一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IMQ2024-87,IE2024-165,MVE2024-104(2025-03)

談話解析を用いた因果関係に基づく評価構造の自動構築

宮嶋 大輔† 杉本 匡史†† 橋本 翔‡ 長田 典子†

†関西学院大学理工学部/感性価値創造インスティテュート 〒669-1330 兵庫県三田市学園上ヶ原 1 番
††阪南大学国際学部 〒580-0032 大阪府松原市天美東 5 丁目 4−33
‡ 西南学院大学商学部 〒814-8511 福岡市早良区西新 6-2-92
E-mail: †{daisukemiyajima, nagata}@kwansei.ac.jp ††sugimoto@hannnan.ac.jp ‡s-hashimoto@seinan-gu.ac.jp

あらまし　大規模言語モデル(LLM)と談話関係解析を活用し，製品レビュー等のテキストデータから評価グリッド法に用いられる評価構造を自動的に可視化する手法を提案する．提案手法では，まず EventGraph を用いてテキストデータから因果関係を抽出し，評価項目間の因果ペアを生成する．次に，類似する評価項目をカテゴリ化し，GPT-4 を用いたプロンプトエンジニアリングによって各カテゴリのキーフレーズを生成する．最後に，これらの情報を評価構造可視化システムに入力し評価構造を生成する．実験の結果，提案手法は，カテゴリ生成とキーフレーズ生成の両方において高い精度を達成できることが示された．

キーワード　大規模言語モデル，評価グリッド法，生成モデル，因果分析

Automated Construction of Evaluation Structures Based on Causal Relationships Using Discourse Analysis

Daisuke MIYAJIMA† Masashi SUGIMOTO†† Sho HASHIMOTO†‡ Noriko NAGATA†

† School of Science and Technology, Kwansei Gakuin University
1 Uegahara, Gakuen, Sanda-shi, Hyogo, 669-1330 Japan
†† Faculty of International Studies, Hannan University,
Amamihigashi, Matsubarashi, Osaka, 580-0032 Japan
‡ Faculty of Commerce, Seinan Gakuin University
Nishiara, Fukuoka-shi, Fukuoka, 814-8511 Japan

Abstract　We propose a novel method for automatically visualizing the evaluation structure, as used in the evaluation grid method, from textual data such as product reviews. Our approach utilizes large language models (LLMs) and discourse relation analysis. The method first extracts causal relationships from the text data using an EventGraph, creating causal pairs between evaluation criteria. Subsequently, similar evaluation criteria are clustered into categories, and representative key phrases for each category are generated via prompt engineering with GPT-4. Finally, this information is fed into an evaluation structure visualization system to produce the final evaluation structure. Experimental results demonstrate that the proposed method achieves high accuracy in both category generation and keyphrase generation.

Keywords　Large-scale language model, Evaluation grid method, Generative model, Causal analysis

1. はじめに

産業分野において，ユーザ中心設計の重要性が増している．市場には多様な製品・サービスが溢れ，ユーザニーズは複雑化の一途を辿る中，企業が競争優位性を確立するためには，個々のユーザの価値観や嗜好，さらには潜在的な欲求までを的確に把握し，製品・サービス開発に反映することが不可欠である．

このようなユーザニーズの把握に有効な手法として，評価グリッド法[1]がある．評価グリッド法では，ユーザの評価を，抽象的な価値判断，機能的な価値から，具体的な物理的特性まで階層的に整理し，評価構造（図）として可視化する．また，半構造化インタビューの一種であり，ラダリングと呼ばれる上位概念・下位概念を聞き出す対話手順を用いることで，インタビューから効率よく評価構造を構築できる．この手法は，製品開発やサービスデザインなど，多くの研究開発に用いられ，その有用性が示されてきた．しかし，評価グリッド法は，実施における人的・時間的負荷が高いという問題を抱えている．

