をもって正解データを作成できた業種に関しては精度が良く、そうでない業種については精度が低かった。また、特に専門家がアンケートデータを信頼できると判断したカフェ業種対象のアンケート

データについて、CSIMG はベースラインよりも誤差が小さく、精度が上回ったことから、CSIMG の競合関係を考慮した課題優先順抽出が有用であることが言えると考えている。特にカフェ業種のように顧客が自然と複数社のサービスを利用し比較しやすい業種では、CSIMG が有効に機能する可能性が高い。

一方、CSIMG にはいくつかの課題や限界がある、1つ目は、課題優先順抽出の精度については、来店型保険ショップと自動車保険に関して専門家のコメントにあったような、業種とデータセットが特殊である場合に、精度が悪くなる。具体的には、利用者が、異なる複数の企業によるサービスを利用する機会がほとんど無く、正しく比較してアンケート回答が出来ないような業種については、CSIMG は精度が悪くなる。

2つ目は、個別項目の満足度について、企業間でばらつきが小さいことの理由が少なくとも2つあり、CSIMG では、それらを区別できないことである。

理由の1つ目は、そもそも改善に限界があり、その企業の業種全体として個別項目の満足度が低い、つまりばらつきが小さい個別項目が存在する。例えば、カフェ業種のアンケートで登場する「店内の温度・空調・換気(分煙対策も含め)」という個別項目は、どの企業も一定の設備投資を行えば同じ程度の満足感を提供できるものであり、それ以上の満足感をユーザーに与えることは困難であるため、企業によってアンケートの満足度に差がでにくい。同様に、保険業種のアンケートにある「補償内容に対する保険料」という項目は、ユーザー目線ではより多くの金銭を得たいというニーズがあるため改善すべき項目となりがちだが、利用者から徴収する保険料には限界があるため、各社間で満足度のばらつきは小さくなる。

このように、例え重要であったとしても、そのサービスに関して打てる施策に物理的な限界があったり、利益追求における制約があったりすると、企業間のばらつきは小さくなる。

理由の2つ目は、業種によっては、カフェ業種における「店内の清潔さ」など、満足度が下がることが総合満足度に深刻な悪影響を及ぼす項目が存在するためである。これらは、各企業とも、一定の満足度を維持することは必須であり、結果として、アンケートの満足度のばらつきも小さくなっている。こうした、満足度を維持することが必須の項目(以後、最低限維持項目と記載)は、投資に対する改善効果が期待できないからと言って、投資対象ではないとは言えない。

改善に限界のある項目や、最低限維持項目を、区別できない点については、CSIMG の限界である。

本稿では、以上の、企業のばらつきが小さい個別項目について、それ以上向上させることが難しい項目であるとして一律に扱った。CSIMG を実際に企業が利用する際に、これらを区別する必要がある場合には、ど

の項目が最低限維持項目であるか利用者側で設定するようにするなど、利用側で工夫をすれば、提案手法は十分な情報を利用者に提供できると考えている。

謝辞

本研究では、国立情報学研究所の IDR データセット提供サービスにより株式会社 oricon ME から提供を受けた「オリコンデータセット」を利用した。ここに記して謝意を表す。

参考文献

- [1] 岡本哲弥, 林美玉: 重要度-満足度マトリクスにおける境界の基準化と有意性の適用, 滋賀大学経済学会, 408号, p4-19, 2016
- [2] 相良 英憲, 北村 佳久, 古野 勝志ほか: Customer Satisfaction(CS) 分析を応用した実務実習モデル・コアカリキュラム実施における改善項目の抽出, 医療薬学, 32巻, 4号, p. 295-305, 2006.
- [3] 松本 幸正, 塚本 弥八郎: CS 分析の考え方を導入した授業評価アンケートの分析と授業改善ポイントの定量化, 京都大学高等教育研究, 10巻, 10号, p. 21-32, 2004.
- [4] 株式会社 oricon ME (2019): 顧客満足度調査データ. 国立情報学研究所情報学研究データリポジトリ. (データセット). <https://doi.org/10.32130/idr.10.1>
- [5] John A. Martilla and John C. James, "Importance-Performance Analysis", *Journal of Marketing* Vol. 41, No. 1 (Jan., 1977), pp. 77-79 (3 pages)
- [6] Dolinsky, L. Arthur, "Considering the Competition in Strategy Development: An Extension of Importance-Performance Analysis", *Journal of Health Care Marketing*; Boone 巻 11, 号 1, (Mar 1991): 31.
- [7] Deng, W.-J., Kuo, Y.-F., & Chen, W.-C. (2008). Revised importance-performance analysis: three-factor theory and benchmarking. *The Service Industries Journal*, 28(1), 37-51.
- [8] Abalo, J., Varela, J. and Rial, A. (2006) El Análisis de Importancia-Valoración aplicado a la gestión de servicios, *Psicothema*, 18, 730-737
- [9] Rial, A., Rial, J., Varela, J., & Real, E. (2008). An application of importance-performance analysis (IPA) to the management of sport centres. *Managing Leisure*, 13(3-4), 179-188.
- [10] 大西 智司, 農家満足度調査を活用した農業技術評価 CS ポートフォリオ分析の適用, 香川県農業試験場研究報告, 51号, p.61-66(1999-03)
- [11] 館知也, 後藤千寿, 齋藤康介, 大野佑城, 安田昌宏, 水井貴詞, 小林健司, 佐橋誠, 野口義紘, 寺町ひとみ, 「多変量解析および CS(Customer Satisfaction) 分析による薬剤師の実務実習生に対する講義の評価」, 日本ファーマシューティカルコミュニケーション学会 Vol.12 No.2 2014
- [12] Senso Ormanović, Alen Ćirić, Munir Talović, Haris Alić, Eldin Jelešković, Denis Čaušević, "IMPORTANCE-PERFORMANCE ANALYSIS: DIFFERENT APPROACHES", *Acta Kinesiológica* 11 (2017) Supp. 2: 58-66
- [13] 皆月昭則, 水野凌太郎, 「患者満足度調査による医療機関向けのサービスサイエンス概念の導出」, 情報科学技術フォーラム講演論文集 14 (3), 549-550, 2015 2015-08-24 FIT(電子情報通信学会・情報処理学会) 運営委員会
- [14] 井上 勝雄, 杉山 裕香, CS 分析を用いた女性用通勤靴デザインの経験価値の分析, 日本デザイン学会研究発表大会概要集, 2017, 64巻, p. 204-, 2017