TETDMを用いたテキストマイニング初心者の ツールの選択と操作を助ける操作履歴の提示†

砂山 渡*¹・中江 剛士*²・西原 陽子*³・畑中 裕司*¹

近年、多くの分野でデータを業務に活用する動きが活発になっている。市販の多くのデータ分析ソフトウェアを利用する際には、データ分析を行う人は事前にソフトウェアの利用法を含めたマイニングのための知識を得なくてはならない。通常、分析ソフトウェアは説明書やチュートリアルを備えるが、データ分析の初心者を実問題に対して適切なツールを選択した上で、複数のツールを使って分析を進めることは必ずしも容易ではない。そこで本研究では、テキストマイニングのための統合環境 TETDM を用いて、テキストマイニングの初心者に対して、分析の助けとなるツールや操作を含む他人の操作履歴を提示する機能を提案する。そして、この提示機能が初心者のテキストマイニングのツールの利用方法に与えた効果を検証する実験を行い、被験者が分析結果の解釈の幅を広げられる新しいツールの利用を促す効果が期待できることを確認した。

キーワード：テキストマイニング，データ分析支援，データサイエンティスト育成，操作履歴，TETDM

1. はじめに

近年、多くの分野でデータを業務に活用する動きが活発になっている。データ分析は我々の意思決定の根拠を見出すために行われるが、分析ツールの出力をそのまま知識として活用できるわけではなく、ツールの出力結果に人間が意味を与える解釈を行うことが必要となる。

インターネット上にある大量の Web ページや、業務レポート、アンケート結果やレビュー記事などには、多くのテキストが含まれており、これらをうまく分析データとして活用することが望まれている。分析の目的に対して、適切なツールを選択した上で、複数のツールをうまく使いこなしていく必要が生じるため、テキストマイニングの初心者が実際に分析にとりかかることは必ずしも容易ではない。

さまざまなツールには、説明書やチュートリアルが用意されていることが多く、本研究で用いるテキストマイニングのための統合環境 TETDM [1] においても、チュートリアルが用意されている [2]。しかし多くのチュートリアルは、ツールの開発者側によって用意されたシナリオに沿って操作するいわゆる受動的な体験に留まっている。そのため、ユーザが実際に分析を試みようとする能動的なツール活用場面における、適切なツールの選択やツールの操作を行う経験を補うには十分では

ない。

そこで本研究では、テキストマイニングのための統合環境 TETDM を用いて、テキストマイニングの初心者に対して、分析の助けとなるツールや操作を含む他人の操作履歴を提示する機能を提案する。また、その提示機能が初心者のテキストマイニングのツールの利用方法に与える効果を検証する。長期的に効果的な知的活動を営むために、他人の行動を模倣することは重要であり、効果的に行動を学ぶことができる環境をデザインすることは有用である。本研究ではテキストマイニングのソフトウェアの一つとして TETDM を用いるが、本研究で提案する機能や知見は、類似するテキストマイニングソフトウェアにも応用可能と考えている。

以下本論文では、2 章で関連研究について述べる。3 章で提案する機能の詳細について述べ、4 章で実装した機能を評価するために行った実験の結果と考察を述べる。最後に、5 章で本論文を締めくくる。

2. 関連研究

本研究ではテキストマイニングの初心者に対しツールの選択と操作を伝達することを支援する。本章では、まずツールの選択および操作と手続き的知識と宣言的知識との関係について触れた上で、ツールの選択と操作の伝達支援に関する研究について述べる。最後に、本研究のベースとして用いるテキストマイニングのための統合環境 TETDM を用いた操作履歴の再現機能について述べる。

2.1 手続き的知識と宣言的知識

本節では、ツールの選択および操作と手続き的知識と宣言的知識との関係について述べる。

ツールの選択と操作は人に伝達可能な知識の一種と考えられる。本研究では認知心理学における宣言的知識と手続き的知識という用語を用いて研究の位置付けを明らかにする。

† Presentation System of Operating Histories to Support Choice and Operation of Tools for Beginners of Text Mining Software TETDM
Wataru SUNAYAMA, Tsuyoshi NAKAE, Yoko NISHIHARA,
and Yuji HATANAKA

*1 滋賀県立大学工学部

School of Engineering, The University of Shiga Prefecture

*2 滋賀県立大学大学院工学研究科

Graduate School of Engineering, The University of Shiga Prefecture

*3 立命館大学情報理工学部

College of Information Science and Engineering, Ritsumeikan University

宣言的知識は事実に関する知識である。テキストマイニングの知識に照らし合わせると、宣言的知識はよくある例題を解くのに適したツールの選択、およびツールの操作に相当する。テキストマイニングの宣言的知識の伝達については、既存研究においてチュートリアルシステムが提案され、その有用性が示されている [2]。

これに対し手続き的知識は、宣言的知識が経験や練習を通じて変換されることにより得られる知識である [3]。テキストマイニングの知識に照らし合わせると、手続き的知識は例題とは異なり、実社会での問題を解くのに適したツールの選択、およびツールの操作に相当すると考えられる。既存研究においてはテキストマイニングの手続き的知識の伝達を支援したものは少ない。本研究では、提案する機能により、テキストマイニングの手続き的知識である、ツールの選択やツールの操作の伝達を支援する。

2.2 ツールの選択と操作の伝達に関する研究

本節では、ツールの選択と操作の伝達支援に関する研究について述べた上で、本研究の位置付けを説明する。

ユーザ自身の操作履歴を用いて、ツールの選択と操作の支援を行う研究は以前から行われている。携帯端末のメニューを操作履歴に基づきカスタマイズして使いやすくする手法 [4] や、ユーザが自らの繰り返し操作をマクロとして扱いたいときに、操作を視覚的に可視化する手法 [5] が提案されており、自らの操作の中から自動的にマクロを生成することもできる。しかしこれらの研究は、自身の過去の操作を振り返るものであるため、自身が行っていない新しい操作を提示するものにはなっていない。