一方，自然言語処理分野では，文章から因果関係を把握する技術が進展しており[2, 3]，因果分析タスクにおける有用性が示されつつある．ここで，評価グリッド法で得られる評価構造図は，ユーザの評価における因果関係のネットワークとして捉えることが可能である．つまり，評価項目間の関係性は，ある評価が別の評価の原因となる，あるいは結果として導かれるといった因果的な繋がりとして解釈できる．

この点に着目し，本研究では，談話関係や LLM を活用し，評価グリッド法で得られる評価構造の可視化の

自動化・効率化を図る．具体的には，インタビュー記録や製品レビュー等のテキストデータから，これらの技術を用いて因果関係を自動抽出し，評価構造を構築する手法を開発する．

2. 関連研究

本研究に関連する先行研究として，因果分析技術，および評価グリッド法の効率化に関する取り組みについて説明する．

2.1. 因果分析

因果分析手法は，大規模言語モデル(LLM)を用いた因果分析[2]と談話関係解析に基づいたルールベースの因果分析[3]の 2 つに大別される．後者において，特定のイベント間の因果関係や時間的関係をグラフ構造で表現する EventGraph[3]に関する研究が注目を集めている．EventGraph は，テキストデータからイベント間の因果関係を抽出し，それらを構造的に表現するグラフ構造である．この技術は，複雑な事象の理解や推論への応用が期待されている．従って，評価グリッド法で得られる階層性を持つ評価構造の表現や分析にも有用であると考えられる．

2.2. 評価グリッド法の効率化

評価グリッド法は大きく分けて，①インタビュー対話による評価項目およびそれらの因果関係抽出，②類似した評価項目のカテゴリ生成，③カテゴリ毎のキーフレーズ生成，④評価構造図の生成，の４つのステップからなる．本手法の人的・時間的負荷を軽減し，その適用範囲を広げるため，自動化に向けた取り組みが試みられてきた．

代表的な研究である E-grid[4]は，④（評価構造図の生成）をグラフ理論に基づき行うことで，評価項目の重要度に応じて可視化の粒度を調整するシステムであり，その有用性が示された．

一方で著者らは，①（インタビュー対話）の代わりに，①’顧客レビューなどのテキストデータから評価構造を自動構築する試みを進め，これまでアプレイザル評価表現辞書に基づく方法[5]や，単語の係り受け関係に基づく方法[6]を提案した．さらに，近年の大規模言語モデル（LLM）を活用することにより，本来の①であるインタビュー対話自体，および②，③の自動化を果たし，①から④までの全体を評価グリッド法インタビューシステム（Evaluation grid method interview system: EGi[8]）として開発した．本システムでは，①をマルチエージェント LLM，②を BERTopic と感情分析[7]，③を LLM[8]を用いて実装している．なお④はグラフ理論ベースの可視化機能[4]に加え，さまざまなインタフェース機能を有している（EGi-Visualizer サブシステム: EGi-V）．

これらの研究は，評価グリッド法の自動化に向けて大きく前進しているものの，以下のような課題が残されている．まず，①’テキストデータからの評価構造自動構築においては，因果関係の抽出精度や評価項目の抽象化レベルの統一性に改善の余地がある．また，②カテゴリ生成と③キーフレーズ生成の精度向上が求められられている．

3. 提案手法

本研究では，LLM と談話関係解析を組み合わせ，テキストデータから評価構造を自動かつ高精度に構築する手法を提案する（図 2）．本手法は，①’因果関係抽出，②カテゴリ生成，③キーフレーズ生成，④評価構造生成の 4 つのステップで構成される．①’因果関係抽出では，EventGraph を用いて談話関係を解析する．②カテゴリ生成では，LLM による高精度な意味ベクトル化と Affinity Propagation[9]によるデータ駆動型クラスタリングを組み合わせ，サンプル数の事前指定を不要とし高精度なクラスタリングを達成する．③ キーフレーズ生成では，ドメイン固有知識を活用することにより，評価構造の解釈性を向上させる．

