

教育機関における TETDM の活用事例報告

Practical Application in Use of Total Environment for Text Data Mining among Educational Institutions

梶並知記*¹
Tomoki Kajinami高間康史*²
Yasufumi Takama砂山渡*³
Wataru Sunayama*¹神奈川工科大学情報学部
Faculty of Information Technology, Kanagawa Institute of Technology*²首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University*³広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

This paper reports practical application in use of total environment for text data mining (TETDM) among educational institutions. The TETDM has been developed for non-specialist users. It supports text mining and learning of mining methods. It allows users to develop new mining tools. This paper shows the utility of TETDM through practice of text data mining and development of mining tools.

1. はじめに

本稿では、教育機関において、テキストデータマイニングのための統合環境 (TETDM) [砂山 14] を活用した事例を報告する。

TETDM の構造上の特徴は、(1) テキストデータをマイニング処理するツール (マイニング処理ツール) と、処理結果を可視化するツール (可視化ツール) の、2 種類のツールに分離している、(2) 形態素解析といった前処理が自動的に行われる、(3) オープンソースで公開されており、テンプレートとなるソースコードや、既存ツールを拡張する形で新規ツールを実装できる、3 点である。

マイニング処理ツールや可視化ツールは、原則的に、1 つのツールにつき 1 つの役割といった比較的小規模なものとなっている。TETDM の想定ユーザにはテキストデータマイニングの非専門家が含まれているが、1 つのツールがシンプルであることに加え、特徴 (1) により、ユーザは、色々なツールを組み合わせることで様々な角度から文書の分析がし易い。また、ツールを作成する際に、マイニング処理と可視化のどちらか片方のみ作成することができる。特徴 (2) により、TETDM に標準で組み込まれている変数、メソッドを利用することで、語の品詞を特定したり、語を数え上げたりできるため、ツールを自作する際にユーザが前処理部分についてコーディングする必要はない。特徴 (3) により、TETDM はユーザにテキスト分析作業の支援のみを提供するのではなく、新たな分析ツールの、比較的容易な実装環境を提供している。

本稿では、上記の TETDM の特徴と、高等教育機関で行われている教育活動の間に親和性があると考え、TETDM を教育に活用した事例を報告する。報告事例に基づき、TETDM を教育現場において活用する際の、有用性や問題点を議論する。

2. 教育現場における TETDM の活用意義

2.1 教育活動の分類

本稿では、高等教育機関におけるテキストデータマイニングに関連する教育を、目的や内容、形式面に着目し、「技術習得型」と「技術提案型」の 2 つに分類する。

技術習得型では、テキストデータマイニングとは何か、どんなことができるのかについて大ざっぱに学生に理解してもらう段階から、技術要素を理解してもらう段階まで含む。授業の形式としては、座学が基本となるが、個々の技術要素の理解をより確かなものとするため、プログラミングを行い、理論通りの結果が得られるかどうか確認するといった実験や演習も含む。また、試験的なマイニングの課題が与えられ、課題の解決を図るグループワークも含むが、課題解決に用いる手法や、実装するシステムそのものに新規性を求められることはない。

技術提案型では、テキストデータマイニングに関する新しい手法の提案とプロトタイプシステムの実装、その評価を含む。アルゴリズムの提案のみのものもあれば、データ分析作業そのものを支援することを目的としたユーザインタフェースの提案や、既存手法の組み合わせ新規性のあるシステム全体を提案するものも含まれる。主に、専門性の高いゼミナール形式授業、卒業研究以上で行われる研究教育の現場を想定する。

2.2 TETDM の特徴と親和性

2.1 節で述べた、技術習得型と技術提案型の教育活動において、なんらかのソフトウェア (システム) を用いた支援を行う場合、特に要求される点は以下になると考える。

