3K2-NFC-3-9

テキストデータマイニングのための統合環境TETDMの公開と活用

Total Environment for Text Data Mining

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
ダヌシカ ボレガラ[*9] Danushka Bollegala

[*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]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

In this challenge, we develop and distribute an integrated environment to flexibly combine multiple text data 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

本チャレンジ 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]．現在 3 年目に入る 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 版の公開を行った．TETDM は以下の 3 つを達成目標として掲げている．

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を含め，より広い範囲で多くの人に活用してもらえる統合環境を目指す．2. では，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，先の 1. と 2. の達成を必要条件として，一般の人が多くのツールを連携させて使う中で，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．

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 ホームページ：http://tetdm.jp

これらの目標が達成されることによって，利用者側の利便だけではなく，研究開発者ならびに社会にとっても，以下の効果を見込むことができる．

- 関連技術の収集が容易になり，関連技術との比較検討や機能拡張が容易になる．
- 試験的なシステムを含めた，多くの技術の実用化や再利用が行われる．
- 研究成果を一つのモジュールとして配付することを，研究のモチベーションにつなげられる．
- 他の作成されたモジュールに刺激を受け，新しいモジュールの研究開発意欲の増加，モジュールの活用案の発想など創造意欲の増加が期待できる．

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．統合環境は「基本テキスト処理部」「テキストデータ／連動処理データ」「連動制御処理部」から構成される制御処理部と，「マイニング処理モジュール」と「可視化インタフェースモジュール」のペアによる出力パネルからなる．

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したい場合には，それらを 1 つのテキストファイルに結合した後に入力できる．

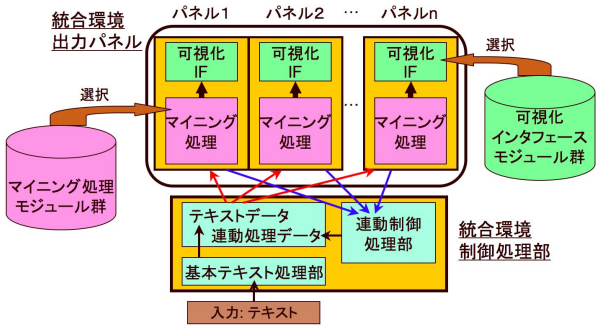

図 1: TETDM 統合環境構成図

表 1: データ構造：テキストデータ（TextData 型）のメンバー変数とメンバー関数の例

メンバー変数／関数名	意味
originalText	入力テキスト
segment[]	セグメント情報
segment[].segmentText	セグメントのテキスト
segment[].sentenceNumber	セグメントの文数
segment[].wordNumber	セグメントの単語数
sentence[]	文情報
sentence[].sentenceText	文のテキスト
sentence[].word[]	文内の単語
sentence[].wordNumber	文内の単語数
keyword[]	キーワード情報
keyword[].partOfSpeech	キーワードの品詞
keyword[].frequency	キーワードの頻度
fileSave()	テキスト保存

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，特定の文字列をテキスト中に挿入し，その文字列をもとに分割する*2.

2.2 基本テキスト処理部

基本テキスト処理部では，入力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，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（表 1）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）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 現時点では，形態素解析器「茶筌」[ChaSen] で未知語と判定される文字列で，かつ設定で指定された文字列（デフォルトでは「スナリパラフト」）によりセグメントに分割している．

表 2: マイニング処理モジュールの例

名称	処理内容 (ペアとなる可視化モジュール番号)
マイニングなし	処理なし (8,11)
エディタ	テキストエディタ (1)
長文チェック	文章中の長い文を抽出 (4)
光と影データ	各文の主題との関連度を評価 (2,3)
要約 (展望台)	文章要約 (展望台システム)(1,5)
関連チェック	2 つのセグメント間の関係を評価 (4)
単語間関連度	キーワードマップのデータ作成 (6)
タグデータ	文のタグクラウド用データ作成 (7)
キーフレーズ	タグクラウド用データ作成 (API を利用)(7)
川下りラベル	文章の主題についての一貫性を評価 (9)
アノテーション	文章中の着目すべき文を抽出 (10)