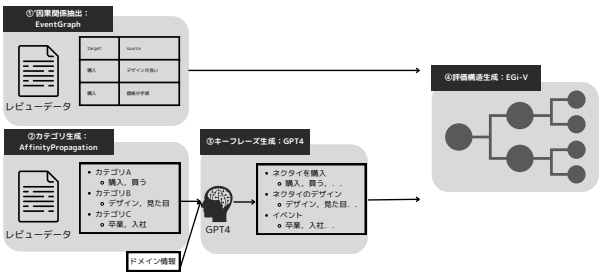

図２ 提案手法の概念図

3.1. 因果関係抽出

まず，テキストに対して EventGraph を適用し，文や節の間の談話関係を解析する．本研究では，EventGraph の出力のうち，談話標識が「原因・理由」と判定されたものを評価項目の因果ペア（評価項目（原因）と評価項目（結果））として抽出する．例えば，「価格が安いため購入した」というレビューから，「価格が安い」と「購入した」という評価項目の因果ペアを抽出する．

3.2. カテゴリ生成

カテゴリ生成では，因果関係抽出で得られた評価項目（原因・結果）を，LLM として，日本語の短文に特化した SentenceTransformer[10] モデルである sonoisa/sentence-bert-base-ja-mean-tokens-v2 を使用して数値ベクトルに変換する．これは，単語間の意味的な近さを計算し，類似した評価項目を特定するためである．単語を数値ベクトルとして表現することで，ベ

クトル間の距離や角度を計算することができ，意味的に近い単語は近いベクトルとして表現される．次に，ベクトル化された評価項目を Affinity Propagation を用いてクラスタリングする．Affinity Propagation は，データ間の類似性に基づいて自動的にクラスター数を決定する手法であり，事前にグループ数を指定する必要がない．さらに，高品質なクラスター結果を生成しつつ，計算効率も優れているため，大量の評価項目を効果的に処理できる．ここでは事前にカテゴリ数を予測することが困難なため，本手法が有用である．

3.3. キーフレーズ生成

各カテゴリに適切な名前を付ける．ここでは，OPENAI 社が開発した LLM である GPT-4[11]を用いて，各グループに含まれる評価項目を解釈し，人間が理解しやすいカテゴリ名を作成する．具体的には，解釈するクラスタがどのドメインのものなのかを予め指定したプロンプトを GPT-4 に与えることで, GPT-4 の事前知識を活かしたキーフレーズ生成を行う．与えたプロンプトを図 3 に示す．

```
<命令>
以下のクラスタを解釈してキーフレーズをつけてください.
なおキーフレーズは〇〇が△△の形にして（例：服が可愛い、乗り心地が
良い、使い勝手が悪い）
        <フレーズ群>。
{all_result}
        </フレーズ群>
</命令>
<制約>
・（服が可愛い）などキーフレーズのみ出力してください
・　クラスターの要素をそのまま抜き出すのではなく、このクラスターが
何を意味しているのかを解釈した上でキーフレーズをつけてください
・　地名・人名など固有名詞を表すクラスターはそのまま抜き出して下さ
い
・改行コードなどを入れないでください
・これは{domain}に関するクラスターです。それを考慮してください。
・キーフレーズは１つのみにしてください
・キーフレーズは見て意味がわかるようにしてください
</制約>
"""
```

図 3　キーフレーズ生成のプロンプト

3.4. 評価構造可視化

前処理，カテゴリ生成，キーフレーズ生成の結果を，本学が開発した評価構造可視化システム EGi-V に入力することで評価構造を自動生成する．

4. 性能評価実験

本章では，提案手法の有効性を検証するために，カテゴリ生成，キーフレーズ生成・評価構造可視化の 3 点について評価実験を行い，その精度を検証する．

4.1. カテゴリ生成

4.1.1. 実験データ

本実験では，評価グリッド法を用いて実際に収集された 3 種類のデータセットを使用した．各データセットは, 異なるドメイン（Performance, Tourism, Fashion）における評価項目で構成されており，手作業でカテゴリ分類とカテゴリ名が付与がされている．データセットの詳細は表 1 の通りである．

