テキスト分析における試行錯誤の支援に向けて — TETDMのインタフェースに関する一考察 —

Towards Supporting Trial and Error Process on Text Analysis — Redesign of the Interface for TETDM —

大塚 直也[1*]　　松下 光範[1]
Naoya Otsuka[1]　　Mitsunori Matsushita[1]

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

Abstract: 本研究の目的は探索的にテキスト分析を行うユーザの支援である．ユーザの行うテキスト分析は一般に one-shot のプロセスではなく，試行錯誤を繰り返しつつ有益な情報を発見するプロセスである．現在、様々なテキストマイニングモジュールを切り替えながらこのような分析を行う基盤として TETDM が提案されているが，現在の実装では円滑にモジュール切り替えを行えるインタフェースになっておらず，試行錯誤を伴う分析を行う場面で利用するには十分ではない．そこで我々は，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，改良インタフェースのプロトタイプを実装したので報告する．

1 はじめに

現在、膨大なテキストデータを日常的に利用できる環境が整ってきている。それに伴い、それらのテキストデータを閲覧だけに利用するのではなく、分析して有益な情報を発見したり新しい知見を見つけたりするための資源として利用することにも期待が高まっている。テキストマイニングは、このような期待の下で研究が進められている技術であり [1]、構造化されていないテキストデータから新たな知識を取り出すことを目的としている。テキストマイニングでは自然言語処理技術やデータマイニング技術、情報可視化技術などの多様な技術が適応的に利用されており、その結果が重要文の抽出や文書集合のクラスタリング、文書要約などに役立てられている。

一般にテキスト集合の中から新しい知識を発見する場合、人は様々な観点からテキストを処理し、その結果を解釈しながら探索を進めていく [2]。すなわちテキストマイニングによる知識発見は、ゴール志向の one-shot の探索プロセスで達成されるタスクではなく、試行錯誤を繰り返しながら有益な情報を発見する探索的な情報アクセスを必要とするタスクと捉えることができる。

しかし現状では、大量のテキストデータの中から有益な知識を得たいと考えている人が持つ様々な要求に、汎用的かつ十分に対応できる情報アクセス手段は乏しい [3]。また、その分析結果を解釈する際にも利用できる視覚化手法は様々に提案されているが、日常的に使えるレベルのシステムは未だ十分とは言えない [4]。

このような問題を解決するため、汎用的なテキストマイニングの実行を目指した統合環境（Total Environment for Text Data Mining; 以下 TETDM と略す）が提唱されている [5]。TETDM では、世の中に分散しているテキストマイニングツールを画一的に扱える統合環境を構築することによって、テキストデータを扱うユーザの創造的活動を支援するツールの提供を目指している。

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必要とする情報探索に適したものになっているとは言い難い。そこで本研究では、TETDM の目的に沿って、ユーザの試行錯誤を伴う創造的活動をより円滑に支援するための環境に改善することを目指している。このような観点の下、本稿では TETDM のインタフェースが満たすべき要件についての考察を行い、改良インタフェースのプロトタイプを実装したので報告する。

2 TETDM の概要

本節では、これまでに進められている TETDM の概要と、そこで利用可能なマイニングツールの例を示す。

*連絡先： 関西大学 総合情報学部
〒 569-1095 高槻市霊仙寺町 2-1-1
E-mail: k589149@kansai-u.ac.jp

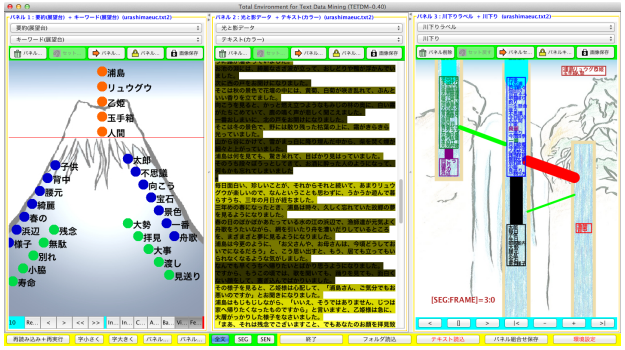

図 1: 複数の可視化結果を表示した時の TETDM の例

2.1 TETDM

これまで、テキストマイニング及びその要素技術について、多くの研究成果が発表されている [3]。しかし、これらの研究により提案されたツールは、論文用の試験的なものが多く実際に世の中で使われている技術はごく一部に限られている。

