

テキストデータマイニングのための統合環境TETDMの活用と実践

Practical Application in Use of Total Environment for Text Data Mining

砂山 渡^{*1} 高間 康史^{*2} 西原 陽子^{*3} 徳永 秀和^{*4} 串間 宗夫^{*5} 阿部 秀尚^{*6}
 Wataru Sunayama Yasufumi Takama Yoko Nishihara Hidekazu Tokunaga Muneeo Kushima Hidenao Abe
 梶並 知記^{*7} ダヌシカ ボレガラ^{*8} 佐賀 亮介^{*9}
 Tomoki Kajinami Danushka Bollegala Mitsunori Matsushita

^{*1}広島市立大学大学院情報科学研究科 Graduate School of Information Sciences, Hiroshima City University
^{*2}首都大学東京システムデザイン学部 Faculty of System Design, Tokyo Metropolitan University
^{*3}立命館大学情報理工学部 College of Information Science and Engineering, Ritsumeikan University
^{*4}香川高等専門学校 Kagawa National College of Technology
^{*5}宮崎大学医学部附属病院医療情報部 Medical Informatics, University of Miyazaki Hospital
^{*6}文教大学情報学部 Faculty of Information and Communications, Bunkyo University
^{*7}神奈川工科大学情報学部 Faculty of Information Technology, Kanagawa Institute of Technology
^{*8}東京大学大学院情報理工学系研究科 Graduate School of Information Science and Technology, The University of Tokyo
^{*9}大阪府立大学工学研究科電気情報系専攻 Graduate School of Engineering, Osaka Prefecture University

In this challenge, we develop and distribute an integrated environment to flexibly combine multiple text mining techniques. Text mining techniques include numerous tasks such as salient sentence extraction, keyword extraction, topic extraction, textual coherence evaluation, multi-document summarization, and text clustering. Although tools that individually perform one or more of the above-mentioned tasks exist, it is difficult to integrate and activate multiple tools for a particular task. We attempt to provide the flexibility to integrate numerous tools that exist in the community in our proposed text mining environment. Users can use a customized version of the proposed text mining environment for their specific tasks, thereby concentrating solely on their creative work.

1. はじめに

近未来チャレンジテーマ TETDM (テトディーエム)^{*1}は、複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し、社会的創造活動を支援できる環境としての提供を目指している [砂山 11, 砂山 13]。現在 4 年目に入った TETDM は、統合環境の正式公開を 2015 年 4 月と目標を定めて進める中、2011 年 12 月に試用 α 版、2013 年 5 月に β 版の公開を行った。TETDM は以下を達成目標としている。

1. 幅広い利用者と開発者の参入
2. モジュール間での相互インタラクション
3. 知識創発のための基盤環境の構築

1. では、テキストマイニングの研究を直接活用する人だけではなく、一般のネットを利用する人たちや、プログラミングの学習をしたい人たちなど、利用者と開発者の両面で、初心者、初学者の参入が可能な環境づくりを目指している。2. では、統合環境内のモジュールとして、多くのマイニングツールや可視化ツールが集められたときに、それらが独立に動作す

連絡先: 砂山渡, 広島市立大学大学院情報科学研究科, 731-3194
 広島市安佐南区大塚東 3-4-1, TEL082-830-1705

^{*1} TETDM サイト: <http://tetdm.jp/>

るだけではなく、互いに連携して動作する環境を目指す。3. では、多くのツールを連携させて使う中で、利用者側の発想を促すプロセスの明確化と積極的な思考支援を行っていく。

2. TETDM 統合環境の構成

本章では、TETDM 統合環境の構成 (図 1) について述べる。本環境は、多くの開発者によって作成されるモジュール集合を、協調的に連動させることができる点が新しい。特に、一般的に出力に用いられる可視化インタフェースを入力インタフェースとしても用いることができ、任意のモジュールが、モジュール間で共通の観点に基づいて、協調的な再処理動作を行える点が新しく、これによりインタラクティブな分析作業に没入できる環境を目指している点が、既存のソフトウェアでは実現できていない特徴となっている^{*2}。

2.1 入力: テキスト

本環境には単一のテキストを入力として与える。複数のテキストを入力する場合には、フォルダを指定することで、フォルダ内のテキスト集合が 1 つのテキストに結合させ入力とする。

入力されたテキストは「セグメント」「文」「単語」の 3 つに分割して扱う。「単語」へ区切る際は、形態素解析器を用いて単語に分割する。この際、指定した品詞の単語だけを、キー