ユーザ自身の操作ではなく、他者の操作を参考にする研究に、運転操作における他者の視線の履歴を用いて言葉では困難なスキルの伝達を支援する研究 [6] や、算数が苦手な子供に対し、算数が得意な子の模倣をさせることにより、学習への動機付けを与え、学力を高めるための研究 [7] がある。これらは、言葉では伝えにくいスキルに関わる操作を、他者の履歴や模倣によって伝達するものとなっている点では本研究に類似する一方で、運転や勉強においては正解と呼べる理想的な操作が存在している。しかし本研究においては、テキストマイニングによって新たな知識やアイデアを発見する過程を支援することを目的としており、正解となる知識やアイデアが存在せず、またそれに到達するための正解と言える操作が存在するわけではない。探索的なデータ分析によって知識を獲得するためには、さまざまな試行錯誤が必要 [8] で、ユーザにとって未経験の操作として、他者の操作履歴を参考にしてもらうことには意味がある。

ユーザが未経験の他者の操作を参考にできる研究には、CGの作成において他者の操作履歴を可視化して共同作業を支援する研究 [9] や、任意のソフトウェアの操作履歴をワンストロークで表現して可視化し、目的的操作を検索しやすくしたシステム [10] がある。ユーザが新たな操作方法を学んでいくことができる可能性がある点では共通しているが、学ぶべき操作の優先順位を与えるものとはなっていない。

ユーザに提示する情報の優先順位を与える研究としては、アプリケーションを利用した時刻と位置についての利用履歴を元に、時間と場所の情報から推薦するアプリケーションの優先順位を与える研究 [11]、ユーザの過去の行動履歴を元に、ユーザの行動範囲内にあるユーザの興味に合致する場所の優先順位を与える研究 [12]、他者の学習成果物や学習履歴などの蓄積データを元に、ユーザに類似する学習履歴を持つユーザの e ポートフォリオの優先順位を与える研究 [13] などがある。これらの研究においては、ユーザの履歴に合致する情報を提供するための優先順位を与えているが、本研究においては、ユーザの履歴に合致しない、新しい操作を提示する点が異なっている。

すなわち本研究においては、正解と呼べる知識、ならびに正解と呼べる操作があらかじめ分からないテキストマイニングを題材として、ユーザが未経験の操作を重視して他人の操作履歴に優先順位を与えて提示する点が新しい。

2.3 TETDM と操作履歴の再現

本節では、本研究のベースとして用いているテキストマイニングのための統合環境 TETDM と、TETDM を用いた操作履歴の再現機能について述べる。

TETDM はテキストマイニングのためのフリーソフトとして提供¹されており、マイニングのための処理ツールと処理結果を可視化するための可視化ツールとを組み合わせる [1]。処理ツールは約 60 種類、可視化ツールは約 30 種類が用意されており、これらのツールを選択的に利用することで、多様な目的に対応した分析を行うことができる。また、分析結果とその解釈を記録する「結果と解釈」の登録機能、ならびに集めた解釈を統合して知識につなげるための知識創発インタフェースを有している。

この TETDM を利用して、分析操作を他人に伝えるための操作履歴を再現するためのインタフェースが提案されている [14]。操作履歴は、ツールの切り替えを行った時点から、次にツールの切り替えが行われるまでの履歴を 1 つの単位として、その間のマウスの移動履歴、ボタンの操作履歴、「結果と解釈」の登録内容、が保存される。保存された操作履歴は、インタフェースを通じてその操作内容を再現することができる。テキスト分析の操作を知識として伝達することは容易ではないため、先の関連研究に挙げた操作の模倣による伝達を狙ったものとなっている。ただし操作時間が極端に短いものや、何も操作が行われなかったものは履歴として保存されない。

3. テキストマイニング初心者のためのツールの選択と操作を助ける操作履歴の提示機能

本章では、テキストマイニング初心者のためのツールの選択と操作を助ける操作履歴を提示する機能について述べる。本研究においては、この機能をテキストマイニングのための統合環境 TETDM に実装する。

テキスト分析のためには、汎用的な知識につながる手がかりとして、ツールの分析結果を幅広く集める必要がある。しか

1 TETDM サイト: (URL) <https://tetdm.jp>

図1 操作履歴を用いたツールや操作の提示システム

し、ユーザが思いつきでツールを利用しているだけでは、利用するツールに偏りが生じてしまう。またツールの利用時においても、ツールの利用方法が人によって異なる可能性があり、特に初心者においては、あまり利用されないツールが存在していることがある。そのため、ユーザが分析のために行ってきた操作履歴を元に、そのユーザがまだ利用していないツールや操作を提示する機能を実装する。

操作履歴を用いたツールや操作の提示の枠組みを図1に示す。新しい操作を提示してもらいたいユーザは自身の操作履歴をシステムに入力として与える。システムには他のユーザの操作履歴集合がデータベースに集められており、それらの操作履歴の中から、ユーザが利用していないツールやユーザが行っていない操作を含む履歴をユーザに提示する。

このとき、すべてのツールや操作を提示対象とすると提示の幅が広くなり過ぎることになり、分析の目的に合致しない提示が行われてしまう。そのため、特定の分析の目的を設定し、その目的達成のために行われた操作履歴を対象に提示を行う必要がある。また、目的が同じでも分析対象データが異なる場合、行われる分析操作が変わる可能性がある。そこで本研究においては、あらかじめ分析課題を用意し、同一のデータを対象として、その課題を解くための操作履歴集合を用いる。