これらのツールは各研究者によって独自に構築されていることが多いため、個別に導入しなくてはならず、異なるツール間のデータ受け渡しを行うために手作業でデータフォーマットを整えたり、複数のツールの処理結果を比較するために新たに比較用のインタフェースを実装したりするといった手間が必要であった。そのため、テキストデータを用いて情報を多角的に分析したい場合や従来のマイニングツールと新たに開発したマイニングツールの特徴を比較したい場合に、複数のツールを用いることが容易ではなかった。

このような問題を解決するために、TETDM プロジェクトでは世の中に分散しているテキストデータマイニングツールを画一的に扱える環境の構築を目指し、それに向けた統合環境 (図 1 参照) の構築を進めている[1]。

前述したように、TETDM の目的は同一環境上で複数のテキストマイニングツールを扱えるようにすることである。そのため、同一のインタフェース上で複数のツールを組み合わせたり、横に並べて表示することができる仕様になっている。利用者は、自らの興味や目的に応じてツールを選択し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。また、この統合環境上では複数のマイニングツール間のインタラクションを想定しており、あるツールの出力結果を他のツールの入力として用いたり、出力結果に対しての操作が並列に表示されている他のツールの処理結果にも即座に反映されたりするようになっている。

TETDM に入力されたテキストファイルは、共通に行われる前処理 (例えば形態素解析など) の後、各モジュールによって処理されるようになっている。モジュールとは、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、これを TETDM に追加していくことで、取り扱えるテキストマイニング手法を拡張できるようになっている。テキストマイニングの要諦である二つの技術、すなわちマイニング技術と可視化技術は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できる。

現状の TETDM のインタフェースは、横に並べられた複数のツールパネルから構成されている。その一例を図 1 に示す。この図では、3 つの異なるツールが同時に表示されている。それぞれのツールパネルの上部には 2 つのドロップダウンリストがあり、それを通じてマイニングモジュールと可視化モジュールを選択することができる。

利用可能なマイニングモジュールと可視化モジュールの一覧を表 1、2 に示す。現在、既にいくつかのテキストマイニングツールが TETDM にモジュールとして実装されており、利用可能になっている。

2.2 利用可能なマイニングツールの例

TETDM で利用可能な分析手法とモジュールについて例をいくつか示す。

テキストの主題関連度を可視化するシステムとして「ひなたシステム」がある [6]。このシステムは文書中のすべての名詞と各文の関連度を計算し、各文の背景色を関連度が低いものほど黒く、高いものほど黄色く彩色する。主題と関連する箇所や全体に対する位置や割合を確認できるようにすることで、ユーザの文章理解の支援を試みている。TETDM では「光と影データ」(表 1-1 参照) というマイニングモジュールと「分布」(表 2-1 参照)、「テキスト (カラー) 」(表 2-2 参照) の 2 つの可視化モジュールを組み合わせることで、このひなたシステムを実現できる。

また、テキストの主題に基づく一貫性評価を行えるシステムとして「川下りシステム」が考案されている [7]。川下りシステムでは、テキストとその主題を表す単語の集合を入力とし、テキスト中で使われている各単語に主題との関連を表すラベルを与え、それに基づいた単語分布を可視化することによりテキストの主題に基づく一貫性評価を行える。TETDM では、マイニングモジュール「川下りラベル」(表 1-2 参照) と可視化モジュール「川下り」(表 2-3 参照) によって実装されている。また、入力に用いられる主題を表す単語の集合は、統合環境上に実装された展望台システム [8] の出力結果を用いている。

[1]TETDM の最新版は TETDM プロジェクトのホームページ (`http://tetdm.jp/`) 上で公開されている (平成 24 年 10 月 26 日現在の最新版は 0.40 である)。

表 1: 主な利用可能マイニングモジュール

	モジュール名	内容
1	光と影データ	文章中の各文の主題関連度を評価
2	川下りラベル	テキストの一貫性評価
3	要約 (展望台)	特徴的な単語の抽出、重要分抽出
4	主語抽出	文書内の主語を抽出
5	単語チェック	文章作成時のチェック支援
6	長文チェック	文書中の長文の提示
7	テキスト分析	他モジュールの出力結果の集約
8	Twitter	Twitter の検索結果を入力テキストとする
9	単語間関連度	各単語の単語間関連度を出力
10	タグデータ	単語の出現頻度によるタグクラウド用データを出力
11	アノテーション	アノテーションを付与すべき文の抽出

表 2: 主な利用可能可視化モジュール

	モジュール名	内容
1	分布	テキスト中の各文の評価値を横向きの棒グラフで表示
2	テキスト (カラー)	テキストの各文や単語に背景色をつけて表示
3	川下り	川下りラベル用
4	キーワード (展望台)	キーワードを富士山画像の上に配置
5	HTML テキスト	html 形式でテキストを表示
6	キーワードマップ	キーワードを 2 次元平面上に配置
7	タグクラウド	単語の出現頻度によるタグクラウド