表 1　データセットの詳細

評価項目データセット	評価項目数	平均単語数	手作業でのカテゴリ数
Performance	681	9.8	41
Tourism	772	2.9	110
Fashion	471	1.1	115

4.1.2. 実験方法

提案手法を含め，以下の 4 つのクラスタリング手法の精度を比較した．

(1) G-Means[12]
エンコーディング: Tf-idf
K-Means 法をベースとし，統計的検定に基づき最適なクラスター数を自動決定するアルゴリズム．

(2) スペクトラルクラスタリング[13]
エンコーディング: Tf-idf
データの類似度行列の固有ベクトルを用いてデータを低次元空間に写像しクラスタリングを行う手法．

(3) BERTopic[14]
エンコーディング: Tf-idf
Sentence-BERT を用いて文書をベクトル化し，UMAP で次元削減後，HDBSCAN でクラスタリングを行い，c-TF-IDF でトピックを抽出する手法．

(4) 提案手法
エンコーディング: Sentence-Transformers
SentenceTransformers を用いて文書埋め込みを行い，Affinity Propagation でクラスタリングを行う手法．

各手法によって生成されたクラスタを，手作業で付与された正解カテゴリと比較し，以下の 5 つの評価基準でクラスタの質を評価した．これらの評価は，研究者 3 名による合議で決定した．

(i) 概ね一致: 生成されたクラスタが正解データのカテゴリと概ね一致している．

(ii) 複数(2～3)のカテゴリが統合: 正解データの複数のカテゴリが 1 つのクラスタに統合されている．

(iii) 分割されたカテゴリ: 正解データの 1 つのカテゴリが複数のクラスタに分割されている．

(iv) 新規カテゴリ: 正解データには存在しないが，新しいカテゴリとして妥当なクラスタが生成されている．

(v) 解釈不可: クラスタ内の評価項目に一貫性がなく，意味のあるカテゴリとして解釈できない

実験の結果を表 2 に示す.

表 2 カテゴリ生成結果

(a) Performance

		提案手法	G-means	スペクトラル	BERTopic
比較結果	(i) 概ね一致	**51.4**	33.3	55.4	51.6
	(ii) 複数 (2 3) のカテゴリが統合	**29.9**	45.6	21.5	25.0
	(iii) 分割されたカテゴリ	**16.8**	0.0	6.2	9.4
	(iv) 新規カテゴリ	**0**	1.8	4.6	7.8
	(v) 解釈不可	**0**	19.3	12.3	6.3

(b) Tourism

		提案手法	G-means	スペクトラル	BERTopic
比較結果	(i) 概ね一致	**81.4**	12.9	53.3	61.6
	(ii) 複数 (2 3) のカテゴリが統合	**10.6**	50.0	22.2	22.2
	(iii) 分割されたカテゴリ	**7.5**	1.4	6.7	5.1
	(iv) 新規カテゴリ	**0.6**	0.0	7.8	4.0
	(v) 解釈不可	**0**	35.7	10.0	7.1

(c) Fashion

		提案手法	G-means	スペクトラル	BERTopic
比較結果	(i) 概ね一致	**67.6**	9.7	58.1	59.7
	(ii) 複数 (2 3) のカテゴリが統合	**22.5**	25.8	25.8	26.4
	(iii) 分割されたカテゴリ	**8.8**	0.0	0.0	0.0
	(iv) 新規カテゴリ	**0**	0.0	3.2	1.4
	(v) 解釈不可	**1.9**	64.5	12.9	12.5