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れに喩えて表示
10. アノテーション表示	注目すべき文を明示
11. 連動可視化	注目する単語，文，セグメントを表示
12. セグメント木	セグメント間の関係を木構造で表示

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．連動制御には次の 2 つがあり，複数のモジュールの連携を図る．

1. 可視化連動：複数の可視化インタフェース上の表示を同時に変更する
2. 処理連動：複数のマイニング処理を同時に実行する

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．現在までに作成された，モジュールの例を表 2 に示す．

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．現在までに作成された，モジュールの例を表 3 に示す．

3. 統合環境の利用方法

本章では，利用者が TETDM 統合環境を用いる方法について述べる．

*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．

3.1 統合環境のインストール

統合環境の基本テキスト処理の 1 つとしている形態素解析を，現在は外部プログラム「茶筌」[ChaSen] を利用して行っている．そのため，事前に「茶筌」をインストールし，実行プログラム chasen のパス設定を行っておく必要がある．また，JAVA の実行環境がインストールされている必要がある．

3.2 統合環境の操作方法

利用者は図 2 のように任意の数のパネルを横に並列に並べることができ，画面下部のボタンで，各パネルで用いる「マイニング処理モジュール」と「可視化インタフェースモジュール」をそれぞれ選択してセットできる．

3.3 想定する統合環境の利用場面の例

本統合環境を利用する場面として，現在以下の 4 つを想定している．

1. 文章理解支援：Web 上のテキストや論文などの入力に対して，キーワード，要約，単語の出現分布などを提示し，文章の理解を支援する．
2. 文章作成支援：レポート，メール，ブログなど作成した文章を入力して，話のつながり，一貫性，段落間構造や単語の出現頻度などを提示し，文章の作成を支援する．
3. 文章評価支援：レポートや小論文など評価対象となる文章集合を入力して，文章間の類似度，課題との関連度，文章中の意見文などを提示し，文章の評価を支援する．
4. 文章分析支援：アンケートの自由記述文，仕事の日報など傾向を分析することで今後の活動の改善が図れる文章集合を入力して，文章の分類結果，ポジティブあるいはネガティブなキーワード，感情表現などを提示し，文章の分析を支援する．

4. 統合環境のモジュールの作成方法

本章では，統合環境内で動作する「マイニング処理モジュール」と「可視化インタフェースモジュール」を，開発者が作成する方法について述べる．統合環境内のひとつのパネルは，この両者のペアで構成され，両者をともに作成することもできるが，統合環境内の既存のモジュールの利用を前提として，どちらか一方だけを作成することもできる．また，テキスト処理全般のプログラムで必要とされる前処理部分のコードを記述する必要がなく,「マイニング処理」や「結果の描画」など本質的な部分に限ってプログラムを組めば動作する．

4.1 マイニング処理モジュールの作成方法

マイニング処理モジュールは，統合環境が提供する雛形の MiningModule クラス（JPanel 型を継承している）を継承して作成する．図 3 にモジュールの処理内容と処理順序を示す．図 3 左の 3 つのメソッド（関数）をオーバーライド（上書き）すると，統合環境内で動作する．

4.2 可視化インタフェースモジュールの作成方法

可視化インタフェースモジュールは，統合環境が提供する雛形の VisualizationModule クラス（JPanel 型を継承している）を継承して作成する．図 3 右の 3 つのメソッドをオーバーライドすると，統合環境内で動作する．

