2E3-NFC-01a-1

テキストデータマイニングのための統合環境TETDM

Total Environment for Text Data Mining: TETDM

砂山 渡[*1] Wataru Sunayama
高間 康史[*2] Yasufumi Takama
西原 陽子[*3] Yoko Nishihara
徳永 秀和[*4] Hidekazu Tokunaga
串間 宗夫[*5] Muneo Kushima
阿部 秀尚[*6] Hidenao Abe
梶並 知記[*7] Tomoki Kajinami
松下 光範[*8] Mitsunori Matsushita
ボッレーガラ ダヌシカ[*9] Danushka Bollegala
佐賀 亮介[*10] Ryosuke Saga
河原 吉伸[*11] Yoshinobu Kawahara
川本 佳代[*1] Kayo Kawamoto

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

[*9]School of Electrical Engineering, Electronics, and Computer Science, Liverpool University

[*10]大阪府立大学工学研究科
Graduate School of Engineering, Osaka Prefecture University

[*11]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka 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

人工知能学会全国大会の近未来チャレンジテーマとして進められている TETDM（Total Environment for Text Data Mining: 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きた [砂山 14]. 近未来チャレンジでは，5 年以内の目標達成を必要条件としており，2010 年度の全国大会で採択された本テーマは，今回の 2015 年度の全国大会で 5 年間の期限を迎えている．本チャレンジでは，次の 3 つを目標に掲げて活動を行ってきた．

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の実現
3. 知識創発のための基盤環境の構築

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/

1 については，日本語を対象として，卑近なデータやモジュールを扱えるようにすることで，テキストマイニングの専門家のみが使えるツールとしてではなく，広く電子テキストを扱う人，ならびに簡単なプログラムの作成が可能な人が，それぞれ利用者，開発者として参入できる環境の枠組みを目指してきた．

2 については，独立に作成された複数のモジュールを並列に並べることができ，ユーザの操作に対して．それらが協調的に動作したり，協調的な表示の切り替えができる環境とすることを目指してきた．

3 については，テキストマイニングの処理やその結果を可視化して提示するところに留まるのではなく，表示される結果の「解釈」，ならびに解釈の結果をまとめて「創発」を実現するところまでの機能を含めた環境とすることを目指してきた．

このそれぞれのチャレンジに対して，それに応えられる環境として TETDM を完成させた．また実社会での活用に向けての取り組みを進めており，一定の成果も得られている．

以下本稿では，2. で構築したテキストマイニングのための統合環境 TETDM について述べる．3. でこの 5 年間の取り組みと成果，4. で今後の展望について述べ，5. で締めくくる．

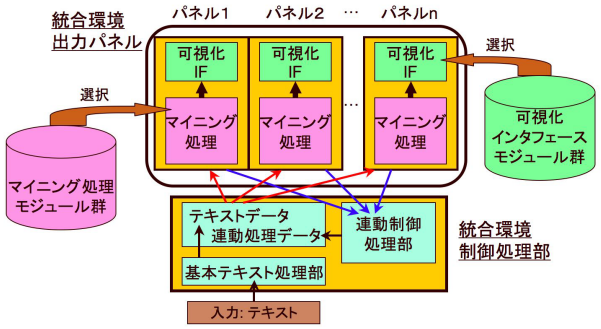

図 1: TETDM 統合環境構成図

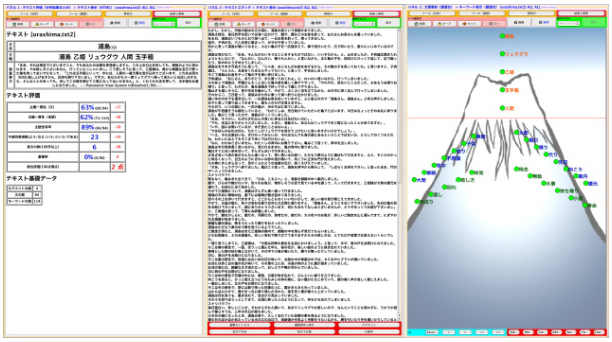

図 2: TETDM 統合環境

2. 統合環境 TETDM の構成

統合環境 TETDM [*2] の構成を図 1 に示す．TETDM は，あるテキストを入力したときに，その分析処理を行う処理モジュールと，処理結果を可視化して出力する可視化モジュールを複数備えている．これらのモジュール群は，単独の開発者によって提供されるものではなく，任意の開発者が作成できる．