^{*2} 既存のソフトウェアとの関係は、[砂山 11] を参照頂きたい。

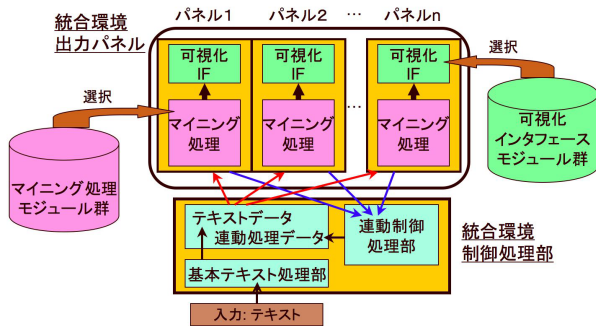


図 1: TETDM 統合環境構成図

表 1: テキストデータ (TextData 型) のメンバー変数の例

メンバー変数	意味
originalText	入力テキスト
segment[].segmentText	セグメントのテキスト
segment[].sentenceNumber	セグメントの文数
sentence[].sentenceText	文のテキスト
sentence[].word[]	文内の単語
sentence[].wordNumber	文内の単語数
keyword[].partOfSpeech	キーワードの品詞
keyword[].frequency	キーワードの頻度

ワードとして取り扱う。「文」に区切る方法は、テキスト中の句点記号(「。」や「.」)をもとに分割する。「セグメント」に区切る方法は、テキスト中の特定の文字列をもとに分割する。

2.2 基本テキスト処理部

基本テキスト処理部では、入力されたテキストにデータマイニングの前処理に相当する基本処理を施して、環境内のデータ構造として扱うテキストデータを生成する。現時点では、形態素解析、単語の出現情報や頻度情報の計算、キーワード、文や段落間の関連度計算を行っている。

2.3 テキストデータと連動処理データ

入力されたテキストに前処理を施した後のデータを保持するためのデータ構造としてテキストデータを用意している。このテキストデータ (TextData 型インスタンスの text) (表 1) が、各モジュールへの入力となる。

合わせて、モジュール間の連動処理に関わるデータ構造を用意する。現時点では、あるモジュールで注目しているデータを共有するためのデータ構造 (Focus 型) を、TextData 型のサブクラスとして実装している。

2.4 連動制御処理部

連動制御処理部では、統合環境上で動作するモジュール間の連動を制御し、また連動を実施する処理を行う。これにより、多様なモジュールの柔軟な組合せを実現し、各モジュールの単体としての使用にとどまることなく、組合せての使用を前提に、スムーズに深い考察が行える環境としての利用を見込んでいる。連動動作には、複数のモジュールが共通のデータに焦点を当てて、再処理や可視化を行う「フォーカス情報による連動」、他のモジュールの処理を実行する「オプションによる連動」、他のモジュールの処理結果を取得する「データ取得による連動」の 3 種類が実装されている。

表 2: マイニング処理モジュールの例

名称	処理内容
1. エディタ	テキストエディタ
2. 長文チェック	文章中の長い文を抽出
3. 光と影データ	各文の主題との関連度を評価
4. 要約 (展望台)	文章要約 (展望台システム)
5. 川下りラベル	文章の主題についての一貫性を評価
6. テキスト分析	テキストの評価結果をまとめる
7. 主語抽出	テキストの各文から主語を抽出
8. 国語辞書	オンライン辞書から単語の意味を取得
9. Twitter	twitter からつぶやきを抽出
10. 単語チェック	特定の単語の出現をチェック

表 3: 可視化インタフェースモジュールの例

名称	可視化内容
1. テキスト	テキストを表示
2. テキスト (カラー)	文の背景色とともにテキストを表示
3. 分布	横向きの棒グラフを表示
4. Html テキスト	html 形式でテキストを表示
5. キーワード (展望台)	展望台システムによるキーワード表示
6. キーワードマップ	キーワード集合を関連度に応じて表示
7. タグクラウド	各文の単語をタグクラウド形式で表示
8. セグメント独自性	独自性の高いセグメントを明示
9. 川下り	文章の一貫性を川の流に喩えて表示
10. レーダーチャート	データをレーダーチャートで表示

2.5 マイニング処理モジュール

マイニング処理モジュールは、統合環境内のテキストデータをもとに、テキストの理解や分析に役立つ情報をテキストから抽出する。またマイニングという言葉にこだわることなく、テキストに何らかの処理を施すモジュールも対象とする。現在までに 20 以上の処理モジュールが作成、公開されている (表 2)。マイニング処理モジュールの処理結果は、可視化インタフェースモジュールに渡されて出力される。