技術習得型に関しては、使い易さと、結果の素早い確認が可能な点が望まれる。対象とする学習者は、テキストデータマイニングの非専門家であるため、手軽に使えて、学習した技術要素に関する出力結果をすぐに確認できることが望ましい。さらに、実際の授業時間が、卒業研究などと比較して少ないこともあり、理論の確認を行う演習・実験などでも、学習者の実装の負担を減らすことが重要である。TETDM では前処理が自動的に行われる点、また、TETDM では比較的小規模のツールを組み合わせる点、テンプレートとなるソースコードや既存ツールの改造によって試験的な実装ができる点が特徴として

連絡先: 梶並知記, 神奈川工科大学, 〒 243-0292 神奈川県厚木市下荻野 1030, 046-291-3235, kajinami@ic.kanagawa-it.ac.jp

あり、技術習得型教育と親和性が高いと考える。

技術提案型に関しては、処理と可視化を分離した実装、既存手法の再利用や既存手法との比較が可能な点が望まれる。アルゴリズムに特化した研究において、インタフェースはそれほど重要ではなく、その逆に、分析作業を支援するインタフェースに関する研究においては、マイニング処理のアルゴリズムの重要度は下がる。TETDM では、処理と可視化の2種類のツールにわかれており、どちらを重視した研究であっても活用できるため、親和性が高いと考える。また、技術習得型に関して述べた部分でも触れたが、前処理のコーディング負荷が少ない点、既存ツールすなわち既存手法を再利用したり、既存ツールの出力するデータを利用した処理の実装が容易であり、TETDM は技術提案型教育とも親和性が高いと考える。

3. 活用事例

3.1 技術習得型の事例

3.1.1 要素技術の学習支援

文献 [梶並 13] では、情報系の3年次生を対象とした、テキストデータマイニングに関する基本的な要素技術を理解してもらうゼミナールにおいてTETDMを利用している。題材の1つとして、文書の類似度に関連する内容を採用している。教授者が文章をベクトルで表現することの意味を図表を用いて学習者に示し、学習者に簡単な文章のcos類似度を手計算で行ってもらい、その後マイニング処理ツールを実装、理論通りの類似度となっているか確認してもらう、演習や実験に類する流れである。学習者は、TETDMであらかじめ用意されているcos類似度の変数を利用し、マイニング処理ツールを作成する。実際には、1からプログラミングするのではなく、ひな形として用意されているマイニング処理ツールを変更する。また、結果の表示には、既存の可視化ツールを利用する。

TETDM の、処理ツールと可視化ツールに分離している構造により可視化ツールの実装を行わなくて済んだこと、前処理が自動で行われること、標準で用意されているメソッドにより類似度計算できることが有効に働き、学習者が、学んだ理論をすぐに試せた（＝自分の理解が正しいか確認できた）例である。

3.1.2 探索的情報検索の支援

[徳永 14] では、環境問題をテーマとしたレポート作成のための情報収集・分析にTETDMを利用している。この演習で、学習者は、TETDMに用意されている、文章の主題に関連する文をハイライトしたり主題を表す観点語を抽出したり、要約を表示したりするツールなどを利用し、情報分析と学習を行い、コンセプトマップを作成する。Web検索は検索サイトを用い、スニペットや本文をTETDMへの入力データとしている。

学習者は、TETDMの環境を用いることで、複数のツールを連動させながらインタラクティブに知識の修得や再分析を行うことができる。特に、観点となる単語の選択と知識の取得を繰り返す際に、観点語に関するWebページの要約文と観点語に関する文（ハイライト表示される）を、インタラクティブに確認できたことが、効率的な知識の修得に役立っていると報告している。また、学習者によって、活用するツールの種類に差がみられたと報告している。