TETDM（図 2）は，独立した複数のパネル内に，処理モジュールと可視化モジュールを 1 つずつペアとしてセットすることで動作する．各パネルにセットされる異なる開発者によって作成されたモジュールが，それぞれ独立に動作するだけではなく，それらを協調的に連動させることができる点が TETDM の特徴となっている．以下で各部の説明を述べる[*3]．

2.1 入力：テキスト

TETDM に入力されたテキストは「セグメント（文章または段落）」「文」「単語」の 3 つに分割して扱われる．「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，テキスト中に挿入する特定の文字列，あるいは指定単語をもとに分割する．その後，単語の出現情報や頻度情報の計算，キーワードやセグメント間の関連度計算を行った結果をデータ構造に格納したテキストデータ（TextData 型の変数）を生成し，各モジュールへの入力とする．

*2 統合環境 TETDM は，フリーソフトかつオープンソースの環境として TETDM サイト上で公開されており，誰でも利用できる．
*3 構成の詳細は [砂山 14] を参照していただきたい．

図 3: チュートリアルウインドウ

2.2 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 30 以上の処理モジュールが作成，公開されている．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．

2.3 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[*4]．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 型とその一次元配列，二次元配列 (String 型以外) の 11 種類）を定め[*5]，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 30 以上の可視化モジュールが作成，公開されている．

2.4 チャレンジ内容に関する機能

本チャレンジに関わる TETDM の機能について説明する．

2.4.1 初心者参入のためのチュートリアル

TETDM においては，幅広い利用者と開発者に利用してもらうことを意図し，そのスムーズな利用と開発を促すために，チュートリアルを TETDM 内に実装してきた [川本 15]．利用者や開発者はチュートリアルにおいて，画面下部に並べられた課題を表す宝箱を選択することで，課題に挑戦する．類似する課題のグループが MISSION としてまとめられており，各 MISSION の最後の課題をクリアすることにより，次の MISSION の課題を選択できるようになる．本チュートリアルは，新規の利用者や開発者が，TETDM の利用や開発のイメージをつかむために，利用や開発の具体的方法を実践を通じて理解してもらうことを意図して用意した．

2.4.2 モジュール間の相互インタラクション

モジュール間の相互インタラクションを実現するための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，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，「フォーカス情報による連動」「オプションによる連動」「データ取得による連動」の 3 種類が実装されている．

*4 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．
*5 TETDM は Java で実装されている．

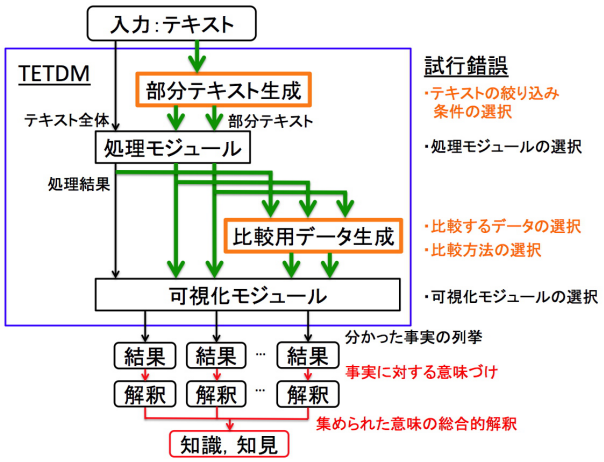

図 4: 知識創発の枠組み

フォーカス情報による連動では，ユーザが，統合環境の出力内で着目（マウスで触れるまたはクリックした）しているデータの情報を，TextData.Focus 型の変数に保存して，モジュール間で共有することで，複数のモジュールが同一のデータに焦点を当てた処理や表示の協調動作を実現する．

オプションによる連動では，各モジュール内において，他のモジュールの処理を，そのモジュールの ID とオプション番号を用いて呼び出して動作させることができる．これを実現するため，可視化インタフェースモジュール，ならびにマイニング処理モジュールにおいて，それぞれの処理を指定のメソッド内に，オプション番号とともに記述する仕様となっている．

データ取得による連動では，マイニング処理モジュール内において，他のモジュールの処理結果を利用したい場合，他の処理モジュールを，モジュール ID とオプション番号を用いて実行させ，その処理結果を取得することができる．

2.4.3 知識創発支援

TETDM における知識創発の枠組みを図 4 に示す．枠組みの中で特に重要なのは，結果を得た後に，それを解釈し創発につなげるプロセスと考え，TETDM 内に次の 3 つのプロセスを支援する機能を実装した．