4.2. キーフレーズ生成

4.2.1. 実験データ

本実験では，4.1.の実験で用いた 3 つのデータセット（Performance，Tourism，Fashion）を対象とした．各データセットは既にクラスタリングされ，手作業でキーフレーズが付与された状態のものを使用した．

4.2.2. 実験方法

被験者 3 名に対し，評価項目データごとに手作業と提案手法のキーフレーズを比較してもらい，その類似度を 5 段階で評価してもらった．全カテゴリ数のうち 類似度の合計が 15 点満点中 12 点以上のカテゴリの割合をキーフレーズ生成の正解率として算出する．

4.2.3. 実験結果

各データセットにおける各手法のキーフレーズ生成の正解率を表 3 に示す．表中の太字は，各データセットにおいて最も高い精度を示した手法を表す．

表 3 キーフレーズ生成結果

	提案手法	TextRank	YAKE	EGi	Tf-idf	EmbedRank
Performance	**77.3**	53.5	17.1	61.0	56.1	75.6
Tourism	**90.4**	46.4	30.9	63.6	40.9	83.6
Fashion	**85.5**	70.4	47.8	80.0	62.6	75.7

4.3. 評価構造生成

4.3.1. 実験データ

本実験では，楽天市場におけるビジネス用メンズネクタイの商品レビューデータを用いた．レビューデータは，2020 年 1 月 31 日までに投稿されたもので，商品数は 147 点，レビュー総数は 515 件であった．

4.3.2. 実験方法

提案手法を用いて，レビューデータから評価構造を自動構築した．生成された評価構造は，EGi-V を用いて可視化した．可視化の際には，Katz 中心性が 0.0676 以上のノードのみを表示することで，重要な評価項目とその因果関係に焦点を当てた分析を可能にした．生成された評価構造に対しては，視覚的な妥当性の確認と，因果関係の正しさに関する定量評価を行った．定量評価では，抽出された評価項目ペア（因果ペア）を以下の 3 つの基準で分類した．

(1) 因果関係あり: 2 つの評価項目の間に明確な因果関係が認められる．

(2) 同値: 2 つの評価項目がほぼ同じ意味を表している．

(3) 因果関係なし: 2 つの評価項目の間に因果関係が認められない．

4.3.3. 実験結果

提案手法によって生成された評価構造の例を図 3 に示す．図 3 は，重要な評価項目とその因果関係に焦点を当てた結果を表している．

定量評価の結果，抽出された因果ペアのうち 66.7%が「因果関係あり」，7.5%が「同値」と判定され，合計で 74.2%が意味的につながりのある関係を捉えられていることが示された．

5. 考察

5.1. カテゴリ生成

実験の結果，提案手法は多くの指標において他のクラスタリング手法を上回る性能を示した．特に，Toursism と Fashion において，提案手法で生成されたクラスタは，手作業で分類された正解データと高い一致率を示し，解釈不可能なクラスタの割合も非常に低かった．この結果は，提案手法で用いた SentenceTransformer による文書埋め込みと Affinity Propagation によるクラスタリングが，評価項目の意味的類似性を捉え，適切なクラスタを生成する上で有効