2.6 出力: 可視化インタフェースモジュール

可視化インタフェースモジュールは、マイニング処理モジュールによる出力結果を可視化する^{*3}。可視化インタフェースモジュールでは、入力として受け取れるデータ型 (boolean, int, double, String 型とその次元配列、二次元配列 (String 型以外) の 11 種類) を定め、そのデータの意味を表す整数型変数との組合せにより、マイニング処理モジュールからデータを受け取ることができる。現在までに 20 以上の可視化モジュールが作成、公開されている (表 3)。

2.7 統合環境の動作例

「浦島太郎」のテキストを入力とした処理結果の例を図 2 に示す。左端のパネルの「テキスト分析」モジュールは、右の 6 つのモジュールから「データ取得による連動」により処理結果を得て、html 形式でまとめて表示する。左から 2 つ目のパネルの「キーワード (展望台)」においては、山の上部にある線より上に、ユーザがマウスで配置したキーワードに焦点を当てることができ、「フォーカス情報を用いた連動」により連動に対応しているパネル (左から 1, 3, 5 番目) の出力が更新される。このように、複数のパネルを用いて、利用者の興味に応じた出力結果をインタラクティブに取得できる環境となっている。

^{*3} マイニング処理モジュールの結果によらず、統合環境のテキストデータを可視化するモジュールであっても良い。

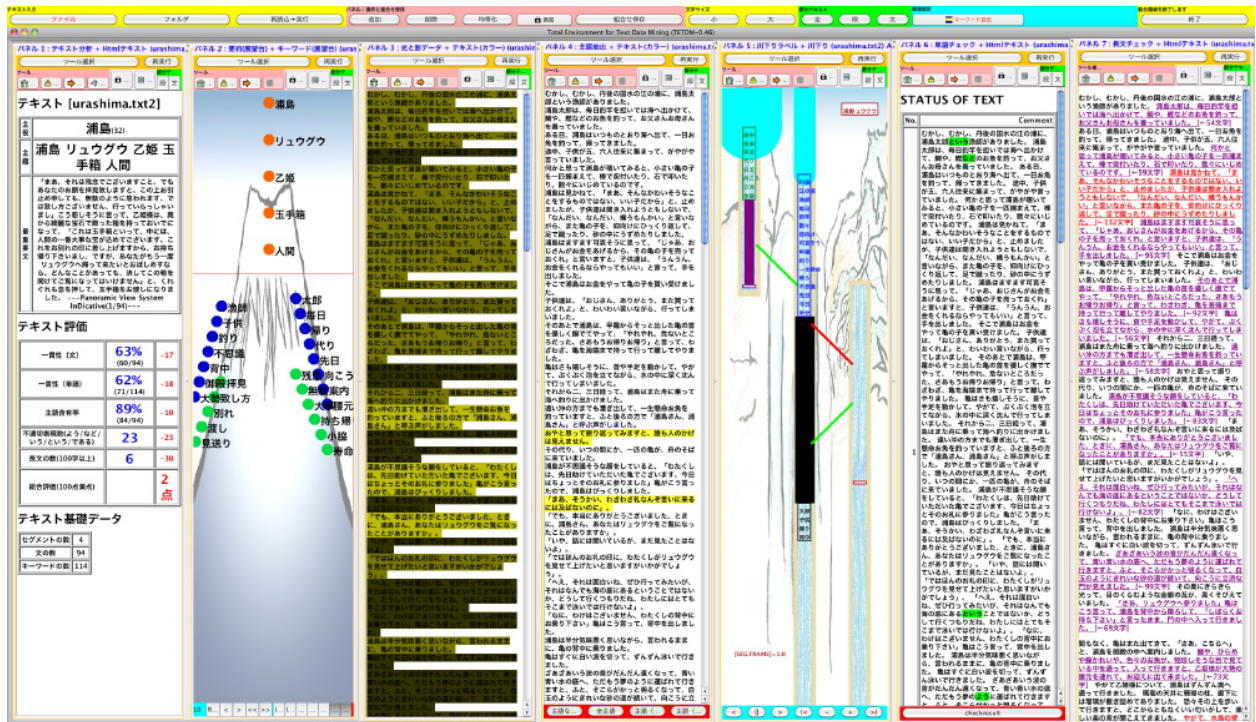


図 2: TETDM 統合環境 (Ver.0.46) の出力画面 (7つのパネルにセットされているマイニング処理と可視化インタフェースモジュールは、左から順に表 2 と表 3 の番号のペアで、(6,4)(4,5)(3,2)(7,2)(5,9)(6,4)(2,4))