以下の各節では、テキスト分析の流れとTETDMのユーザインタフェースについて説明した上で、提案する機能の詳細について述べる。

3.1 テキスト分析の流れ

TETDMを用いたテキスト分析は、分析の目的と分析対象となるテキストデータが用意された後、主に以下の手順で行われる。

1. テキストデータの入力
2. ツール選択
3. テキストデータ処理（処理ツール）
4. 処理結果の可視化（可視化ツール）
5. 結果の収集
6. 結果の整理と統合による知識創発

すなわち、1.でデータが入力された後、2.で分析に用いるツールを選択する。TETDMでは、分析のための処理ツールが約60種類、分析結果を可視化するための可視化ツールが約30種類用意されており、それらから選択して利用する。なおTETDMのような統合環境が用いられない場合には、様々な分

図2 TETDMのユーザインタフェース

析が行えるツールを個別に入手して分析処理を行う必要がある。3.の処理ツールによる分析処理をもとに、4.の可視化ツールによる分析結果を眺めながら、5.でデータの背後に隠れている知識（因果関係または相関）を解き明かす手がかりとなり得る分析結果を収集する。最後に6.で集めた分析結果の集合を整理、統合する中で、汎用的な知識を探る。

データ分析において大局的な知識を獲得する手順は、推理小説で探偵が謎を解く様子に喩えることができ、データからわかる手がかり（結果）を1つでも多く集め、それらをもとに推理（整理、抽象化）していくことで大局的な真理（知識）に結び付けることできる。

本研究においては、5.で多様な分析結果を収集するために、2.のツール選択や、3.や4.のツールの操作について、ユーザがまだ利用していないツールや操作を参照できる履歴を提示する。

3.2 TETDMのユーザインタフェース

TETDMのユーザインタフェースを図2に示す。TETDMのウィンドウは、上部にツール全体の操作を行うためのメニューウィンドウがあり、下部に処理ツールと可視化ツールのペアを割り当てて利用するためのパネルが複数配置される（図2では3つのパネルが並んでいる）。

ユーザが行う主な操作は、TETDMの環境を操作するためのメニューウィンドウ内のボタンとパネル操作ボタン、ならびにパネルにセットされたツールを操作するためのツール操作ボタンからなる。

各パネルには、処理ツールと可視化ツールを1つずつセットして利用する。処理ツールは、入力されたテキストデータに対して、何らかのマイニング処理を行うツール、可視化ツールは、マイニングの処理結果を表示するためのツールを表す。表1にTETDMの処理ツールの例を示す。ユーザはこれらのツールの中から、ユーザの分析の目的に関するツールを選択して利用する。

またTETDMには、前節で述べた「5.結果の収集」と「6.結果の整理と統合による知識創発」のための、専用のインタフェースが設けられている。結果を登録するためのインタフェースを図3に示す。このインタフェースは、メニューウィンドウを通

表 1 TETDM の処理ツールの例

ツール名	ツールの処理
文章要約	文章中の重要文やキーワードを表示
主題関連文	文章の主題に関係する文をハイライト
主題関連語	文章の主題に関係する単語をハイライト
主語なし文抽出	文章中で主語がない文をハイライト
長文抽出	文章中の長い文をハイライト
単語冗長文抽出	文章中で同じ単語を繰り返し使っている文をハイライト
あいまい単語抽出	文章中のあいまいな単語をハイライト
失礼単語抽出	文章中の失礼になる可能性がある単語をハイライト
意見理解文抽出	文章中の意見を理解する助けとなる文をハイライト

図 3 結果の登録インターフェース

じて表示させることができ、気になる結果が見つかった時にその結果を登録して記録することができる。ユーザは最後に、このようにして集めた結果について、類似する結果を整理してまとめ、抽象化を繰り返すことで大局的な知識の獲得に結びつける。²

TETDM には初心者向けのチュートリアルが用意されており、これらのインターフェースを含む TETDM の基本操作、ならびに 3.1 節で述べたテキスト分析の流れに沿った分析手法を体験的に学んでもらうことができるようになっている。³ 本論文で提案する操作履歴の提示機能を利用するユーザは、この初心者向けチュートリアルを終了したユーザを想定している。

3.3 操作履歴の提示方法

操作履歴は、利用したツール名、マウスやボタンの操作履歴、分析結果の登録の有無、テキストデータの絞り込み機能の利用の有無から構成される。また、利用するツールが切り替えられた時には、別の操作履歴に分けられる。

ユーザに提示する操作履歴に優先順位を与えるための、操作履歴の評価方法の概要を図 4 に示す。各操作履歴には、新しい分析操作（操作点）、新しいツールの利用（ツール点）、分析結果の登録回数（結果点）、テキストの絞り込み機能の利用（絞り込み点）の 4 項目による評価点を与える。各項目には 100 点満点で評価点を与え、新しい分析操作と、新しいツールの利用は、ユーザの操作履歴とデータベース内の各操作履歴とを比較し、その利用内容の異なる部分について点数を与える。分析結果の登録回数と、テキストの絞り込み機能の利用は、操作履

図 4 操作履歴の評価

表 2 ボタンの評価点

結果の登録に未利用のボタン	1 個	2 個	3 個
評価点	30	40	50

表 3 操作時間の評価点

操作時間	1 分未満	2 分未満	3 分未満	4 分未満	5 分未満	5 分以上
評価点	50	40	30	20	10	0

歴の有効さを表す絶対評価としてそれぞれの実施回数を元に、データベース内の各操作履歴に点数を与える。⁴

これらの指標は、ユーザが参照することでユーザに新たな操作方法の示唆を与えることができる手続きを含む履歴を評価するために用いる。履歴を作成するユーザや、分析対象とするデータによっては、これらの指標の数値が少ないからといって有効な手続きを含む履歴でないことを表すものではないが、履歴を参照するユーザにとって、具体的な操作が多く行われたわかりやすい履歴を提示するための指標とする。⁵ 以下の各節で、各評価指標の詳細について述べる。