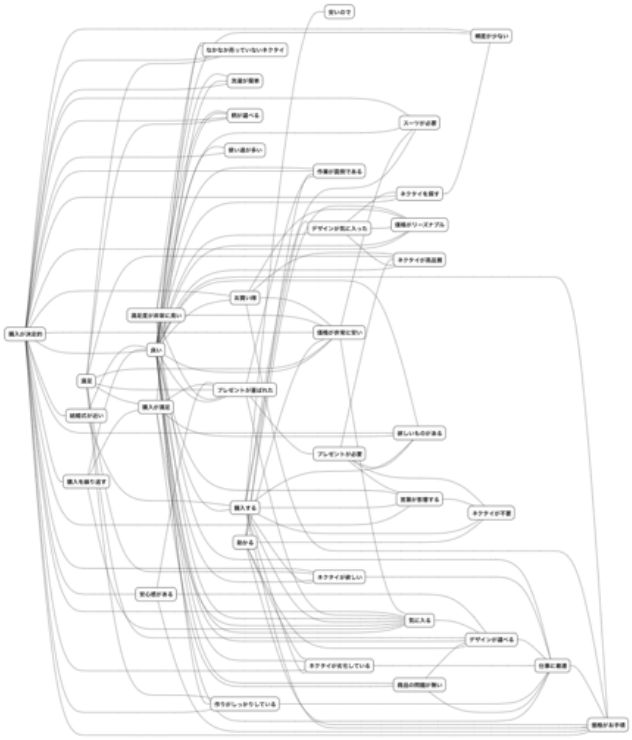

図 3　生成された評価構造図(Katz 中心性=0.0676)

であることを示している．

表 4 に BERTopic と提案手法の比較結果例を示す．例えば，BERTopic の「主張が弱い」カテゴリには，「主張が激しい」という極性が逆の要素が含まれているのに対し，提案手法ではこれらが分かれていることがわかる．また，「クール・さわやか」についても，BERTopic ではカテゴリが分かれていなかったのに対し，提案手法では手作業による分類と同様に分かれていることがわかった．さらに，「フォーマル・カジュアル」についても，提案手法ではカテゴリが正しく分かれていることが確認された．

一方， Performance においては，提案手法はスペクトラルクラスタリングに次ぐ精度となった．Performance の評価項目は，他のデータセットに比べて表現の多様性が高く，高次元の空間に分布していると考えられるため，スペクトラルクラスタリングの方が高い精度を示した可能性がある．しかし，提案手法で生成されたクラスタは手作業で分類されたカテゴリを細分化する傾向が見られ，これは，提案手法がより詳細なレベルでの評価構造の分析を可能にする可能性を示唆している．また，手作業においては，クラスタを分割するよりも統合する方が負荷が低いため，この特性は実用上の利点となり得る．

表 4　BERTopic と提案手法の比較例

カテゴリ	手作業	BERTopic	提案手法
主張が弱い	主張が強くない 主張弱くない 主張が激しくない あまり主張しない 主張が弱い	主張が強くない 主張弱くない 主張が激しくない あまり主張しない 主張が弱い 主張が激しい	主張が強くない 主張が弱くない 主張が激しくない あまり主張しない 主張が弱い
クール	クール クール系 さわやか	フレッシュ クール系 さわやか	クール クール系 さわやか
さわやか	さわやか さわやかに見える	さわやか さわやかに見える	さわやか フレッシュ さわやかに見える さわやかで かわいらしい
フォーマル	フォーマル かっちりしている	服装に気を付けている フォーマル	かっちりしている きっちりしている
カジュアル	カジュアル カジュアルな感じ かちっとしてない	カジュアルな感じ	カジュアル カジュアルな感じ カジュアル カジュアルな 感じ

5.2. キーフレーズ生成

実験の結果，全ての結果で他手法を上回る精度になったことがわかった．理由としては，大規模言語モデルのファインチューニングによるものだと考えられる．従来は，文章のベクトル表現を元に，キーフレーズ生成を行なっていたが，GPT の事前知識を使うことで，ドメイン知識を把握した上で，キーフレーズ生成を行うことで，手作業と同等のキーフレーズ生成を行うことができたと考察できる．

5.3. 評価構造生成

視覚的評価と定量的評価の結果から，提案手法は評価構造を自動的に構築する上で有効な手法であることが示された．特に，EGi-V による可視化の結果，上位概念には「購入が決定的」「満足」「良い」といった感情的な評価項目が，下位概念には「頻度が少ない」「バリエーションが豊富」「価格がお手頃」といった具体的な評価理由が抽出されていることが確認できた．これは，評価グリッド法におけるラダリングで得られるような，評価とその背後にある要因との関係性を捉えられていることを示唆している．また，「普段使いに適している」から「使い道が多い」，「価格が非常に安い」から「お買い得」といった，意味的に妥当な因果関係も抽出されていることが確認できた．

