分析対象と目的に基づき データ分析手法の選択を支援するインタフェースの検討

Supporting Interface for Choosing Text Mining Methods of TETDM Designed for Analysis Unit and Purpose

西原陽子 [1*]　深井俊樹 [1]　山西良典 [1]
Yoko Nishihara[1]　Toshiki Fukai[1]　Ryosuke Yamanishi[1]

[1] 立命館大学情報理工学部
[1] College of Information Science and Engineering, Ritsumeikan University

Abstract: Text data analysis can be conducted by using text mining software. Instructions and tutorials of the software tell users how to use the software. TETDM, one of the text mining software, also has tutorials for users. Even if the users are the beginners of text mining or data analysis, they would be able to use TETDM after finishing the tutorials. However, the users need much time to finish all of the tutorials because there are many quizzes that can be solved by using tools of TETDM. Moreover, though the users can do text mining by referring the quizzes of tutorial, they also need much time to find quizzes that are related to their own analysis unit and purpose. We believe that it is necessary to support users for finding appropriate tools in shorter period. This paper proposes a supporting interface for choosing text mining methods of TETDM considering analysis unit and purpose. Once a user decides the analysis unit and purpose of text mining, the interface shows appropriate methods of TETDM. We had evaluation experiments with the proposed interface. We asked participants to answer questions of text mining by using the proposed interface. Experimental results showed that the participants answered questions more correctly and in shorter periods.

1　はじめに

多くの分野で収集されたデータを分析し，業務に活用しようとする動きが進められている．データの分析が手作業ではなく，パソコン上のツールを用いて行われることも増えてきた．大学においてもデータサイエンスの講義や演習が開講されるようになり，ツールを用いたデータ分析のスキルを身につけることの需要が高まってきている．

パソコン上のツールを用いてデータ分析を行うと，大量のデータを短時間で処理することができるようになる．しかし，データ分析を行うためには，目的や対象に対して適切なツールを選択し，ツールを使いこなす必要がある．仮に分析する人がデータ分析の初心者であると，自分の分析したい対象や目的にとって最適なツールがどれかが分からず，分析が進められないこともあると考えられる．

パソコン上のツールには説明書やチュートリアルが付属していることがある．ツールの説明書の多くは，ツールの使い方を説明してくれ，チュートリアルの多くはサンプルデータを用いて，ツールの分析結果の見方などを説明してくれる [3].

テキストマイニングのソフトウェアである TETDM にもチュートリアルがあり，スーパーライトモード，ライトモード，通常モード，拡張モードと 4 種類のチュートリアルが用意されている．テキストマイニングやデータ分析の初心者であっても，チュートリアルを全てクリアすると，TETDM が使えるようになるとされている．しかし，チュートリアルの問題数は多く，全てをクリアするまでに多くの時間がかかってしまう．また，チュートリアルを参考にして，TETDM のツールを使おうとすると，ツールの使用が説明されているチュートリアルの箇所を探す必要があり，時間がかかってしまう．ユーザが分析したい対象と目的を持ち，ツールを使いたいと思っても，その説明がある場所を探すことが大変になっている．より短時間でツールを探せる仕組みが必要と考えられる．

そこで本研究では，ユーザの分析したい対象と目的に

*連絡先：立命館大学情報理工学部
滋賀県草津市野路東 1-1-1
E-mail:nisihara@fc.ritsumei.ac.jp

応じて，テキストマイニング手法の選択を支援するインタフェースを提案する．本研究ではテキストマイニングのソフトウェアである TETDM のツールの選択を支援することとする．分析したい対象は，本研究では単語，文，段落，文章の４つのうちのいずれかとする．選択肢として提示するテキストマイニング手法は，TETDM に含まれている手法のいずれかとする．TETDM のチュートリアルを利用する場合と，提案インタフェースを利用する場合を比較し，提案インタフェースにより短時間で分析ツールにたどり着け，分析が行えることを目標とする．

多くの選択を支援するシステム，ホテルの検索システム（例えば，https://www.booking.com/）やレストランの検索システム（例えば，https://www.gnavi.co.jp）などにおいては，ユーザが必要とする条件を入力させることにより選択肢を絞り込み，絞り込んだ選択肢をユーザに提示している．本研究でも同様に，ユーザが必要とする条件として分析したい対象と目的の２つを入力させて，利用可能なツールを絞り込みユーザに提示することとする．