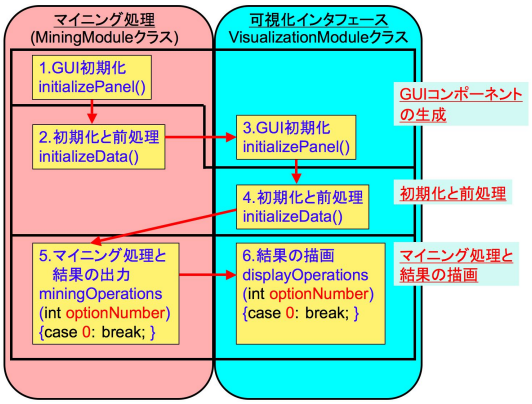

図 3: 各モジュールの処理内容と処理順序

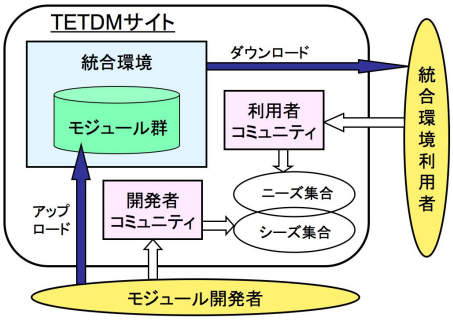

図 4: TETDM サイトの全体構成

4.3 想定するモジュール作成場面の例

モジュールを作成する場面として，現在は主に以下の 3 つを想定している．

1. 一般的な処理や可視化方法の実装を目的とした場面
2. 研究成果の実用化を目的とした場面
3. 大学の講義や演習において，プログラミングやインタフェース設計の学習を目的とした場面

今後，多くの研究者や開発者によってモジュールが作成されることを期待しており，実験評価が行われた信頼性が高いモジュールはもちろんのこと，基本的な処理を行うモジュール，内容は単純だが面白いモジュール，特定の条件でのみ有効なモジュール，面白いが信頼性を保証しないモジュールなど，さまざまなモジュールを積極的に募る．

5. TETDM サイトとコミュニティの形成

本章では TETDM のサイトと，TETDM サイトを用いたコミュニティ形成に向けて，関連する取り組みについてまとめながら，今後の展開について述べる[*4]．図 4 に TETDM サイトの全体構成図を示す．

TETDM サイトにおいては，利用者と開発者が互いに出会える場所を提供することが最も重要と考えている．TETDM が 1 章で掲げた目標を達成するためには，利用者と開発者の

*4 TETDM 統合環境と，技術的に関連する研究は [砂山 11] を参照していただきたい．

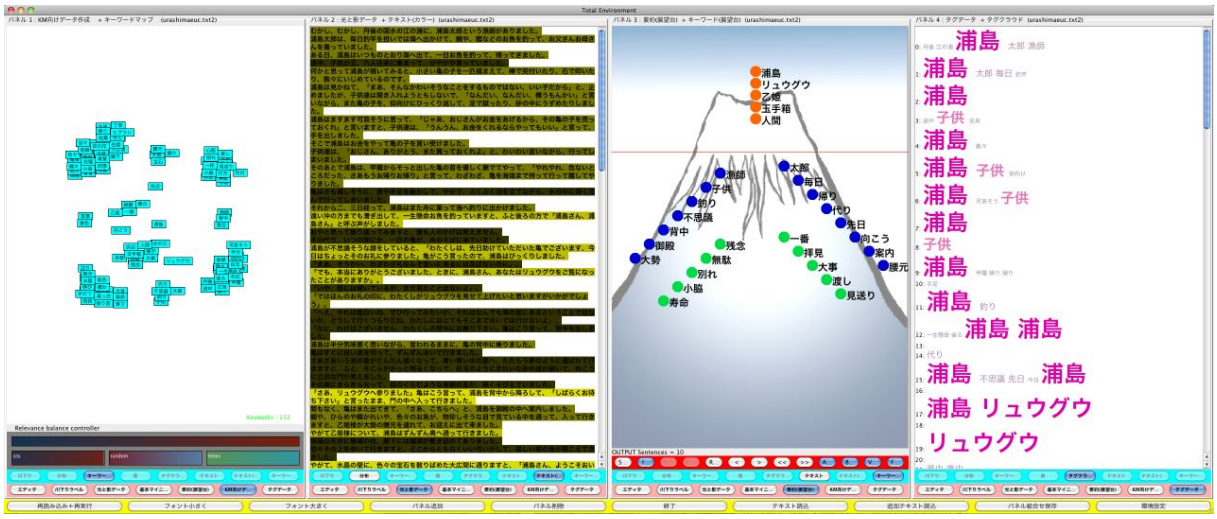

図 2: TETDM 統合環境の出力画面 (左から順に表 3 の 6, 2, 5, 7 番の可視化インタフェース)