3.4 新しい分析操作の評価

操作履歴が含むボタン操作および操作時間に関する評価を分析操作の評価とする。ボタン操作について、単にユーザが利用していないボタンを評価するのではなく、そのボタンが分析に有効に寄与しているかを踏まえた評価を行う必要がある。そこで、操作履歴が含む分析結果の登録の情報を用いて、分析結果の登録が行われた、すなわちユーザが何らかの手がかりを集めることができた操作の履歴を有効な分析が行えた手続きを含む履歴として評価対象とする。

ボタンの評価点を表 2 に示す。ユーザがそれまでの分析操作の中で結果を登録するために用いていないボタンを、他のユーザが結果を登録するために用いている操作履歴について、そのボタン数に応じて評価点を与える。

操作時間の評価点を表 3 に示す。分析操作に要した時間が短

2 後述する本論文の実験では、TETDM の知識創発インターフェースは活用せずに、簡易的に列挙した結果をもとにアイデアをまとめてもらった。

3 例えば、テキスト分析の流れの「5. 結果の収集」として、まずは気になる結果をできるだけたくさん集めること（後述する本論文の実験の課題 1 から課題 3 に相当）、ならびに「6. 結果の整理と統合による知識創発」として、集めた結果を集約して知識としてまとめること（後述する本論文の課題 4 に相当）を体験的に学んでもらう形式となっている。

4 評価点は、各評価指標でおおよそ取りうる値に対して恣意的に決定した。

5 分析対象データや、分析の目的が異なる履歴を集めた場合は、推薦される履歴に偏りが生じる可能性があると考えられるが、現時点では、同一データを用いて、同一の目的に対する分析を行うユーザの履歴集合を推薦対象とした状況を想定している。

表 4 ツールの評価点

結果の登録に未利用のツール	1 個	2 個	3 個
評価点	30	40	50

表 5 分析結果の登録回数の評価点

登録回数	1 回	2 回	3 回	4 回	5 回以上
評価点	50	70	80	90	100

いほど、効果的にツールが操作されて分析が行われていると考えられるため、操作時間が短いほど高い評価点を与える。

3.5 新しいツールの利用の評価

操作履歴で利用されているツールを評価する。ツールについても、単にユーザが利用していないツールを評価するのではなく、ツールの利用によって分析結果の登録が行われた、すなわちユーザが何らかの手がかりを集めることができた操作の履歴を有効な分析が行えた手続きを含む履歴として評価対象とする。

ツールの評価点を表 4 に示す。ユーザがそれまでの分析操作の中で結果を登録するために用いていないツールを、他のユーザが結果を登録するために用いている操作履歴について、そのツール数に応じて評価点を与える。ただし TETDM におけるツールは、各パネルごとに処理ツールと可視化ツールの 2 つがセットされる。評価点は処理ツールと可視化ツールのそれぞれで与えられ、3 つのパネルを利用する操作履歴の場合、操作履歴が含む利用ツールの数は、処理ツールと可視化ツールのそれぞれで最大 3 としてカウントされる。

3.6 分析結果の登録回数の評価

分析結果の登録回数が行われた回数を評価する。分析結果の登録を行ったということは、分析結果から何らかの知識に繋がりそうな手がかりを得たということであり、手がかりが得られる分析を行うことができたと解釈できる。

表 5 に分析結果の登録回数の評価点を示す。分析結果の登録を行った回数に応じた評価点を与える。操作履歴は 3.3 節で述べた通り、ツールの切り替えが行われるごとに分割されるため、1 つの操作履歴の中での登録回数は、その特定のツールの組み合わせを利用した操作によって、より多くの手がかりを得られる可能性があることを示唆するため、高い評価点を与える。

3.7 テキストの絞り込み機能の評価

テキストの絞り込み機能が用いられた回数を評価する。テキストの絞り込み機能とは、入力されたテキスト全体に対する分析ではなく、ある条件を満たす一部のテキストに対する結果を確認するための機能となっている。

テキストの絞り込み回数の評価点を表 6 に、テキストの絞り込みを行ったツール数の評価点を表 7 に示す。テキストの絞り込みを行った回数と、絞り込みを行ったツール数に応じた評価点を与える。絞り込みを繰り返し行っている場合や、複数の

表 6 テキストの絞り込み回数の評価点

絞り込み回数	1 回	2 回	3 回	4 回以上
評価点	30	40	50	60

表 7 テキストの絞り込みを行ったツール数の評価点

絞り込みを行ったツール数	1 つ	2 つ以上
評価点	20	40

ツールで絞り込みが行われている場合、結果を確認するためのより詳細な分析操作が履歴に含まれている可能性があるため、高い評価点を与える。

3.8 新しいツールや操作を含む操作履歴の提示インタフェース

新しいツールや操作を含む操作履歴の提示インタフェースを図 5 に示す。ユーザの操作履歴の入力に対して、前節までに述べた方法により、データベース中の操作履歴に評価点が与えられ、その評価点が高い順に図右側⑤の領域に操作履歴の一覧が表示される。なお表示の順序は、図左側③の領域のボタンにより、重視する評価項目を変更できる。

選択した操作履歴は、ユーザの TETDM の画面上に重畳する形で履歴を動画として再現することができる。すなわち、操作履歴が含むツール構成に TETDM の画面がセットされ、操作履歴が含むマウス操作やボタン操作が自動的に行われる。この際の再現の動画の速度を図左側①のタブで、再現中にユーザのマウス操作を一時的に可能にするために、再現の動画を中断するまでの時間間隔（秒数）を図左側②のタブで設定できる。