定量評価の結果，抽出された因果ペアの 66.7%に明確な因果関係が認められ，7.5%が同義の表現であった．これらの結果は，提案手法が評価項目間の意味的なつながりを高精度に捉えられていることを示している．

6. まとめ

本研究では，LLM と談話関係解析を組み合わせ，テキストデータから評価グリッド法の評価構造を自動構築する新しい手法を提案した．製品レビュー等から談話関係解析により評価項目の因果ペアを抽出した．また LLM を用いたカテゴリ生成により，データ駆動型で高精度なクラスタリングを実現した．さらに GPT-4 の事前知識とドメイン情報を活用し，人間が理解しやすいキーフレーズを生成した．実験の結果，提案手法は高い精度で評価構造を自動構築できることが示され，特に，カテゴリ生成とキーフレーズ生成の精度向上が確認できた．提案手法は，評価グリッド法の適用コストを大幅に削減し，大規模なユーザ理解を可能にする．

今後は，多様なデータでの検証と，抽出された評価構造の活用方法の検討を進める．

謝　辞

本研究は JSPS 科研費 JP22H03681 の助成を受けた．

文　献

[1] 讃井純一郎， 乾正雄， “レパートリー グリッド発展手法による住環境評価構造の抽出－認知心理学に基づく住環境評価に関する研究 (1)－，” 日本建築学会計画系論文， No. 367， pp. 15-22， 1986.

[2] Jing.M, “Causal Inference with Large Language Model: A Survey, ” arXiv preprint arXiv:2409.09822, 2024.

[3] 清丸寛一，植田暢大，児玉貴志，田中佑，岸本裕大, 田中リベカ, 黒橋禎夫，“因果関係グラフ: 構造的言語処理に基づくイベントの原因・結果・解決策の集約,” 言語処理学会 第 26 回年次大会，1125−1128，2020

[4] Onoue, Y., Kukimoto, N., Sakamoto, N., & Koyamada, K.: E-Grid: a visual analytics system for evaluation structures. Journal of Visualization, 19(4), pp. 753-768 (2016).

[5] G.A.Kelly, “The Psychology of Personal Constructs, ” W. W. Norton & Company， New York， 1955.

[6] 山田篤拓, 橋本翔, 長田典子,“レビューデータを用いた評価表現辞書に基づく印象の自動指標化,” 日本感性工学会論文誌, 17(5), 567-576.

[7] 大谷俊太，橋本翔，杉本匡史，長田典子,“単語の係り受け関係に基づく印象評価構造の自動構築,” 第 17 回日本感性工学会春季大会, 2D2-04, 2

[8] 宮嶋大輔，張帆，杉本匡史，佐々木香暖，北野泰成，橋本翔，長田典子,“大規模言語モデルを用いた評価グリッド法に基づくインタビュー対話システム,” 信学技報，vol.123，no.180，MVE2023-20，pp.33-38，2023.

[9] B.J.Frey, D.Dueck, “Clustering by Passing Messages Between Data Points, ” Science, vol.315, no.5814, pp.972-976, 2007.

[10] N.Reimers, I.Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, ” 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.3982-3992, 2019.

[11] OpenAI， “GPT-4 Technical Report, ” arXiv preprint arXiv:2303．08774， 2023.

[12] G.Hamerly, C.Elkan, “Learning the k in k-means, ” Advances in Neural Information Processing Systems, vol.16, pp.281-288, 2003.

[13] A.Y, Ng, M.Jordan, Y.Weiss, “On Spectral Clustering: Analysis and an algorithm, ” Advances in Neural Information Processing Systems, vol.14, pp.849-856, 2001.

[14] M.Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure, ” arXiv preprint arXiv:2203.05794, 2022.