人工知能学会 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(第6回)
SIG-AM-06-07

探索的データ分析支援に向けた TETDMインタフェースの改良に関する基礎検討

Basic study on an Improvement for TETDM Towards the Goal of Supporting Exploratory Date Analysis

井須 弘恵[1] 大塚 直也[2] 松下 光範[1*]
Hiroe Isu[1] Naoya Otsuka[2] Mitsunori Matsushita[1]

[1] 関西大学 総合情報学部
[1] Faculty of Informatics, Kansai University

[2] 関西大学大学院総合情報学研究科
[2] Graduate School of Informatics, Kansai University

Abstract: Knowledge discovery on text mining requires a trial-and-error process so that a user's informational requirements are unclear when they start his or her exploration. Our purpose is to support a user's information seeking behaviour on text mining. In this paper, we observe how a user behave on TETDM : Total Environment for Text Date Mining. According to the experiment, we found that 5 usability problems and 1 problem for TETDM. From obtained results, we sort the system requirements and propose that a design criteria to facilitate a user's information seeking.

1 はじめに

近年、構造化されていないテキストデータから新たな情報や知識を発見するための分析手法として、テキストマイニングが注目されている。テキストマイニングとは、文書集合から新しい情報や知識を発見することのできる「構造化されていないテキストデータからの情報抽出に関する技術」である [1]。

テキストマイニングによる知識発見は、あらかじめ分析のゴールが明確に定まっているようなゴール指向のタスクではなく、有益な情報を発見する探索的な情報アクセス [2] を必要とするタスクと据えることができる。Hearst によると、テキストマイニングのゴールは、データから新たな情報を発見かつ誘導し、 何かしらのパターンが存在するかを調べることである [3]。テキストマイニングは、単なる情報検索の技術、発話解析、語義曖昧性の解決、辞書作成などの自然言語処理技術やテキスト要約などの技術にあるのではなく、それらを利用した「探索的データ解析」 にある。

テキストマイニングの技術は既に多くの研究成果が報告されているが、実際に世の中で使われる技術は一部に限られている。この問題を解決するために、Total Environment for Text Data Mining(以下 TETDM と略す) が提案されている [5]。 TETDM とは、世の中に分散しているテキストマイニングの技術を同一環境上で柔軟に組み合わせて分析を行うことができる統合環境である。現在 TETDM の統合環境は開発途中であり、そのインタフェースはテキストマイニングのような試行錯誤を伴う探索的な情報アクセスに必ずしも適したものではないことが指摘されている [4]。

そこで本研究では、テキストマイニングにおける知識発見のためのユーザの探索行為を支援するために、TETDM のインタフェースが満たすべき要件を整理することを目的とする。そのために、実際に TETDM を用いて、ユーザの情報探索行動を観察する。得られた結果から TETDM が抱える問題点を整理する。また、TETDM インタフェース改良の先行研究として提案されている大塚らのインタフェース (図 3 参照) の有用性を評価するために、ユーザ観察を行う [4]。それら 2 つの実験の結果から、ユーザの探索行為の円滑化のために満たすべき要件を整理し、TETDM のインタフェースの改良指針を提案する。

2 関連研究

2.1 TETDM の概要

TETDM は複数のウィンドウで構成されており、それぞれの画面で異なった分析および結果を表示することが可能である。TETDM ではテキストマイニングのプロセスを、自然言語処理、データマイニング、情報可視化の 3 つのプロセスとして捉えている。TETDM に入力されたテキストは、形態素解析などの前処理の後、各モジュールによって処理が行われる。モジュール

*連絡先：関西大学総合情報学部総合情報学科
〒 569-1095 大阪府高槻市霊山寺町 2-1-1
E-mail: mat@res.kutc.kansai-u.ac.jp