履歴の提示に際しては、図左側④の領域で、ユーザが取り組んでいる課題を選択することで、同課題に取り組んだ別ユーザの履歴の提示が行われる。ユーザはこのインタフェースを用いて、参考にしたい操作履歴を選択し、その再現を行うことで、自身の操作の参考にすることができる。

4. 新しいツールと操作の提示機能の評価実験

本章では、前章で述べたテキストマイニングの初心者を支援するための新しいツールと操作の提示機能の効果を検証した実験について述べる。

4.1 実験の設定

本実験では、TETDM のチュートリアル機能により、初歩的な TETDM の操作とテキスト分析の手順を学んだ 9 名の理系の学生を被験者としてシステムの評価実験を行った。いずれの被験者も、TETDM の基本操作を学んだ程度でマイニングの実践経験が乏しく、テキストマイニングの初心者に相当する。

被験者には、「被験者が映画のプロデューサーであるという立場の仮定のもと、新作映画のアイデアを考える」というマイニング課題を行ってもらった。分析対象となるテキストとして、映画のレビューサイト [15] に投稿されたある子供向けアニメの映画のレビューの中から、高評価のレビュー 75 件、低評価のレビュー 75 件をそれぞれ無作為に抽出した。なお各レ

図5 新しいツールや操作を含む操作履歴の提示インタフェース

表8 レビューテキストの分析課題

課題	課題内容
1	レビュー全体を特徴付ける頻度が高い単語を見つけ、その単語がレビュー内でよく使われた理由を調べることで、多くの人がこの映画について抱いている印象を教えてください。
2	〔低評価〕のレビューを特徴付ける頻度が高い単語を見つけ、その単語が〔低評価〕レビュー内でよく使われた理由を調べることで、多くの人がこの映画について抱いている不満を教えてください。
3	〔高評価〕のレビューの中から、この映画についてユーザが不満に思っている点を見つけて教えてください。
4	集めた印象と不満から、〔次回作の制作〕に向けて、どのような改善案が考えられるか教えてください。

ビューには、映画に対する総合評価が1から5の5段階で与えられており、得点が4, 5のレビューを高評価レビュー、得点が1, 2のレビューを低評価レビューとした。

実験は2日間で行い、2日間とも全被験者に共通の同一のレビュー集合に対して、表8の4つの課題に順番に取り組んでもらった。これは、提案機能が同一データで同一目的の操作履歴の利用を前提としていることによる。課題は最も基本的なテキスト分析として、レビューによく含まれているいずれかの単語に着目してレビューの傾向を調べ、次回作の制作に活かせるアイデアを考えてもらうものとした。すなわち、課題1でレビュー全体の傾向を把握するために、レビュー全体の高頻度の単語を元に、映画について抱かれている印象を調べてもらった。課題2で、低評価のレビューについて、低評価につながっている単語を調べ、映画に対する不満を調べてもらった。課題3で、高評価のレビューの中からも、不満が挙げられている点を調べてもらった。課題4で、課題1から3で調べた点を総合して、次回作への改善策を考えてもらった。

課題1から3の回答は、TETDMの「結果と解釈」を登録す

る機能を用いて、そこに答えを記入してもらった。課題4は、集めた結果の一覧を眺めながらアイデアをまとめてもらった。

実験の1日目においては、通常のTETDMを用いて課題に取り組んでもらった。実験の2日目においては、1日目の全被験者の操作履歴を操作履歴データベースに登録した上で、前章で述べた新しいツールと操作の提示機能を用いて課題に取り組んでもらった。すなわち2日目には、各被験者は自分以外の8名の1日目の操作履歴の中から、操作履歴を提示してもらうことができる。なお操作履歴データベースは、課題1から課題3で別々に用意されており、各課題に取り組む際には、該当課題のデータベースから操作履歴の提示を受けることができる。

本実験では、以下の2つの項目を評価する。

1. 提案インタフェースにより、テキストマイニングの新しいツールの利用や操作を促すことができたか。
2. 新しいツールの利用や操作を学ぶことにより、アイデアの幅を広げることができたか。

1. を評価するために、操作履歴の閲覧数と、新しく利用されたツールと操作の数を評価する。これらの数値が大きくなった場合、新しいツールと操作の利用が促進されたと評価する。

2. を評価するために、アイデアの記述文の具体性を評価する。アイデアが具体的に記述されれば、表現に使われる単語の種類数が増えると考えられる。本実験では名詞と形容詞の数を1日目と2日目で比較する。1日目より2日目の方が単語の種類数が増えた場合、アイデアの幅を広げることができたと評価する。

4.2 新しいツールと操作の利用

提案インタフェースにより提示される操作履歴が、被験者の新しいツールの利用につながったかを検証するために、2日目の実験で被験者に提示された操作履歴の上位12件を対象として、被験者が操作履歴を再現して閲覧した回数と、被験者が1

表 9 2 日目の実験における操作履歴の閲覧数と結果の登録のために新しく利用したツールと操作の数

被験者	課題 1			課題 2			課題 3			合計		
	閲覧数	ツール数	操作数	閲覧数	ツール数	操作数	閲覧数	ツール数	操作数	閲覧数	ツール数	操作数
A	6	2	0	5	1	0	5	3	2	16	6	2
B	7	2	1	1	1	0	4	1	2	12	4	3
C	5	0	0	0	0	0	0	1	1	5	1	1
D	5	3	1	0	2	0	2	1	2	7	6	3
E	0	1	1	3	1	3	3	1	3	6	3	7
F	11	2	0	2	2	1	5	3	4	18	7	5
G	5	0	1	8	2	2	1	1	3	14	3	6
H	5	1	1	0	2	0	7	1	3	12	4	4
I	0	0	0	0	0	0	4	0	0	4	0	0
合計	44	11	5	19	11	6	31	12	20	94	34	31