1. 結果へのアノテーション付け（着眼）
2. 事実への意味付け（解釈）
3. 解釈の一般化（創発）

結果へのアノテーション付け（着眼）では，TETDM における可視化ツールの出力結果上に，何らかの意味があり，解釈を与えるべきと考えられる箇所に「！」の印を付けるアノテーション付けを行う機能を設けた．これにより，思考の材料の収集を支援する．

事実への意味付け（解釈）では，先に与えた着眼点が表す「結果」とその「解釈」をペアで入力できるフォームを用意した．具体的な出力を見ながら入力でき，直感的に結果から解釈につなげられる環境を用意した．

解釈の一般化（創発）では，集められた複数の解釈結果を端的かつ簡潔に表す支援を行うことで，汎用的な知識の創発を支援する．すなわち，入力された「解釈」を分類，整理して複数の解釈を統合した解釈（知識）を生成するための創発支援インタフェースを実装した．創発支援インタフェースにおいては，複数の「解釈」を自由に移動，配置できるようにした上で，関係がある複数の「解釈」を統合した新たな「解釈」の生成を繰り返すことで，最終的に 1 つの「解釈」を得られる環境とした．

表 1: コアメンバー数（各年度全国大会時）の推移

	コアメンバー
2010 年度全国大会	1
2011 年度全国大会	7
2012 年度全国大会	9
2013 年度全国大会	10
2014 年度全国大会	12
2015 年度全国大会	12

表 2: 処理モジュールと可視化モジュール数の推移

時期	Ver.	処理モジュール	可視化モジュール
2011 年 12 月	0.20	7	9
2012 年度全国大会	0.35	20	21
2013 年度全国大会	0.50	28	29
2014 年度全国大会	0.57	30	34
2015 年度全国大会前	0.62	38	35
公開準備中		19	16

3. TETDM の 5 年間の歩みと成果

2015 年 5 月に，TETDM バージョン 1.0 の公開を行った[*6]．TETDM バージョン 1.0 においては，ライトユーザ向けのものも同時に公開しており，幅広いユーザに積極的に活用してもらえる環境とした．知識創発の枠組みまでを含む，テキストマイニングの統合環境を作成する当初の目的は，基本部分で達成できたと考えている．

表 1 に，近未来チャレンジ TETDM の初年度からのコアメンバー数の推移を示す．最初はチャレンジャー 1 名のみからのスタートであったが，本チャレンジの内容に賛同，同調していただける方を募り，年々メンバーが拡大されてきた．

表 2 に，統合環境 TETDM に含められたモジュール数の推移を示す．2011 年に最初のバージョンを公開した際には，わずか 16 のモジュールのみを含む状態であったが，年を経るごとに新たなツールが開発されていき，現時点では準備中のものを含めると 100 以上のモジュールが動作する環境となった[*7]．また TETDM は，バージョン 1.0 公開前の 2015 年 3 月までに，のべ 6000 件以上ダウンロードされている．

3.1 実社会における成果

実社会における成果として，まず大学や大学院における教育現場において実際に活用が行われている [梶並 15]．講義や演習において TETDM を活用し，プログラミングやテキストマイニングの考え方を学生に理解してもらうことに役立てられている．また教員の立場においても，実験や演習のレポート評価や，卒論を含む文章指導支援に活用されている．

*6 本稿作成時 (2015 年 3 月) においては予定．

*7 バージョン 1.0 以降においては，モジュール数が多くなりすぎると，適切にモジュールを選択できなくなる恐れがあるため，基本的なモジュールのみを統合環境とともに配付し，必要に応じてその他のモジュールを追加してもらうことを想定している．

医療の現場においても，電子カルテの分析支援ツールとして活用がなされ [串間 15, 山崎 15]，新人とベテランの看護師の，電子カルテの書き方の違いを分析し，新人に対する電子カルテの書き方の指導などに役立てられている．

3.2 TETDM を利用したアプリケーションの開発

TETDM 上での実装を仮定したアプリケーションがさまざまに開発されている [西原 15, 高間 15, 山手 13, 徳永 13]．入力となるテキストに対して共通となる前処理を実装する必要がないことや，既存ツールとの連携，複数パネルを用いたツール間の協調動作を行う上で，本環境の効果が高いことが伺える．