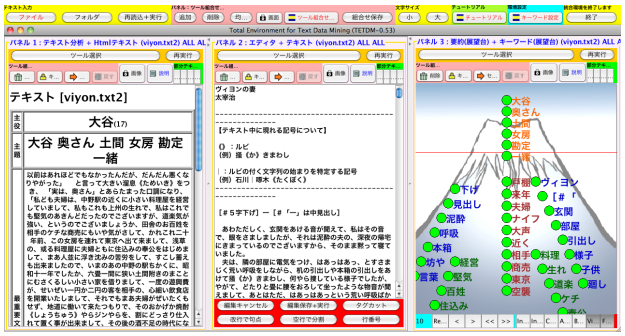

図 1: TETDM の概観

とは、テキストマイニングを行う機能ごとにまとめられた要素であり、これを TETDM に追加していくことにより、扱える技術を拡張することができる。ユーザは、自らの興味や目的に応じてツールを選択してそれらを自由に組み合わせたり、処理結果を比較することによって、多角的なテキスト分析を行うことができる。

TETDM では、マイニングツールと可視化ツールの 2 つの技術が統合環境内のモジュールとして実装されており、それらをユーザが切り替えることで処理内容と処理結果の表示手法を変更することができる。そのため、TETDM を用いて分析を行う場合、ユーザはマイニング処理ツールと可視化ツールをそれぞれ 1 つずつ選択し、組み合わせて結果を比較しながら分析を進めていく。ユーザは様々な分析手法の中から適当な手法を選択して分析を行い、得られた結果を解釈・考慮して次の分析手法を試みる、といった探索プロセスを繰り返す。

現在のバージョン 0.54 では、利用可能なマイニング処理ツールと可視化ツールはいずれも 28 種類ずつであるが、組み合わせのパターンによって 45 種類の分析が可能である。マイニング処理ツールと可視化ツールは単体では機能せず、それぞれ対となるモジュールを必要とする。図 2 は、ユーザがマイニング処理ツールを選択する場面において、「光と影」が選択された状態である。左の列がマイニング処理ツールのリスト、右の列が可視化ツールのリストである。オレンジ色で示されているものが、 推奨されたツール同士の組み合わせである。マイニング処理ツールである「光と影」に対しては、「キーワード選択」、「スコア分布」、「テキスト (カラー)」の 3 種類の異なった可視化結果が推奨されている。

2.2 試行錯誤を支援するインタフェース

前節でも述べたように、テキストマイニングの本質は探索的なアプローチにあり、分析を進めていく中で、自分の求める情報要求を精緻化・明確化していく。しかし大塚らは、現在の TETDM インタフェースは必ず

図 2: 「光と影」を選択した状態

しもこのような探索的な分析に適していないことを指摘し [4]、ユーザの試行錯誤の円滑化を目的とした新しいインタフェースを提案している (図 3 参照)。大塚らは、モジュール切り替えのためのインタフェースは、直感的な操作をできることが望ましいとした。また、複数のテキストマイニングツールを用いて多面的に分析を行うことを想定して、分析内容が多岐に渡り、作業が煩雑になる可能性があると指摘した。これを解決するために、ユーザは自身の分析プロセスを常に把握し、現在行っている分析に引き続いてどのような分析ができるのかを把握できることが必要であると考えた。このような要求に応えるため、提案インタフェースでは、各モジュールをノード、それらを処理の順に繋ぐ線をリンクとするグラフ表現が採用されている (図 3 左参照)。提案手法では、各モジュールを表すノードをマウスで直接操作することによって、ツールの選択・切り替えを行えるようになっている。このようにオブジェクトを直接操作することにより、直感的なモジュール切り替えを実現でき、ユーザは試行錯誤を妨げられることなく分析を進めていくことができると考えられる。

このような先行研究を受けて、本研究では、TETDM インタフェースが満たすべき要件を明らかにするために、TETDM を実際に用いた実験を行い、ユーザの情報探索行動を観察する。また、大塚らのインタフェースの有用性を測り、TETDM インタフェースの満たすべき要件を明らかにするために実験を行う。

3 従来インタフェースを用いた観察

上述したように、本研究の目的は、TETDM を用いてユーザの情報探索を観察し、 TETDM インタフェースが探索的データ分析を行うユーザを支援するにあたって考慮すべき点を発見することである。 探索的データ