ユーザの嗜好を捉えた上で選択支援を試みる研究も多数行われている．階層分析法（AHP）を用いた研究 [1, 4] や，ユーザの嗜好の点数付けから評価軸を作成する研究 [2] などもある．本研究ではデータ分析の初心者のためのツール選択支援システムについて検討をする．初心者はツールを熟知しておらず，熟知していない状態ではツールに対する嗜好は存在しないと考えられる．したがって，嗜好を用いた選択支援については対象とせず，データ分析に必要な条件が入力されたときの選択支援を対象とする．

2　提案インタフェース

提案インタフェースの概要を図 1 に示す．提案インタフェースは，サイドメニュー，メイン画面の２つから構成される．サイドメニューの一番上には分析対象である「単語」「文」「段落」「文章」の４つが示されている．ユーザが分析したい対象を選択すると，図 2 に示すように，分析目的が示される．分析目的を選択すると，サイドバーに利用可能な手法が表示される．ユーザが手法を選択すると，手法に必要な TETDM のツールにポップアップがつけられる．ポップアップをクリックすると，図 3 に示すように，手法の実行例が表示される．提案インタフェースを用いることにより，ユーザは分析対象と目的に合致する手法を選択していくことが可能になる．

図 1: 提案インタフェース概要．分析対象や目的を表示するサイドメニュー（左）と，対象や目的に利用可能なツールを表示するメイン画面（右）の２つにより構成される．

図 2: 分析目的の表示例．図では対象として「文」が選択された時に，目的として「特定の単語の抽出」「単語間の関係の抽出」「単語の情報の抽出」「その他」が表示されている．

図 3: 手法の実行例の表示例．図 1 のメイン画面に表示されているポップアップをクリックすると，手法の実行例が表示される．

2.1 分析手法の分類方法

ユーザがテキストマイニングの対象と目的を入力した時に，利用可能な手法を表示するため，TETDM 内の手法を分類する．著者らは初めに，TETDM に含まれる手法を分析対象により分類し，その後分析目的により分類を行った．本研究における分析対象は単語，文，段落，文章の 4 つになり，TETDM に含まれる手法をまず 4 通りに分類した．続いて，分析目的により詳細な分類をした．本研究における分析目的は対象の属性，対象間の関係，対象が持つ情報の 3 つになる．分析対象と分析目的による分類クラスを表 1 の左 2 列に示す．

TETDM ではデータを分析する際に，分析手法と分析結果の可視化手法と 2 つの手法を設定する必要がある．分析目的ごとに手法のペアが設定されており，分析手法と可視化手法を分けて分類することが困難であった．そこで，本研究では分析手法にだけ注目し，分析対象と目的を考慮して分析手法のみを分類した．分析手法のペアとなる可視化手法は，分析手法が分類された先に自動的に振り分けた．

手法の分類結果を表 1 に示す．分析対象ごとに分析手法を分類し，その後，分析目的ごとに分析手法を分類した．共通する分析目的が存在しなかった分析手法については，その他というクラスを設け，そこに分類した．

3 評価実験

提案インタフェースによる選択支援について評価する実験を行なった．実験は以下の手順で行なった．

表 1: 手法の分類結果

対象	目的（抽出物）	手法の種類数（例）
単語	特定の単語	9（専門用語抽出など）
	単語間の関係	2（単語間関連度など）
	単語の情報	7（単語情報まとめなど）
	その他	3（辞書オンラインなど）
文	特定の文	6（意見文抽出など）
	文間の関係	0
	文の情報	1（文情報まとめ）
段落	特定の段落	1（テキスト分類）
	段落間の関係	6（段落順序評価など）
	段落の情報	2（セグメント情報まとめ）
	その他	1（テキスト集合評価アプリ）
文章	特定の文章	0
	文章間の関係	0
	文章の情報	5（テキスト評価など）
	その他	3（テキスト評価アプリなど）

1. 実験者は被験者を実験群と統制群に分ける
2. 各群の被験者は，TETDM のチュートリアルを 2 日間に渡って行う
3. 各群の被験者は，実験者が出題するテキストマイニングの問題に回答する

実験者は第 2 著者であった．被験者は情報理工学部に所属する大学生 14 名であった．実験群 7 名，統制群 7 名であった．TETDM を利用した経験はなく，データ分析の経験も少ない被験者であった．

TETDM のチュートリアルは 4 種類あり，それぞれスーパーライト，ライト，通常，拡張モードとなっている．1 日目にスーパーライトとライトモードを行い，2 日目に通常と拡張モードを行なった．チュートリアルにはツールの説明以外にも，データの入力方法や TETDM の機能についても説明があり，TETDM を用いたデータ分析を行うためにはチュートリアルを行うことが必要であった．そのため，実験群と統制群の両群がチュートリアルをこなした．