3. モジュール開発とテキスト分析の実践

本章では、開発した TETDM 統合環境を用いて現在までに実践した内容を述べる。

3.1 教育機関における学生向け講義、演習での実践

A 大学情報学部 3 年次学生 (4 名) を対象にした、情報検索やテキストマイニングに関するセミナーにおいて、TETDM を活用している。このセミナーは、情報検索やテキストマイニングに関する理論を、実際にプログラムを作成して確認することを目的としている。

題材の 1 つとして、ベクトルを用いた文章の表現を学び、2 つの文章間のコサイン類似度を求めるプログラムを作成してもらった。これにより、2 つの文章からなるテキストファイルを読み込み、作成したモジュールによって求められたコサイン類似度と、手計算で求めたコサイン類似度が一致するか確認することで、テキストマイニングに関する理解を深めてもらうことができた。

また同セミナーでは、「マイニング結果をユーザにどのように提示すべきか」、「ユーザへの提示の仕方に適するマイニングは何か」を学習してもらうための実践を行った。結果として、テキストの主題語⁴が含まれている文やセグメントの割合、主題語を主語とする文の割合など、主題語の出現割合を抽出するものが作成された。

B 大学の大学院博士前期課程の演習で、6 名が 1 チームになり、「Twitter から社会をセンシングするツールを作る」というコンセプトで、実質 3 ヶ月程度でツール (モジュール) の仕様策定と作成を行ってもらった。

結果として、標準語 vs. 方言や、同じ事を表す二つのスラングなど、対となる単語を含むツイートを集めて、どちらが多いかを

*4 既存の処理モジュール、要約 (展望台) の出力結果を利用した。

表す割合を可視化する、というツールの作成に取り組んだ。作成には、ツイートを取得する既存モジュール (TweetExtraction) を利用した。また、TETDM では複数のパネルを並列に表示しての利用が可能のため、(a) キーワードを指定し、それを含むツイートを検索するパネル、(b) 複数の (a) のパネルから検索結果件数を受け取り、それぞれの割合を可視化するパネル、の組合せで動作する環境を作成している。

C 大学の大学院博士前期課程の講義の演習では、この講義を履修している人が個別に、「文章の作成に役立てられる、テキストを評価するツールを作成」する課題に取り組んだ。

結果として、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た。

3.2 教育機関における学生による研究利用の実践

D 高等専門学校では、専攻科生 1 名が特別研究 (大学の卒論に相当) として、探索型検索 (exploratory search) を支援する環境の開発を半年間で行った。得たい情報が不明確で、検索結果より知識を得ながら検索を進めていく探索型検索においては、検索結果のスニペットやリンク先のページを読み、次の検索クエリを探し出すことが行われる。この読むという手間をなくし、ユーザの目的にあったキーワードを推薦するシステムを TETDM のモジュールとして開発した。

C 大学では、特別研究 (卒業研究) として、テキストマイニング初心者のための TETDM におけるツール選択支援を目指したインタフェースの開発を行った^{*5}。TETDM (Ver.0.43) においては、処理モジュールと可視化モジュールがそれぞれ 20 以上あるが、使用目的に対して、即座にツールをセットするこ

*5 TETDM は全体がオープンソースとなっており、独自のインタフェースを開発することも可能となっている。



図 3: 専門用語抽出支援ツール

とが難しい状況となっている。そこで全てのパネルにツールをセットするための環境として、これらツールの一覧を確認でき、可能な処理ツールと可視化ツールの組合せを確認できるインタフェースの開発を行った。

3.3 シーズとニーズのマッチングの実践

本節では、TETDM を用いて、利用者と開発者のニーズとシーズのマッチング [砂山 13] を実践した例について述べる。なお、本節で述べる 2 つのツールのベースは、それぞれ異なる大学の学生が特別研究 (卒業研究) として行った。

3.3.1 看護記録からの専門用語抽出支援ツールの開発

看護記録を対象としたテキストマイニングのために、専門用語辞書を構築する作業を支援する目的でツールの開発を行った。開発した専門用語抽出支援ツールは、TETDM の 2 つのパネルを利用し、マイニング処理モジュールと可視化モジュールのペアとして、各パネルに「用語抽出モジュール + TermDisplay (図 3 右)」と「TermHighLight + Html テキスト*6 (図 3 左)」を割り当てて利用する。