このように TETDM では、同一のマイニング処理モジュールからの出力に対し、ユーザの目的に応じて異なった可視化手法を適用できる。既存ないし将来の研究成果によるテキストマイニングツールを同一環境上に集め、画一的に扱える環境を構築することにより、複数ツールを用いた分析を行う際に余計な作業をする必要がなくなり、分析作業に集中できることが期待される。また、多くの研究成果の実用化の際や複数のマイニング手法の比較の際に利用が見込まれる。

3 既存のインタフェースの問題点

前述したように、テキスト分析は試行錯誤を伴う探索的な情報アクセスであるため、ユーザは様々な分析手法の中から適当なひとつの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。そのため、TETDM のインタフェースは、こうした探索プロセスを円滑に行えるものでなければならない。しかし、現在の TETDM のインタフェースは、そのようなプロセスを円滑に行うには十分ではないと考えている。

以下にその問題を 3 点に分けて説明する。

1. **モジュール選択の問題**

   現在の TETDM では、分析に用いるモジュールを選択するためのインタフェースとして、ドロップダウンリストが用いられている (図 2 参照)。これを用いてモジュールの選択や切り替えを行う場合、ユーザが利用可能なモジュールを把握するためには、格納されているリストを展開する必要がある。また、現在の実装では単純なリストが採用されているため、利用可能なモジュール全体を概観することができない。これは、どのような分析が可能かを熟知していないユーザにとっては、作業の円滑な遂行の妨げになる。

2. **モジュールの入出力データの問題**

   TETDM では、複数のテキストマイニング技術をシステム内のモジュールとして扱うことにより、それらを組み合わせて多角的なテキスト分析を行える。しかし、現在利用可能なツールには、

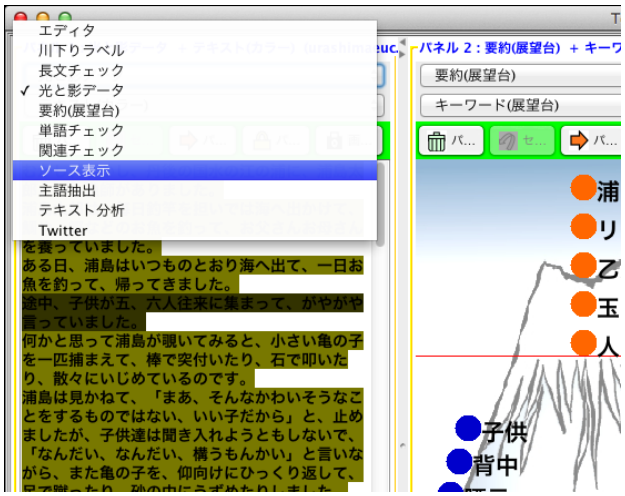

図 2: ドロップダウンリストによるモジュール切り替え

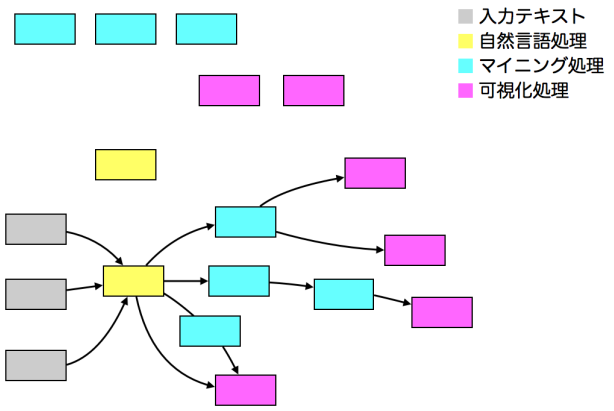

図 3: 提案するインタフェースの概念図

マイニングモジュールと可視化モジュールが一対一の対応となっているものや、一つのモジュールで複数のマイニング処理を行なっているものがある (表 1 参照)。それらのモジュールは、特定のモジュールと組み合わせることのみを前提としており、汎用性が低く、他のモジュールと柔軟に組み合わせることができない。

3. **データフローの問題**

   TETDM に入力されたテキストデータは、形態素解析などの前処理の後、マイニングモジュール及びそれと組み合わされている可視化モジュールによって処理される。それに加えて、他のモジュールの出力を入力としたり、あるモジュールに対する操作が他のモジュールに反映される仕組みが用意されているため、TETDM を用いたテキスト分析では、入力されたテキストデータに対する処理の流れが煩雑となる。しかし、現在の TETDM では、組み合わされて用いられているモジュール以外の処理の流れが明示されておらず、ユーザは自身の分析行為が他の分析結果に及ぼす影響を把握することが困難である。そのため、ユーザの意図しない操作や処理結果が暗黙的に引き起こされる可能性があり、分析の妨げになる。