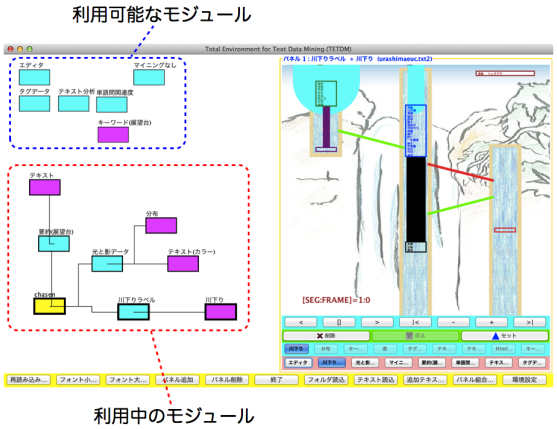

図 3: 大塚らによる提案インタフェース [4]

分析では、ユーザによってその分析方法が異なるため、同じデータを分析対象にした場合でも、すべての分析者が同じ結論にたどり着くことは期待できない。そのため、その行為を支援するシステムの評価においては、タスクの達成度や正答率といった指標を用いることができないことから、本実験では、ユーザが実際に TETDM を使用し、ユーザがどのように分析を行うかについて観察を行う。

3.1 実験の概要

実験では、 TETDM (ver. 0.53) を用いてユーザ観察を行った。実験参加者は情報学を専攻する大学院生 2 名（男性 2 名)、4 年生大学の情報学部に通う 1 名 (男性 1 名)、 3 年生 1 名（女性 1 名）と社会人 1 名（女性 1 名）の計 5 名であった。事前説明では、TETDM インタフェースの改良を目的としていることを実験参加者に説明し、ユーザの立場から問題点や改善点を指摘することを求めた。実験の課題は、TETDM を用いて小説『ヴィヨンの妻（太宰治著)』から登場人物を発見し、発見した人物の特徴 (性格や年齢、職業など) について説明することとした。実験は、時間制限を設けず、参加者の考えがまとまった時点で終了とした。これは、テキストマイニングがゴールがある課題ではなく、一定の結論に至ったと自覚することがゴールとなるような課題と捉えられるためである。実験を始める前に、参加者に対して、 TETDM の使用経験と、課題とする小説の読書経験について確認を行ったところ、参加者全員が過去に TETDM を使用したことがなく、課題の書籍が未読であることが確認された。参加者は実験中にメモを取ることが許され、実験中の様子は VTR で記録された。また、実験終了後に、分析を進めるにあたって障害となった点を聞き取るために、 (1) どのように分析を進めたか、 (2) 分析を進める上で障害と感じた点はあるか、の 2 つの質問を用意した。

3.2 実験の結果

以下ではユーザ観察を行った順に、参加者を A, B, C, D, E と記す。実験参加者が TETDM を操作する様子を観察したところ、TETDM インタフェースの 4 つの問題が明らかになった。

ツールの組み合わせに関する問題

TETDM では、マイニング処理ツールと可視化ツールをそれぞれ一つずつ選択し、組み合わせることで分析結果を得ることができる。事後インタビューにおいて、質問 (1) に対しては、全員から、様々なツールの組み合わせを試行したのちに、使用する分析手法を決定し、分析を開始したという回答が得られた。しかし、ユーザ観察の結果、全員が推奨されていないツールを選択し、正しい処理結果が得られなかったことが観察された。また、ツール選択の際、処理の順序が無視されていた。本来であれば、まずマイニング処理ツールを選択してから可視化ツールを選択する必要があるが、可視化ツールを最初に選択する様子が全員に観察された。これについて事後インタビューで参加者 C は、ツールの組み合わせがよく理解できなかったと述べた。また、参加者が使用するツールの組み合わせのペアは、 45 種類ある中で 2、3 種類程度と限られたものであるということが分かった（表 3.2 参照)。

処理の進行状況の提示に関する問題

事後インタビューにおいて、 5 名中 4 名が、自分が選択したツールの処理が正しく行われているかどうか分からなかった回答した。参加者 A は、「処理に時間がかかるものだと、使えないのか使えるのかが分からないので、処理中であることを表すものが欲しい」と述べた。また、ユーザ観察の結果、 5 名中 3 名が、処理結果が表示される前に別の分析へ移行する様子が観察された。