日目に用いなかったツールや操作を利用した回数を表 9 に示す。表 9 の合計値について、閲覧数とツール数の相関係数が 0.73 と強い正の相関があり、閲覧数と操作数の相関係数が 0.36 と弱い正の相関があったことから、操作履歴の閲覧が、新しいツールや操作の利用を促す効果があったことが示唆される。

課題 1 と課題 2 で新しい操作数が少なかった理由としては、課題 1 と課題 2 においては、頻度が高い単語を探すことが起点となっており、頻度が確認できるツールにおいては被験者の操作が必要なかったためと考えられる。一方で課題 3 においては、結果の収集に用いられるツールが特定されておらず、さまざまな操作を試しながら結果を集める必要があったため、他の被験者の操作が多く試されたと考えられる。

また被験者 C や被験者 I においては、ほとんど新しいツールや操作が行われない結果となった。両被験者は、操作履歴の閲覧数も少なかったことから、他人の操作履歴に頼らずに、自身の考えで分析を進めようとしたことが伺える。他人の操作を参考にしたいと考えるのは、自身の操作に自信が持てない時と考えられるため、より難しい分析を行う際には、これらの被験者においても他人の操作を参考にする機会が増えるのではないかと考えられる。

課題 1 においては、ツール「フォーカス情報」を新たに使った被験者が 3 名いた。このツールは単語を選択した時に、その単語を含むレビューをハイライトするツールであり、課題 1 の被験者が着目した単語が使われた理由を探るのに適したツールとなっていた。また課題 2 においては、ツール「キーワード集合比較」を新たに使った被験者が 4 名いた。このツールは単語を選択した時に、レビュー内でその単語とともに使われやすい単語、およびともに使われにくい単語を表示するツールであり、課題 2 の不満を表す単語を、複数の単語の関連を元に探るのに適したツールとなっていた。

また新たに行われた操作の面においても、課題 2 と課題 3 では、レビューに含まれる失礼な単語⁶をハイライトする操作を行った被験者が 6 名、課題 3 では、レビュー中の意見文をハイライトする操作を新たに行った被験者が 4 名いた。すなわち映画の改善点を探るために不満を表すと考えられる失礼な単語を

表 10 課題 4 で被験者がまとめた「新作映画のアイデア」に含まれる名詞または形容詞の種類数

被験者	1 日目	2 日目	増減
A	24	18	-6
B	8	16	8
C	9	27	18
D	21	56	35
E	6	13	7
F	23	26	3
G	12	7	-5
H	30	35	5
I	25	37	12
合計	158	235	77
t 検定	$t = -1.47, df = 16, p = 0.080$		

探す操作、ならびにさまざまなレビュー内のコメントの中から注視すべき意見を確認するための操作が行われたことが確認できた。

以上のことから、提案インタフェースが提示する操作履歴が、新たなツールの利用ならびに操作を促す効果があったことが確認できた。

4.3 実験によって得られたアイデアの比較

被験者が新しいツールや操作を用いることで、課題 4 で回答してもらったアイデアの幅を広げる効果があったかを確認するため、アイデアに含まれる名詞または形容詞の種類数を数えた結果を表 10 に示す。

2 日目の名詞または形容詞の種類数は、1 日目に比べて増加する傾向が見られた（平均値の差の t 検定、 $t(16) = -1.47, p = 0.080$ ）。このことから、提案インタフェースによる操作履歴の提示による、新しいツールや操作の利用によって、提案されるアイデアの内容の幅が大きくなったことがわかる。

また表 11 に、被験者が 2 日目に課題 4 でアイデアの回答に用いた単語（名詞または形容詞）のうち、課題 1 から課題 3 の回答に用いた単語を利用した割合、およびこの割合と、各被験者の表 9 の閲覧数、ツール数、操作数の合計値との相関係数を示す。

表 11 の単語の重複率が高ければ、被験者が課題 1 から課題

6 使用すると一般的に失礼になる可能性がある単語として、ツール「失礼単語抽出」が用いる失礼単語辞書に登録されている単語。

表 11 被験者が 2 日目の課題 4 の回答に、課題 1 から課題 3 の回答に用いた単語を利用した割合（単語の重複率：名詞または形容詞の種類数）と、各被験者の表 9 の閲覧数、ツール数、操作数の合計値との相関係数

被験者	単語の 重複率	閲覧数	ツール数	操作数	ツール数 + 操作数
A	28%	16	6	2	8
B	19%	12	4	3	7
C	0%	5	1	1	2
D	34%	7	6	3	9
E	38%	6	3	7	10
F	38%	18	7	5	12
G	14%	14	3	6	9
H	34%	12	4	4	8
I	14%	4	0	0	0
相関係数	—	0.37	0.72	0.58	0.78

3 で集めた手がかりを元に、課題 4 のアイデアを生成しようとしたことが伺える。逆にこの数値が低ければ、集めた手がかりよりも被験者自身の頭の中で論理を展開してアイデアを生成しようとしたことが伺える。また、表 11 の単語の重複率との相関係数の値によると、操作履歴の閲覧数が多いだけでアイデアの生成につながる手がかりを獲得できるわけではなく、操作履歴を閲覧した上で実際に新しいツールや操作を試すことで、アイデアの生成につながる手がかりを獲得できる可能性が示唆される。

表 11 で最もツール数と操作数の合計が多かった被験者 F は、1 日目に比べて課題 1 から課題 3 の回答を踏まえたアイデアを生成することができていた。また被験者 D や被験者 E は、閲覧数が少ないながらも多くのツールや操作を新しく試すことで、被験者 F と同様に課題 1 から課題 3 の回答を踏まえたアイデアを生成することができていた。