これらの問題を解決し、ユーザの探索行為がより円滑に支援できるインタフェースのデザイン指針について次章で述べる。

4 デザイン指針

本研究では、3 章で指摘した問題を解決するために、TETDM のインタフェースが満たすべき要件をユーザが (1) モジュール切り替えを円滑に行えること、(2) 各モジュールの利用状況を把握できること、(3) モジュール間の処理の流れを把握できること、の 3 つと考えた。

本節では、このようなインタフェースを設計する際に我々が採用したデザイン指針について述べる。

4.1 ユーザの利用に関する指針

TETDM では複数のツールを用いて多面的に分析を行うことを想定しているため、分析内容が多岐にわたり、そのための作業が煩雑になる可能性がある。そのため、ユーザは自身の分析過程を常に把握し、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なくてはならない。

このような要求に応えるため、本研究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、とするグラフ表現をインタフェースに採用することとした。図 3 にその概念図を示す。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、3 章で述べたようなモジュール選択の問題を解消できると考えている。

4.2 データの制約

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。各プロセスは必ずしも 1 回の処理のみで終わるのではなく、異なる自然言語処理、データマイニングを複数回行うことも想定される。これは、テキストマイニングにおけるデータの流れが複数の処理から構成されることを意味する。提案する

インタフェースではグラフ表現を採用するため、この一連のプロセスをテキストデータの入力から可視化結果までの一本のパスとして捉えることができるようになる。

各モジュールには入力と出力があり、それによっての組み合わせ可能なモジュールの対が決定される。例えば表 1-2 の「川下りラベル」モジュールは plainText を入力とし、ラベルが付与されたテキストを出力する。そして、この出力結果は表 2-3 の「川下り」モジュールでのみ利用される。

現在は、モジュールの粒度がまちまちで、単一のモジュールで複数の処理を行ったり、限定されたモジュールのみとデータのやりとりを行ったりする場合もある。そこで、各モジュールは単一の処理のみを担うものとして整理し、それらの入出力のデータ形式を標準化することで、柔軟にモジュールを選択し、組み合わせて利用できるように粒度を整えることにした。これにより、モジュール間の連携の統一化が図られ、より柔軟にモジュールを組合わせることができるようになるため、3 章で指摘したモジュール間の入出力データの問題やデータフローの問題が改善されるようになると考えている。

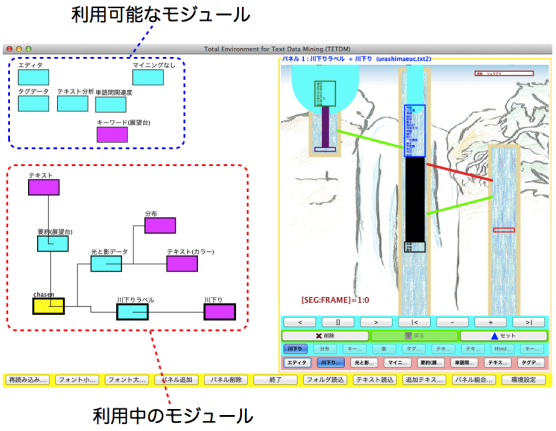

図 4: 実装プロトタイプ

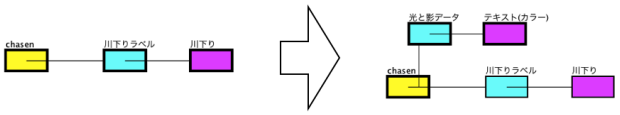

図 5: 「川下り」(左図) に「光と影」を加えた時の様子

5 実装

本節では、実装を行った改良インタフェースのプロトタイプについて述べる。なお、プロトタイプシステムは TETDM version 0.32 をベースとして実装した。図 4 に、現在実装したプロトタイプのスナップショットを示す。この実装プロトタイプは、モジュール切り替えのための改良インタフェース (図 4、画面左) と、既存システムによるツールパネル (図 4、画面右) から構成されている。

改良インタフェースでは、モジュールの利用状況をグラフとして表現し、各モジュールを表すノードを直接操作することによってモジュール切り替えを行える。

既存の TETDM のインタフェースでは、複数のマイニングツールを横に並べて表示できたが、今回の実装では、同時に複数のマイニングツール並べて表示することはできない。