ツール名称に関する問題

事後インタビューにおいて、 5 名中 4 名がマイニング処理ツールと可視化ツールの名称から分析内容や可視化結果を想像できないため、分析を進める上で障害となったと回答した。 TETDM においてユーザは、可視化ツールであれば、「川下り」や「タグクラウド」、「キーワード (展望台)」、「OKmap」、「セグメント独自性」などといった名称から選択する。しかし、テキスト分析に馴染みが薄いユーザにとって、「キーワード (展望台)(図

図 4: TETDM での可視化結果

表 1: 実験参加者が使用したツールの組み合わせ

実験参加者	使用したツール
A	光と影データ+キーワード選択 テキスト分析+ HTML テキスト (Pr)
B	テキスト分析 (Pr) + HTML テキスト (Pr) 光と影データ+キーワード選択 要約（展望台）+キーワード（展望台）
C	タグデータ+タグクラウド 単語抽出+テキストカラー（Sp）
D	関連チェック（Pr）+キーワード選択 テキスト分析+ HTML テキスト 要約（展望台）+テキスト（Pr）
E	タグデータ+タグクラウド 長文+ HTML テキスト（Pr） 要約（展望台）+テキスト（Pr）

4 左参照)」や「タグクラウド (図 4 右参照)」などのツールの名称から可視化結果を想像することは困難であると想像される。

ユーザの情報要求の解決に関する問題

事後インタビューにおいて、5 名中 3 名が、自分の情報要求に対する適切な分析手法が分からず、ツールの選択に躊躇したと回答した。参加者 B は、「たくさんツールはあるけど、使うツールの組み合わせが決まってくるから、予測変換機能のようなものがあれば良いと思う」と述べた。

4 提案インタフェースを用いた観察

本章では、大塚らの提案するインタフェースの有用性を評価することを目的に実施した実験について述べる。大塚らが提案するインタフェースは、オブジェクトを直接操作することで、ツールの選択・切り替えが行うことができる。直感的なモジュールの切り替えにより、ユーザは試行錯誤を妨げられることなく分析を進めていくことができると考えられる。さらに、処理の流れがリンクされるのでユーザは分析内容を把握できると推測される。そこで、本実験では、提案インタフェースを使用し、ツール選択に関して改善点があるか、分析内容把握できているかの 2 つの観点について、ユーザ観察を通して考察する。

4.1 実験の概要

実験参加者は情報学を専攻する大学院生 1 名（男性 1 名）、4 年生大学の情報学部に通う 2 年生 2 名（女性 2 名）、4 年生（女性 2 名）の計 5 名であった。実験の参加条件、実験で対象とした小説、課題は TETDM を用いた実験 (3.1 節参照) と全て同様である（ 3.1 参照）。実験終了後のインタビューでは、提案インタフェースの有用性を測るために (1) ツールの切り替えは円滑に行えたか、 (2) 分析内容を把握できたか、 (3) 分析を進める上で障害と感じた点はあるか、の 3 つの質問を用意した。

4.2 実験の結果

以下ではユーザ観察を行った順に、参加者 A, B, C, D, E と記す。実験参加者がインタフェースを操作する様子を観察したところ、推奨されていないツールの組み合わせを選択する様子は全員に観察されなかった。加えて、 (1) の質問に対して全員が、ツールの組み合わせを迷うことなく、切り替えもスムーズに行えたと回答していることから、ツール組み合わせに関する問題は改善されたと考えられる。しかし、 (2) の質問に対して、5 名中 4 名が分析内容を把握できなかった、自分が行っている分析内容をあまり意識しなかったと回答していることから、分析内容の把握については改善されていないことが示唆される。さらに、 (3) の質問に対して得られた指摘や意見から 4 つの問題が明らかになった。以下に、それぞれについて説明する。

ユーザとシステム間のインタラクションの問題

事後インタビューにおいて、5 名中 4 名が分析の過程で気になった特定のキーワードや調べたい事柄が、元のテキストデータ内のどの位置にあるか知りたかったと述べた。例えば、タグクラウド（図 4 右参照）では、頻出度の高い単語（e.g., 大谷、奥さん）が拡大して表示されるが、それらの単語は原文とリンクされておらず、原文中のどの位置に出現しているか把握することができない。ユーザ観察からも、 5 名中 4 名に可視化結果をクリックする様子が見られた。