一方で、他人の履歴を閲覧せず新しいツールや操作をほとんど行わなかった被験者 C と被験者 I については、表 10 においてアイデアの記述の具体性が増しているものの、課題 1 から課題 3 の回答内の単語をあまり用いていなかった。すなわち、多様な結果を元にアイデアをまとめようとするのではなく、自身の頭の中で論理を展開してアイデアを生成しようとしていたことがうかがえる。

これらのことから、提案するインタフェースを用いてさまざまな操作履歴を確認した上で、実際に新しいツールや操作を試すことが、多くの結果を元に多様なアイデアを生成することにつながる可能性が示された。

4.4 操作履歴の質について

本実験においては、参照の対象となる操作履歴が初心者の操作に限られている。初心者以外の操作履歴が含まれる場合の利点としては、より分析に効果的なツール選択や分析操作が含まれる可能性が挙げられる。逆に初心者でないユーザによる操作履歴においては、ツール選択や分析操作の意図が初心者には伝わらない可能性が生じる。本実験においては、同じ初心者の操作履歴として操作内容が伝わりやすく、操作履歴の再現に基づ

いてそれを模倣することが容易だったと推測される。

一方で初心者の操作履歴においては、効率が悪かったり無駄な分析が行われている可能性がある。表 9 の新しく利用されたツール数や操作数の数は、閲覧数の 3 分の 1 程度に留まっており、すべての操作履歴が即座に新しいツールや操作の利用につながっているわけではないことが伺える。⁷ 現時点では、操作履歴間の類似度は考慮していないため、全く同一の操作履歴があったとしても、同じ評価点が与えられてしまう。そのため、操作履歴間の類似度に基づいて、提示する履歴に多様性を持たせることを検討していきたい。

模倣によって操作を学んでもらうためには、一般的に熟練者の操作を模倣してもらうことが効果的な場合が多い。しかし、2 章で述べたように、本論文が題材とするテキストデータの分析においては、発見すべき正解となる知識、あるいは知識に結び付けるための正解となる分析操作が存在しない。これはアイデアの生成とその適用の効果が、未来の予測を含んでいることによっている。また、分析するユーザの背景知識に応じて、同じ出力を目にした場合でも異なる考えをもつことが考えられるため、ある熟練者の操作が、必ずしも初心者にも効果的とは限らない状況がある。

また、同一データで同一の分析の目的に対する熟練者の操作履歴は、実際のデータ分析の場面では用意できない、あるいは用意する意味がない可能性が考えられる。例えば企業や組織の中で、あるデータの分析を試みる場合に、熟練者がいるのであれば、その熟練者がリーダーシップをとって分析を行えば良いため、本実験のように熟練者がいない状況でも、効果的に分析が行える環境として提案機能を活用できる可能性を示したことは意味があると考えられる。

加えて、データ分析の初心者が熟練者となるためには、特に望ましい操作だけを学ぶのではなく、失敗も含めた様々な操作を含めて試行錯誤しながら体験的に学ぶことも重要と考えられる。そのため、ユーザが未経験の熟練者でない他のユーザの操作を様々に試すことができる環境としても意味があると考えられる。

5. おわりに

本研究では、テキストマイニングのための統合環境 TETDM を用いて、テキストマイニングの初心者に対して、ユーザが用いていないツールや操作を含む操作履歴を提示するインタフェースを提案した。提案インタフェースが、ユーザの新しいツールや操作の利用を促進し、マイニングの最終結果として得るアイデアの内容の幅を広げられる可能性を実験により検証した。

今回の課題として、参照の対象となる操作履歴の質を改善するために、初心者以外の操作履歴を含む場合の提示方法や、操作履歴間の類似度を元に、多様な操作履歴が提示されるように改良していくことが挙げられる。

7 2 日目の課題 3 で各被験者が結果の登録に用いたツールの種類数は 1 種類から 3 種類、被験者全体でのべ 8 種類となっており、同一データに対する操作履歴の閲覧によって、被験者の行動が一方に制約されることはなかった。

また、集められた操作履歴のいずれかを、実際の分析対象データに適用する自動分析や、より多くの操作履歴が集められる環境においては、多くの操作履歴に共通する分析操作を典型的な分析手順として抽出し、それを実際の分析対象データに適用する自動分析が検討できる。加えて、閲覧対象となっている各操作履歴の典型的な分析手順からのずれを明示することにより、閲覧する操作履歴の選択支援も検討していきたいと考えている。