ユーザが自分で処理ツールと可視化ツールの組み合わせを選択し、インタラクティブに分析できるTETDMの特徴が有効に働いた例である。

3.1.3 GUI開発体験

A大学では、テキストマイニングそのものの学習ではなく、JavaによるGUI開発について体験することを目的とした授

業で、TETDMを利用している。Javaの基礎を学習済かつテキストマイニングの知識がない学生を対象にしたゼミナール形式の授業である。この授業では、大学院生が教授者となり、TETDMに標準で付随する基本ツールの動作とソースについて説明、実際にツールを利用してみる他、学習者がソースの一部を書き直してツールの動作がどの様になるかを確認する流れである。授業内で学習者に与えられた課題は、既存ツールを拡張し、独自性を加えたツールを作成することである。学習者によって作成されたツールは、結果的に既存モジュールの軽微な改変となっているが、教授者はこの理由を、学習者のテキストマイニングについての知識不足故に、学習者自身がどのようなツールを開発するか発想するのが困難だったのではないかと推測している。

この事例は、TETDMの、ひな形となるツールが用意されていることや、既存ツールの拡張を許す仕様となっていること、また軽微なツール改変であってもなんらかの結果が出力できるといった点が有効に働いている事例である。

3.1.4 テキスト評価支援

B大学の大学院では、文章の作成に役立てられる、テキストを評価するシステムの開発を目的とした演習に、TETDMを利用している。学習者によって、ツイートを取得するツール、文章の段落ごとの文字数バランスの調整を支援するツールや、文章の語尾や曖昧表現の使用量によって文章のタイプを動物で表すツールなどが作成されている。この授業の特徴は、前年度の成果物を活用している点にある。前年度に作成されたツールの出力データを利用したり、ツールを改造したりして、新たなツールを作成している。具体的には、ツイートを取得するツールの原型は、前年度の学習者によって作成されたものである。

小規模ツールの集合体であるTETDMの構成と、ツール間の連携やツールの改造を許す仕様が有効に働いている例である。

3.1.5 処理ツールに限定した作成体験

文献 [梶並 13] では、工学系3年次生のゼミナールにおいて、特定のツールと連携させることを想定したツール作成タスクを学習者に課した。具体的には、教授者が可視化ツール「レーダーチャート (RaderChart)」を学習者へ提示し、学習者はレーダーチャートで可視化するのにふさわしいと考えるデータを出力するマイニング処理ツールを作成する。

漠然と、マイニング処理ツールを作成せよとの課題ではなく、教授者が利用する可視化ツールを明確にし、学習者にマイニング処理ツール作成の方向性を定めやすくしている。TETDMの、処理ツールと可視化ツールが分離していることと、ツール間の連携ルールが明確であることが有効に働いた例である。

3.1.6 Webからのセンシング支援

A大学では、大学院生を対象にした、テキストマイニングを活用したWebからのセンシングツールの開発を目的とした授業において、TETDMを利用している。この授業は、一般的な情報工学系の大学院生だけでなく、芸術系に所属する大学院生も参加しているゼミナール形式である。

学習者にはTETDMの説明とともに、インフルエンザの流行分析など、Twitterの解析に関する研究の概要を説明した後、開発するツールについて自分たちで検討してもらう。検討の結果、Twitterのリアルタイム性を利用し、指定した単語を含むツイート数をカウントし、イメージピクトグラムを用いて可視化するツールを開発した。開発の際、Twitterの検索・単語抽出を行う処理ツールと、ピクトグラム表示を行う可視化ツールに役割分担をしている。

少人数グループによる共同開発において、処理と可視化ツールに分離しているTETDMの構造が有効に働いた例である。

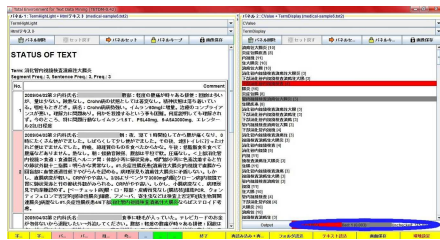


図 1: 専門用語辞書作成支援システム [高間 13]