手順の 3. で問題に解答する際，実験群は提案インタフェースを用いた．統制群は TETDM のみを用いた．

テキストマイニングに用いた文章を表 2 に示す．学生が興味を持ちやすいタイトルで，文字数が約 1000 ずつ異なる文章を 7 本用意し，実験を行なった．

出題した問題を表 3 に示す．これらの問題はテキストマイニングの一般的な問題として用意した．

表 2: 実験で用いた文章のタイトル

番号	タイトル	文字数
1	あおり運転の心理	2118
2	ガンにならないのはどんな人か？	3177
3	「すぐに返信しない男」と「既読スルーを我慢できない女」の脳の違い	4323
4	ゲームのやりすぎは本当に「精神障害」なのだろうか？	4470
5	いつまでも消えないタバコ「7 つの神話」の真実を明かそう	7626
6	ウォーキング･デッドのストーリーまとめ	9503
7	ハリーポッターシリーズのストーリーまとめ	21068

表 3: 実験で出題した問題

問題番号	問題文
1	最も出現頻度が高い単語は何か？単語名をこたえよ.
2	第 3 段落で tfidf 値が 1 番高い単語はどれか？単語名をこたえよ.
3	関連度が T 以上のとき，単語 X と関連のある単語はいくつあるか？単語の個数を答えよ.
4	最も独自性の高い段落は何段落目か？段落数をこたえよ.
5	単語 Y に関わっている段落はいくつあるか？段落の個数を答えよ.
6	最も複数の段落と関連（類似）しているのは何段落目か？段落数を答えよ.
7	主題に関連した文は文章中（前半，後半）どちらに多く登場するか？

表 4: 問題の正解率

問題番号	実験群	統制群
1	100%	100%
2	71%	85%
3	57%	0%
4	86%	57%
5	71%	42%
6	28%	0%
7	100%	85%
平均	73%	53%

表 5: 問題の解答時間

問題番号	実験群	統制群
1	58 秒	68 秒
2	98 秒	78 秒
3	515 秒	711 秒
4	201 秒	265 秒
5	252 秒	642 秒
6	361 秒	711 秒
7	164 秒	231 秒
平均	236 秒	387 秒

3.1 実験結果

問題の正答率を実験群と統制群に分けて表 4 に示す．平均の正答率は，実験群が 73%，統制群が 53%であり，実験群のほうが正答率が高かった．

続いて，問題の平均解答時間を実験群と統制群に分けて表 5 に示す．平均の解答時間は，実験群が 236 秒，統制群が 387 秒で，実験群のほうが解答時間が短かった（t=-4.84, p=0.0008 $<$ 0.01）.

3.2 考察

問題の正答率は実験群の方が統制群よりも高かった．実験群は分析する単位と目的を問題文から読み取り，適切な手法を選択できたため，正答率が高くなったと考えられる．

回答時間は実験群の方が統制群よりも短かった．実験群は適切な手法を選択でき，解答が得やすかったため，回答時間が短くなったと考えられる．

以上の結果から，TETDM でのツール選択に提案インタフェースを利用することで，テキストマイニングの問題が正しく解け，さらに短時間で解けることが示された．

4 おわりに

本研究では，テキストマイニングのソフトウェアである TETDM を用いてデータ分析をするときに，手法の選択を支援するためのインタフェースを提案した．提案インタフェースでは，ユーザが分析したい対象と目的を有した時に，それに合致する手法を表示し，選択の支援を行うものである．評価実験を行ったところ，TETDM を使用するときに提案インタフェースを用いることにより，テキストマイニングの問題の正答率が高くなり，また回答時間も短くなることが確認された．本研究では TETDM を対象として研究を進めたが，高度なソフトウェア、システムのユーザを支援する方法についての研究であり，類似する他のソフトウェアにも応用できる知見が得られたと考えている．チュートリアルを解くことにより，TETDM の使い方を一通り学ぶことができるというのは変わらず，今後は提案インタフェースと TETDM のチュートリアルを併用し，データ分析ができる環境づくりを進めていきたい．

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] 井上拡子，原田利宣，ファジィAHP を用いた製品レコメンドシステムの開発，日本感性工学会論文誌，Vol.11, No.2, pp.255-263, (2012).

[2] 西原陽子，赤井れい子，砂山渡，橘啓八郎，積極的思考支援のためのキーワード選好インタフェース，Vol.18, No.5, pp.766-776, (2007).

[3] 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).

[4] 湯本真樹，定性的評価にラフ集合の決定ルールを用いた AHP による商品選択支援システムの開発，電気学会論文誌C，Vol.139, No.9, pp.1080-1091, (2019).