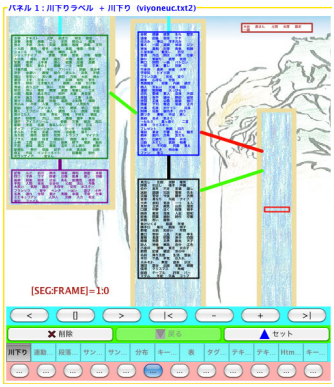

図 5: 川下りの処理結果

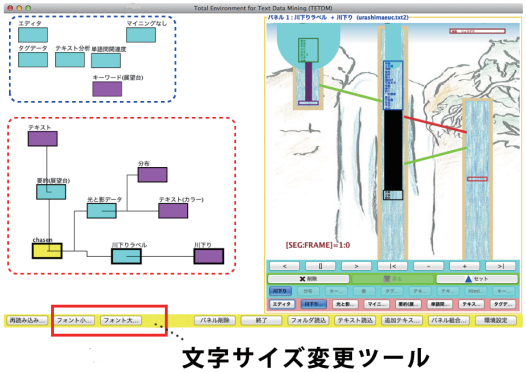

図 6: 文字サイズ変更ツール

ユーザの情報要求の解決に関する問題

事後インタビューにおいて、5 名中 4 名から調べたいことに対してどの手法が適しているのか分からず、様々な分析手法を何度も繰り返したという意見が得られた。参加者 E は「自分の調べたいことに対してどのツールが適しているのか分からない。」と述べている。

ツール間の関係性に関する問題

質問 (2) に対する意見として、 5 名中 4 名が自分の行っている分析内容をあまり意識せず、可視化結果を中心に分析を進めていったと回答した。こととことから、どのような分析手法で可視化を導くかの選択を決定するマイニング処理の部分が、ユーザに意識されなかったことが明らかになった。参加者 C は、 質問 (2) に対して、「どういった過程で可視化結果を導いたかの説明がされていないから、結果を信用することができない」と述べた。

メニューの表示方法に関する問題

この問題は、大塚らがインタフェースを構築する際に従来のインタフェースとは異なるデザインを採用したため、起きた問題である。事後インタビューにおいて、全員が提案システムの結果を表示するウィンドウが小さいため、細かい文字が読めなかったという指摘を行った。 特に一番指摘の多かった可視化ツールが「川下り (図 5 参照)」であり、ユーザ観察からも可視化結果をクリックするなどの様子が全員に見られている。参加者 A は、「クリックすれば画面が文字が大きくなると思い何度もクリックしてしまった」と述べている。フォントサイズは文字サイズ変更ツール (図 6 参照) の位置で変更可能だが、 全員がそのボタンの存在に気付かなかった。

5 デザイン指針

本研究では、 TETDM と提案インタフェースを用いた 2 つのユーザ観察 (3 章および 4 章参照) を通して得られた問題点をインタフェースの問題点と、 TETDM 機能の問題として大別し、それぞれについて解決策の検討を行う。

5.1 TETDM インタフェースの問題

本研究では、 3 章と 4 章で明らかになった問題点を TETDM インタフェースの問題を以下の 5 つにまとめる。

1. 処理の進行状況の提示に関する問題
2. メニューの表示方法に関する問題
3. ユーザの情報要求の解決に関する問題
4. ツールの名称に関する問題
5. ツール間の関係性に関する問題

以上の 5 つの問題点から探索的な分析を行うにあたって TETDM のインタフェースが改善すべき点について検討する。

1 つ目の問題に対する解決策として、プログレスバーを用意することで、ユーザに処理状況のフィードバックを与えることを提案する。ユーザ観察から分かるように、システムから操作に対する反応がない場合、多くの人が繰り返し同じ操作を試みたり、他の操作に移ったりする。そこで、 処理に時間がかかるツールが選択された場合は、その進捗状況を知らせるためのプログレスバーを表示するべきであると考える。

2 つ目の問題に対する解決策として、パネルないしツールバーを用意し、副次的なツールやコンテンツをまとめることを提案する。 ユーザ観察からも分かるように、ユーザは細部ツールへ注意を向けていないこと