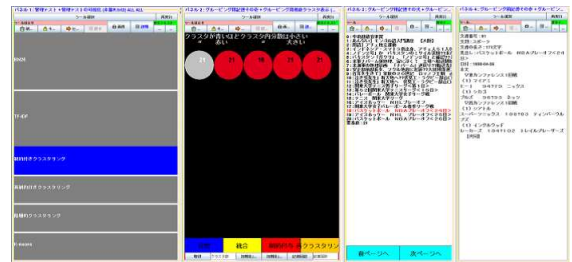


図 2: 制約付きクラスタリングシステム [北村 14]

3.2 技術提案型の事例

3.2.1 専門用語辞書の作成支援

文献 [高間 13] では、看護記録文書を対象にした、専門用語辞書作成支援ツールの開発に、TETDM を利用している。用語の抽出を担当する処理ツール (TermExtraction) と用語をリスト形式で表示する可視化ツール (TermDisplay)、ユーザーの着目した用語が出現している文書中箇所を抽出する処理ツール (TermHighlight) を提案している。図 1 に、実装例を示す。図中の右側にあるパネルに、専門用語がリスト表示され、左側に、用語が出現している箇所を表示されている。

また、TermDisplay との連動に関する処理のみを実装した TermMiningBase ツールを作成し、これを継承することで、異なる用語抽出アルゴリズムを利用することができるようにしている。用語抽出アルゴリズムが異なっても、ユーザーは、同一のインタフェースを利用して分析作業を行うことができる。

処理ツールと可視化ツールで分離している TETDM の構成の他、ユーザーの操作に応じてツール間でデータをやり取りする仕組みが有効に働いた例である。

3.2.2 RDF データの分析支援

文献 [田代 13] では、RDF データベースを対象としたデータ分析ツールを TETDM で実装している。データベース内において、共通の述語を持つ主語の抽出・テーブル作成を司るツール、複数エンドポイント間の共通リソースの抽出を司るツールを提案し、データのリンク関係、すなわちデータ構造の理解を支援する。さらに、時間情報に基づくデータ分析支援のために、時系列データを抽出し、ヒストグラムとして可視化するツールも提案している。ここでは、時間情報が付与されている目的語を抽出し、分析対象としている。目的語に時間情報をもつ述語を選択し、その述語の主語に対して他の述語、目的語を抽出する。抽出した述語と目的語から分析したい組み合わせを選択することにより、その目的語を横軸、時間の階級を縦軸としてヒストグラムを描画する。

この事例では、RDF の解析には Apache Jena を用いており、TETDM 単体ではなく、外部ツールと連携している。

3.2.3 クラスタリング支援

文献 [北村 14] では、インタラクティブに文書クラスタリングを行うシステムの設計・開発に TETDM を利用している。様々な特徴量の計算やクラスタリングのアルゴリズムを、それぞれ処理ツールとして実装している。図 2 は、実装したシステムのスクリーンショットである。一番左のパネルがコントロールパネルであり、ユーザーはここで特徴量ツールとマイニングツールを選ぶ。その右隣から右に向かって順に、クラスタ表示用、クラスタ内の文書一覧表示用、単一文書の詳細表示用パネルである。ユーザーはこれらを用いて文書の詳細を参考にしつつ、クラスタ表示用パネルでグルーピング操作を行い、制約を付与する。その後、制約を考慮して再クラスタリングを行い、

新たな結果を表示する。

この事例では、TETDM のパネルにセットするのは、コントロールパネルのツール、コントロールパネルからの指令を受信するツール、それから可視化ツールとなっており、処理ツールはセットしない。処理ツールは、コントロールパネルから操作することになる。TETDM の基本的な構成である処理ツールと可視化ツールの 1 対 1 対応を崩し、独自のシステム設計パターンに基づき実装している。

3.2.4 文章構造推敲支援

文献 [山手 13] では、TETDM を用いて文章構造推敲支援システムを実装している。ここでは、結論部が先にきて詳細を後から述べるトップダウン構造の文章と、その逆のボトムアップ構造の文章を取り扱っている。それぞれの構造を作成するアルゴリズムと、ユーザーによる構造の変更を支援するユーザインタフェース、即ち提案の核となる部分のみを新規ツールとして実装し、TETDM に標準で付随するツールと組み合わせることで、システム全体を構築している。