謝辞

本研究の一部は科研費（16K00307）の補助を受けて行われました。記して謝意を申し上げます。

参考文献

- [1] 砂山渡, 高間康史, 徳永秀和, 串間宗夫, 西村和則, 松下光範, 北村侑也: “統合環境 TETDM を用いた社会実践,” 人工知能学会論文誌, Vol.32, No.1, pp. NFC-A_1-12, 2017.
- [2] 西原陽子, 中垣内李菜, 川本佳代, 砂山渡: “TETDM を用いたテキストマイニングのスキル獲得を支援するためのチュートリアルシステムの開発,” 知能と情報, Vol.27, No.5, pp. 771-783, 2015.
- [3] 松尾睦, 細井謙一, 吉野有助, 楠見孝: “営業の手続的知識と業績,” 流通研究, Vol.2, No.1, pp. 43-57, 1999.
- [4] Y. Fukazawa, M. Hara, M. Onogi, and H. Ueno: “Automatic mobile menu customization based on user operation history,” *Proc. of the 11th Int. Conf. on Human-Computer Interaction with Mobile Devices and Services*, Sep. 15-18, Bonn, pp. 1-4, 2009.
- [5] D. Kurlander and S. K. Feiner: “A history-based macro by example system,” *Proc. of the 5th Annual ACM Symp. on User Interface Software and Technology*, Nov. 15-18, Monterey, pp. 99-106, 1992.
- [6] 藤本武司, 砂山渡, 山口智浩, 谷内田正彦: “視線行動の可視化による着眼スキル伝達支援,” 人工知能学会論文誌, Vol.19, No.3, pp. 174-183, 2004.
- [7] 江川克弘: “グループ学習で学習苦手手が他の成員を模倣することの有効性の検討—小学校算数科の授業を通して—,” 教授学習心理学研究, Vol.7, No.1, pp. 21-37, 2011.
- [8] 松下光範: “InTREND: ユーザの探索行為と振り返り行為に着目したデータ分析支援システム,” 情報処理学会論文誌, Vol.49, No.7, pp. 2456-2467, 2008.
- [9] G. Salvati, C. Santoni, V. Tibaldo, and F. Pellacini: “MeshHisto: collaborative modeling by sharing and retargeting editing histories,” *ACM Trans. on Graphics*, Vol.34, No.6, pp. 205:1-205:10, 2015.
- [10] T. Nakamura and T. Igarashi: “An application-independent system for visualizing user operation history,” *Proc. of the 21st Annual ACM Symp. on User Interface Software and Technology*, Oct. 19-22, Monterey, pp. 23-32, 2008.
- [11] 松本光弘, 清原良三, 沼尾正行, 栗原聡: “携帯電話における時空間的利用履歴を基にしたアプリケーション推薦システム,” 情報処理学会研究報告 (MPS), Vol.72, pp. 53-56, 2008.
- [12] 倉島健, 岩田具治, 星出高秀, 高屋典子, 藤村考: “行動範囲と興味の同時推定モデルによる地域情報推薦,” 情報処理学会論文誌データベース, Vol.6, No.2, pp. 30-41, 2013.
- [13] 植野真臣: “過去の学習履歴データを利用した e ポートフォリオ・システム,” 情報知識学会誌, Vol.24, No.4, pp. 414-423, 2014.
- [14] 金谷直哉, 砂山渡: “ツールの操作履歴の共有によるテキストマイニング時の発想支援,” 人工知能学会第 12 回インタラクティブ情報アクセスと可視化マイニング研究会, pp. 29-34, 2016.
- [15] Yahoo! JAPAN 映画: <https://movies.yahoo.co.jp> [accessed Jan. 10, 2019]

(2019 年 12 月 24 日 受付)
(2020 年 7 月 21 日 採録)

[問い合わせ先]

〒522-8533 滋賀県彦根市八坂町 2500
滋賀県立大学工学部
砂山 渡
E-mail: sunayama.w@e.usp.ac.jp

著者紹介

すなやま わたる [正会員]
砂山 渡

1995 年大阪大学基礎工学部制御工学科卒業。1997 年同大学院博士前期課程修了。1999 年同大学院博士後期課程中退。同年同大学院助手。2003 年広島市立大学助教授。2007 年同准教授。2016 年滋賀県立大学教授。現在に至る。博士 (工学)。滋賀県立大学地域ひと・モノ・未来情報研究センター兼務。人間の創造活動を支援する研究に興味を持つ。人工知能学会, 電子情報通信学会, 言語処理学会, IEEE, 各会員。

なかがわ つよし [非会員]
中江 剛士

2017 年滋賀県立大学工学部電子システム工学科卒業。2019 年滋賀県立大学大学院工学系研究科博士前期課程修了。現在, 株式会社 NTT データ MSE 勤務。

にしはら ようこ [非会員]
西原 陽子

2003 年大阪大学基礎工学部卒業。2005 年同大学院基礎工学研究科博士前期課程修了。2007 年同研究科博士後期課程修了。博士 (工学)。日本学術振興会特別研究員 (DC1, PD) を経て。2008 年東京大学大学院工学系研究科助教。2009 年同講師。2012 年立命館大学情報理工学部准教授。現在に至る。立命館大学食総合研究センター兼務。ヒューマンコンピュータインタラクション研究。自然言語処理研究に興味を持つ。情報処理学会, 人工知能学会, 日本感性工学会, 各会員。

はたなか りょうじ [非会員]
畑中 裕司

2002 年岐阜大学大学院工学研究科博士後期課程修了。博士 (工学)。同年岐阜高専助手。2007 年同高専准教授。2008 年滋賀県立大学工学部准教授。現在に至る。画像認識, 医用画像解析, 組合せ問題, バーチャルリアリティ, 感性情報処理などの研究に従事。日本医用画像工学会, 医用画像情報学会, IEEE, 画像電子学会, 電子情報通信学会, 日本生体医工学学会, 日本放射線技術学会, 各会員。

**Presentation System of Operating Histories to Support Choice and Operation of Tools
for Beginners of Text Mining Software TETDM**

by

Wataru SUNAYAMA, Tsuyoshi NAKAE, Yoko NISHIHARA, and Yuji HATANAKA

Abstract:

Data analysis has been conducted in business operations in many fields. Analysis of text data like comments of questionnaires and reviews on the Web is also required. Most of the data analysis is conducted by using the software. The data analysts need to acquire knowledge of the software usage and mining operations. Some of the software provide instructions and tutorials about how to use. However, if the analysts are beginners of data analysis, the instructions and tutorials can not tell how to choose appropriate tools and how to operate the tools for their analysis purpose. This paper proposes a presentation system of operating histories including new tools and operations to support for beginners of text mining. We use TETDM as the text mining software in this paper. We experimented the presentation system to verify the effect of the system on tool usage. The experimental results showed that the system could support the participants to use new tools to obtain analysis results.

Keywords: text mining, data analysis support, data scientist training, operating history, TETDM

Contact Address: **Wataru SUNAYAMA**

*School of Engineering, The University of Shiga Prefecture
2500 Hassaka-cho, Hikone-shi, Shiga 522-8533, Japan
E-mail: sunayama.w@e.usp.ac.jp*