改良インタフェース上に表示されている矩形はその時点で利用可能なモジュールを表しており、それぞれがモジュールの利用状況を表すグラフのノードとなっている。各ノードには、対応付けられているモジュールの名称を付与した。TETDM で定められているモジュールは、マイニングモジュールと可視化モジュールの 2 種類であるが、改良インタフェースでは、TETDM によって入力テキストに対し共通に行われる形態素解析などの前処理を表す仮想モジュールを追加し、合計 3 種類とした。TETDM によって行われる形態素解析には ChaSen[9] が用いられているため、仮想モジュールには名称として「chasen」を付与している。線で結ばれているノード同士は、それらが示すモジュールが組み合わせて利用されていることを表している。提案インタフェース上に多数のモジュールが表示されている場合、モジュールの種類が判別し難くなる。そのため、モジュールが行う処理の段階によってそれぞれ異なる色に彩色した。

各モジュールを表すノードは、マウスでドラッグ & ドロップすることにより操作可能である。ドラッグしているノードをそれと組み合わせ可能なモジュールを表すノードに近づけることで線で結ばれ選択状態になる。選択状態のノードの輪郭線は強調されるため、現在用いているモジュールと処理の流れを把握することができる。図 5 の左図では、マイニングモジュールとして「川下りラベル」、可視化モジュールとして「川下り」が組み合わせて利用されているため、それぞれのモジュールを表すノードが線で結ばれている。図 5 の右図では、それに加えマイニングモジュールとして「光と影データ」、可視化モジュールとして「テキスト (カラー)」が用いられている。図 5 のように、複数のマイニングモジュールと可視化モジュールのペアが用いられている場合、可視化モジュールを表すノードをクリックすることで、選択ツールを変更することができる。

6 議論

本研究で実装したプロトタイプでは、複数の処理結果を同時に表示させることができない。提案手法によって、複数ツールの同時表示を行うためには、可視化モジュールを表すノードと処理結果を表示する画面の対応付けを行わなければならない。その実現手法について今後検討する必要がある。

探索的なデータ分析を計算機システムで効果的に支援するためには、探索行為だけではなく振り返り行為についても考慮する必要がある [10]。提案手法では、モジュールの利用状況をグラフとして表現している。そのため、提案手法を用いて、グラフを履歴として保存したり、グラフに対して、ユーザが分析過程で得た知見をアノテーションとして付与できる仕組みを用意することで、探索過程におけるユーザの振り返り行為を支援できるのではないかと考えている。

7 おわりに

本稿では、テキスト分析における試行錯誤の支援を目的とし、TETDM のインタフェースについて考察を行い、それに基づいた改良インタフェースのプロトタイプを既存の TETDM 上に実装した。テキスト分析は試行錯誤を伴うため、TETDM のインタフェースは試行錯誤を円滑に行えるものでなければならないと考えている。このような観点の下で実装した改良インタフェースでは、モジュールの利用状況をグラフで表現し、各モジュールを表したノードを直接操作することによってモジュールの切り替えができるようにした。

今後、TETDM のモジュールの入出力の形式を検討し、より汎用的かつ柔軟に利用できる形式に改善するとともに、被験者実験を通じて、その有効性を検証していく。

謝辞

本研究の遂行にあたり、文部科学省科学研究費 (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] 渡辺勇: ビジュアルテキストマイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 226–232 (2001).

[2] White, R. W. and Roth, R. A.: *Exploratory Search: Beyond the Query–Response Paradigm*, Morgan & Claypool Publishers (2009).

[3] 市村由美, 長谷川隆明, 渡部勇, 佐藤光弘: テキストマイニング: 事例紹介, 人工知能学会誌, Vol. 16, No. 2, pp. 192–200 (2001).

[4] 増井俊之: 情報視覚化の最近の研究動向, 情報の科学と技術, Vol. 54, No. 11, pp. 558–567 (2004).

[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 Data 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).

[6] 西原陽子, 佐藤圭太, 砂山渡: 光と影を用いたテキストのテーマ関連度の可視化, 人工知能学会論文誌, Vol. 24, No. 6, pp. 479–487 (2009).

[7] 砂山渡, 谷川信弘: テキストの主題に基づく一貫性評価と結論抽出への応用, 知能と情報, Vol. 23, No. 5, pp. 727–738 (2011).

[8] 砂山渡, 谷内田正彦: 文章の特徴を表すキーワードを発見して重要文を抽出する展望台システム, 電子情報通信学会論文誌, D-I, Vol. 84, No. 2, pp. 146–154 (2001).

[9] 松本裕治: 形態素解析システム「茶筌」, 情報処理, Vol. 41, No. 11, pp. 1208–1214 (2000).

[10] 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).