右のパネル内では、用語抽出モジュールによる専門用語の抽出結果を、単語リストとして提示する。単語をダブルクリックすることで、赤 (専門用語として適切)、灰色 (不適切)、白 (未選択) の順に切り替わる。代表的な専門用語抽出手法として、複合語へのスコア付けに基づく 4 種類の手法をそれぞれマイニングモジュールとして実装した。左のパネルは、抽出された用語が看護記録中でどのように出現するかを確認するために用いられ、右のパネル内でクリックした単語が文中でハイライト表示される。

3.3.2 新人とベテランの看護記録比較支援ツールの開発

新人とベテランの看護記録の比較支援を目的としたツールの開発を行った [谷 13]。新人とベテランの看護記録に含まれている単語を比較することにより、それぞれの書き方の特徴をもとに、書き方の違いの分析を支援する。

開発した環境では 5 つのパネルを併用する (図 4)。図中一番左のパネルで、新人とベテランの経験年数をチェックボックスで選んで看護記録を絞り込むことができ、図の中央、および右から 2 番目のパネルでは、それぞれ新人とベテランの看護記録の集合をクラスタリングして地図形式で表示する*7。また

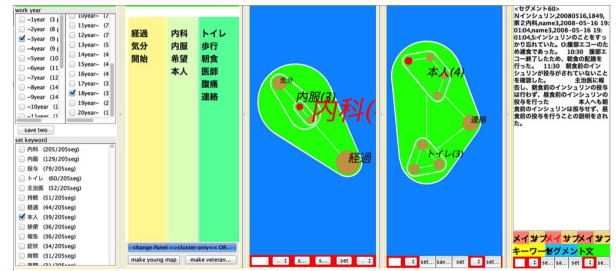


図 4: 新人とベテランの看護記録比較支援環境

一番右のパネルでは、地図中のマウスで選択されたカルテの詳細を表示する。そして、左から 2 番目のパネルで、看護記録に含まれる「新人に特有の単語」「ベテランに特有の単語」「同程度に使われている単語」を表示する。

4. 結論

TETDM では、複数のテキストマイニング、ならびにその周辺技術を柔軟に組み合わせて使える環境を構築し、それらを広く提供することを目指している。本環境により、複数の技術を用いたい利用者の環境が整えられ、ニーズに応じたモジュールを選択し、集中して作業を行えること、また開発者による多くの研究が認知、実用化されていくことを期待している。

今後は、目標の 3. として挙げた「多くのツールを連携させて使う中で、利用者側の発想を促すプロセスの明確化と積極的な思考支援」に重点を置き、知識創発の枠組みを TETDM 内に実装することを目指す。具体的には、処理モジュールと可視化モジュールにおいて、送信/受信するデータの数を制限したプリミティブツールを定義し、プリミティブツール間では任意の組合せと出力が可能になるようにして、組合せの可能性を広げることなどを検討している。

参考文献

- [砂山 11] 砂山渡, 高間康史, Danushka Bollegala, 西原陽子, 徳永秀和, 串間宗夫, 松下光範: Total Environment for Text Data Mining, 人工知能学会論文誌, Vol.26, No.4, pp.483 – 493, (2011).
- [砂山 12] 砂山渡, 濱岡秀平, 奥田澄: 情報収集のためのテキストデータ集合の再帰的クラスタリング, 日本知能情報ファジィ学会誌, Vol.24, No.3, pp.697 – 706, (2012).
- [砂山 13] 砂山渡, 高間康史, 西原陽子, 徳永秀和, 串間宗夫, 阿部秀尚, 梶並知記: テキストデータマイニングのための統合環境 TETDM の開発, 人工知能学会論文誌, Vol.28, No.1, pp.1–12, (2013).
- [湯本 01] 湯本紘彰, 森辰則, 中川裕志: 出現頻度と接続頻度に基づく専門用語抽出, 情報処理学会研究報告 (自然言語処理), Vol.2001, No.86, pp.111–118, (2001).
- [谷 13] 谷恵里香, 砂山渡: 電子カルテにおける新人とベテランの特徴比較支援システム, 第 3 回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資料, pp.37 – 43, (2013).

*6 表 3 の 4 のモジュールをそのまま流用している。

*7 既存システム [砂山 12] をモジュール化して利用している