また現在も，潜在的に TETDM 上での実装が見込まれるアプリケーションの開発が続けられている [後藤 15, 服部 15]．

3.3 TETDM に関連する研究会活動，イベント

TETDM に関連した，全国大会以外での活動についても継続的に行ってきている．人工知能学会の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は，もともと同学会の情報編纂研究会と近未来チャレンジ TETDM が融合して 2012 年度に発足し，その後連携して活動を行ってきており，2015 年 3 月までに 9 回の研究会を実施している．

2013 年 4 月には人工知能学会主催で TETDM の AI ツールセミナーが開催され，2014 年 9 月には広島市立大学主催で公開講座が開催された．2014 年 12 月には，国際会議 SCIS&ISIS2014 において，TETDM のセッションが開催された．

4. TETDM の今後の展望

TETDM の基本形が完成したことにより，今後はその普及と実活用に特に力を入れていきたいと考えている．現在，オープンデータを利用したデータ分析やアプリケーションの開発が注目されている．

利用者の拡大に向けては，一般市民が TETDM をデータ分析に用いることができるように，入手可能なリソースを用いた具体的な利用場面に基づいて，複数のツールを組み合わせて利用する方法を提示するデモやセミナーを実施する．

開発者の拡大に向けても，具体的なモジュールとアプリケーションの開発方法を提示する内容を含めた講習会を，開催していく．例えば，企業が公開しているアンケートデータの分析に TETDM を用いることが考えられ，マイボイスコム株式会社が提供しているインターネットを通じて行ったアンケートの調査結果[*8]等を利用することが考えられる．

TETDM の利用者と開発者の拡大による普及と認知を進める中で，実社会との連携を積極的に行っていく．2015 年度より，広島市との連携事業にも取り組み始めており，その他，自治体や各企業のニーズに応じたアプリケーションの開発や開発支援に取り組んでいく．

5. おわりに

テキストマイニングのための統合環境 TETDM を構築した．本環境は，多くの人々がつながって連携する世の中において，有機的なツールの連携と活用が行えることを目指し，開発を行ってきた．

本チャレンジとしては一つの区切りを迎えたが，世の中のニーズに応えられる，日々活用される環境とするべく，引き続き開発と実践を行っていきたい．

*8 インターネット調査：マイボイスコム http://www.myvoice.co.jp

謝辞

近未来チャレンジ TETDM としての 5 年間を無事に終えることができたことは，本チャレンジに賛同頂いたみなさま，JSAI 近未来チャレンジ担当の方々を初めとする多くの方のサポートによるものと感謝致しております．

参考文献

[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)

[川本 15] 川本佳代，中垣内李菜，西原陽子，砂山渡：TETDM の利用者向けチュートリアルシステムの開発，第 29 回人工知能学会全国大会, 2E5-NFC-01c-1 (2015)

[梶並 15] 梶並知記，高間康史，砂山渡：教育機関における TETDM の活用事例報告，第 29 回人工知能学会全国大会, 2E3-NFC-01a-2 (2015)

[串間 15] 串間宗夫，荒木賢二，鈴木斎王，山崎友義，曽根原登：TETDM を用いた電子カルテ分析支援ツール，第 29 回人工知能学会全国大会, 2E3-NFC-01a-3 (2015)

[山崎 15] 山崎友義，串間宗夫，鈴木斎王，荒木賢二：TETDM を用いた電子カルテテキストデータ分析，第 29 回人工知能学会全国大会, 2E3-NFC-01a-4 (2015)

[西原 15] 西原陽子，梁有烈，柳ヘイ京，福本淳一，山西 良典：議論の進捗を促した発言の抽出と議論の流れの可視化，第 29 回人工知能学会全国大会, 2E4-NFC-01b-1 (2015)

[高間 15] 高間康史，利根川拓馬：TETDM を用いた協調的マルチビューによるインタラクティブ文書クラスタリングの提案，第 29 回人工知能学会全国大会, 2E4-NFC-01b-2 (2015)

[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)

[徳永 13] 徳永秀和：R によるテキストマイニング用 TETDM モジュール開発，第 27 回人工知能学会全国大会, 3B3-NFC-01b-2 (2013)

[後藤 15] 後藤賢悟，砂山渡：AR とテキストマイニングを用いた対話時の好感度推定によるコミュニケーション支援，第 29 回人工知能学会全国大会, 2E4-NFC-01b-4 (2015)

[服部 15] 服部正嗣，藤田早苗，青山一生：テキストから得られる複数特徴量を融合する絵本類似探索法，第 29 回人工知能学会全国大会, 2E4-NFC-01b-5 (2015)