TETDM の、既存ツールとの連携のし易さが、有効に働いた活用例である。

4. 議論

4.1 技術習得型における活用

総じて、処理ツールと可視化ツールに分離している TETDM のシステム構成が有効に働いている。前処理に意識を割く必要がなく、なおかつマイニング処理と可視化処理でツールが分かれていることが、時間的制約のある授業で利用するのに適していると考えられる。

3.1.2 節の事例のように、TETDM に標準で用意されているツールを主に使い、探索的情報検索に伴う分析作業を行うタスクにおいては、ユーザーのデータ分析方針に応じて異なるツールの使われ方をされており、授業内であっても、多様なユーザーのデータ分析方針に対して柔軟に対応できていると言える。また、3.1.1 節で述べたように、要素技術の学習支援において、TETDM に用意されているメソッドを用い、既存の可視化ツールで素早く結果をみられることは便利である。3.1.5 節で述べた、処理ツールの作成演習のように、処理と可視化の 2 種類のツールのうち片方を教授者が用意し、もう片方だけ学習者に作成してもらうような授業が可能である。

一方、TETDM に標準で用意されているツールを使うだけでなく、新たなツールを作成する課題を含んだ授業の場合、既存ツールの軽微な改変で終わった 3.1.3 節で述べた GUI 開発体験の事例から、テキストマイニングの知識如何により、成果に差が出ると考える。3.1.6 節で述べた Web センシング支援の事例においても、教授者は、ツール単位で役割分担して開発

を行えるのは TETDM の利点であるが、学生のスキルによって完成度などに差が出てしまったと報告している。

本稿で述べた事例では、TETDM を用いたことによりどれだけ学習者にとって有益であったか、即ち学習者の技術習得をどれだけ助けたかについて、定量評価には至っていない。3.1.3 節の事例で教授者は、学習者側から TETDM の仕様が難しい、わかりにくいといった意見がでたが、この意見は TETDM を用いない授業と比較した上での意見ではないと報告している。しかしながらこの意見は、TETDM の授業への導入に関して、適切なチュートリアルを設けるといった改善の余地があることを示唆していると考ええる。

チュートリアルの整備と関連するが、ツールを作成せず用意されているツールを用いてデータ分析を行うユーザにとっての使いやすさについては、文献 [井須 14] で検討が行われており、ツール選択を支援するインタフェースが提案されている。ここでは、3.1.2 節の事例のような明確な正解が存在しない調査型のデータ分析と異なり、データ分析から得られる解釈が一意に定まるような分析作業の場合に、ツールの選択の仕方によってはなかなか正解へたどり着けない、または誤った結論を導く可能性があることを考慮している。したがって、使いやすさに関しては今後さらなる議論が必要と考える。

4.2 技術提案型における活用

技術習得型の授業環境と異なり、TETDM 特有の仕様に関する理解を深める時間的な余裕や、学習者の平均プログラミングスキル向上、テキストマイニングに関する基礎知識などがあるため、技術習得型で現れる、使いやすさに関する問題点は表出しにくいと考える。

一方、システムの実装に関して、TETDM を用いることである種自由度が制限されることになるため、提案する手法・システムの規模によっては、単純に処理ツールまたは可視化ツールの実装のみでは収まらない場合がある。実際に 3.2.3 節で述べたクラスタリングシステムの例では、TETDM を開発基盤としつつも、その上で独自のシステム設計モデルに基づいた統合的なシステムを稼働させている。このことから、自動的に行われる前処理や、事前に定まっているツール間のデータ通信方法といった、TETDM の利点を生かしつつ、統合システムそのものを拡張することが容易と言える。また、3.2.2 節で述べた RDF データ分析の事例のように、TETDM 以外の既存ツール、例えばメジャーな R や Weka と連携したツールの開発 [徳永 11] も行われている。このことは、TETDM を慣れ親しんだツールにとって代えるのではなく、補完する活用が可能であることを示している。