が分かる (4.2 節参照)。そこで、文字サイズの変更やチュートリアルの参照といった特定のツールに付随するツールは、パネルを用意してまとめるべきであると考える。

3 つ目の問題に対する解決策として、2 点の案が考えられる。1 点目は、 選択肢のグループ化である。現在のインタフェースにおいて、選択可能なツールを示すリストでは、ツール間の関連性などを考慮せずにツール名が並べられている状態である (図 2 参照)。その中から、ユーザが自身の要求に対して適切なツールを即座に選択することは困難である。この問題を解消する方法として、ツールをグループ化することで、ユーザの選択を支援することが望ましい。2 点目は、ツール推薦機能の付与である。ユーザが行いたい分析に適したツールをインタフェースが推薦することによって、ユーザの試行錯誤は円滑に進められると考えられる。

また、ツール推薦機能を追加することにより、5 つ目の問題のツール間の関係性に関する問題も解決されると想定している。推薦を行う際に、例えば、タグクラウドであれば、単語の頻度により大きさが変わるというように、テキストデータから分析を導き出すまでの過程が説明されれば、テキストマイニングに知識がない初心者でも使いやすいインタフェースになると考えられる。

4 つ目の問題に対する解決策として、分析内容や可視化結果を容易に判断できるようなアイコンをツールリストを示すことを提案する。探索的データ分析では、可視化結果によって洞察を得ながら分析を進めて行くため、ユーザが得られる結果を一目で理解できることが望ましい。

5.2 TETDM 機能の問題

提案インタフェースを用いた実験 (4.2 節参照) でのユーザ観察から、" タグクラウド" (図 4 右参照) や" キーワード (展望台)" (図 4 左参照) などの可視化結果で得られたキーワードから原文にリンクしたいという意見や、特定のキーワードを検索をしたいという意見が得られたことから、TETDM とユーザ間のインタラクションが円滑に行えていないことが予想される。このような問題を解決するために、クエリ機能や検索窓を設けるなどユーザとシステム間の対話を妨げないデザインがなされるべきであると考える。他にも、 従来インタフェースを用いた実験 (3 章参照) での事後インタビューで、「主語抽出ができていないと思う」、「キーワード選択が可能なようだが、それが反映されているように思えない」という意見も挙げられている。このように、TETDM 機能は他にも多くあるため、個々の問題については新たに考察する必要があると考える。

6 おわりに

本稿では、テキストマイニングによる知識発見のためのユーザの探索行為の支援を目的と、ユーザの情報探索行動を観察することで、 TETDM が抱える問題点を整理した。また、先行研究として提案されているインタフェースを用いてユーザ観察を行い、その有用性を評価した。 2 つの実験の結果からユーザの探索行為の円滑化のために TETDM のインタフェースが満たすべき要件について考察を行い、要件を整理した。また、それらを基に TETDM インタフェースの改良指針を提案した。今後は、提案した改良指針に基づいて実装を行い、テキストマイニングを行いたいユーザの探索行為の円滑化について検討していく。

7 謝辞

本研究は科学研究費補助金基盤研究 (C)(課題番号: 22300048) の助成を受けた。記して謝意を表す。

参考文献

[1] Rajiman, M. and Besancon, R.: Text Mining : Natural Language techniques and Text Mining applications, *Proc. 7 th IFIP 2.6 Working Conference on Database Semantics*, pp. 7–10 (1997).

[2] Marchionini,G. : Exploratory Search:From Finding to Understanding, *Communication of the ACM*, Vol. 49. No. 4, pp. 41–46 (2006).

[3] Hearst, M. A.: Untangling text data mining, *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp.3–10(1999).

[4] Otsuka, N. and Matsushita, M.: Graphical Interface that Supports Users' Trial-and-Error Process of Text Mining, *Proc. JSAI2013 International Organized Session: Special Session on Intelligent Data Analysis and Applications*, 1A3-IOS-3a-2 (2013).

[5] 砂山渡, 高間 康史, Bollegala, D., 西原 陽子, 徳永 秀和, 串間 宗夫, 松下 光範: Total Environment for Text Date Mining: テキストデータマイニングのための統合環境, 人工知能学会論文誌, Vol. 26, No. 4, pp. 483-493 (2011).