研究教育において、研究課題の本質的な部分の議論に時間を使い、実装を省力化することは重要と考えるが、TETDM を利用することで、省力化につながるか否か、まだ定量的に評価されていない。教授者側の定性的な評価になるが、技術提案型事例のある教授者は、可視化を含むツールの開発の観点からは例年の (TETDM 未使用) 学習者と比較して、スキルに依存せず開発期間は短い様だと述べている。特に、処理ツールのみを切り替えてアルゴリズムの違いを見たり、あるパネルの有無でインタフェースを変更して比較実験を行ったりは、TETDM の性質に合っていると感じることも述べている。

以上から、技術提案型の教育現場においては、技術習得型で特に有効に働いた利点に加え、TETDM と他のツールの連携と、制限された自由を逆手にとったシステムの拡張が可能な点から、TETDM を有効に活用できると考える。

5. おわりに

本稿では、教育機関において、テキストデータマイニングのための統合環境 (TETDM) を活用した事例を報告した。教育機関で行われている授業を、内容、目的などから技術学習型と技術提案型に分類し、それぞれの型に対して、授業を支援するシステムに望まれる点について論じた。

TETDM は、小規模の処理ツールと可視化ツールの組み合わせから構成され、テキストデータマイニングの非専門家でも、様々な視点からデータ分析可能となっている。さらに、ユーザは前処理のコーディングを特に必要とせず、新たな分析ツールを制作できる。これらの TETDM の特徴は、技術習得型や技術提案型の教育双方と親和性がある。TETDM の有用性を、様々な活用事例に基づいて示した。

しかしながら、今後、特に技術習得型教育において重要度が高いと考えるチュートリアルの充実、学習者の学習効果に関する検証と、技術提案型教育におけるツール同士の比較検討の実践が、課題として挙げられる。

参考文献

- [井須 14] 井須弘恵, 大塚直也, 松下光範: 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, 第 6 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, SIG-AM-06-07, (2014) .
- [梶並 13] 梶並知記: TETDM を利用した情報系専門教育の実践例, 第 27 回人工知能学会全国大会, 3B3-NFC-01a-5, (2013).
- [北村 14] 北村侑也, 高間康史: TETDM を用いたインタラクティブクラスタリングシステムの構築, 第 28 回人工知能学会全国大会, 1H5-NFC-01b-1, (2014).
- [砂山 14] 砂山渡, 高間康史, 西原陽子, 梶並知記, 串間宗夫, 徳永秀和: 統合環境 TETDM を用いたマイニングツールの開発と利用の実践, 人工知能学会論文誌, Vol.29, No.1, pp.100-112, (2014).
- [高間 13] 高間康史, 阿部美里: テキストデータマイニング統合環境を利用した看護記録からの専門用語辞書作成支援ツールの提案, 第 27 回人工知能学会全国大会, 3B3-NFC-01b-1, (2013).
- [田代 13] 田代航一, 高間康史: RDF データベースを対象としたデータ分析支援ツールの提案, 第 5 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, SIG-AM-05-02, (2013).
- [徳永 11] 徳永秀和, 杉村拓哉: R と Weka を活用した TETDM ツールの開発, 第 6 回情報編纂研究会, TETDM-01-SIG-IC-06-07, (2011).
- [徳永 14] 徳永秀和: TETDM による Exploratory Search の評価実験, 第 28 回人工知能学会全国大会, 1H5-NFC-01b-5, (2014).
- [山手 13] 山手砂都美, 砂山渡: トップダウン・ボトムアップな文章構造作成のための推敲支援システム, 第 27 回人工知能学会全国大会, 3B3-NFC-01a-4, (2013).