マッチングとコミュニケーションが必要で，かつ両者の活動を促進する上で重要になる．これにより，利用者側は必要なモジュールを自ら開発する必要がなく，開発者側は現場のニーズをもとに実世界で役立つモジュールを作成できる．

コミュニティの形成と拡大には，コミュニティへの新規参加を容易にする配慮が重要と考えられる．これには，コミュニティ内の人との簡単なやりとりを通じてコミュニティへの帰属意識を高め，徐々に重要な役割を担うように変化していく正統的周辺参加を可能にすることが重要 [Ye 03] で，マスコラボレーションを成功させる 3 つの要因とも関連がある [タプスコット 07]. 3 つの要因とは (1) 参加が容易で，(2) 個人が少しずつ参加でき，(3) 管理のコストが低いこと，と述べられているが，個人が徐々に参加できる配慮が重要なことは共通している．

コミュニティがうまく運用されれば，新しいモジュールの開発の促進につなげられる．コミュニティ内でのコミュニケーションの活性化のためには，電子掲示板やメーリングリストの整備などに加えて，人間関係の形成が重要な役割を果たすとも言われている [高雄 07]．TETDM の活動においても，コミュニケーションをとるためのメディアの整備に加えて，コミュニティの参加者間の関係の明示 [松尾 05, 井上 04]，話題間の関係の明示 [Kamei 01]，各人に適した情報の推薦 [梅木 99] などによりコミュニティへの帰属意識の向上と，参加者の役割に応じた貢献の頻度を高めていきたいと考えている．

6. 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．

この統合環境が単なるツールの一つとして利用されるだけではなく，さまざまなモジュールの組合せをもとに，多くの人の利用方法や利用欲求に関わる創造力を駆り立て，利用者と開発者の双方が意欲的に活動できる大きなコミュニティの形成につながることを願っている．

参考文献

[ChaSen] 松本裕治，山下達雄，平野義隆，松田寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』，Ver.2.4.0，使用説明書，(2007).

[井上 04] 井上智雄，小林哲郎，池田謙一，重野寛，岡田謙一：ウェブ掲示板を対象としたネットワークコミュニティ分析支援システム：CMINER, 情報処理学会論文誌, Vol.45, No.1, pp.131 – 141 (2004).

[Kamei 01] 亀井剛次，ジェットマー・エバ，藤田邦彦，吉田仙，桑原和宏：ネットワークコミュニティの形成を支援するシステム”Community Organizer”における情報提示手法の検討, 電子情報通信学会論文誌 D-I, Vol.J84-D-I, No.9, pp.1440 – 1449 (2001).

[松尾 05] 松尾豊，友部博教，橋田浩一，中島秀之，石塚満：Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会論文誌, Vol.20, No.1, pp.46 – 56 (2005).

[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).

[高雄 07] 高雄慎二，池内哲之：オープンな開発スタイルを目指してーソフトウェア公開プログラム, NTT 技術ジャーナル, Vol.19, No.1, pp.52 – 55 (2007).

[タプスコット 07] ドン・タプスコット，アンソニー・D・ウィリアムズ：ウィキノミクスマスコラボレーションによる開発・生産の世紀へ, 日経 BP 社 (2007).

[梅木 99] 梅木秀雄：ネットワークコミュニティ形成支援技術（「創造的ネットワーク化情報環境に向けて」），人工知能学会誌, Vol.14, No.6, pp.943 – 950 (1999).

[Ye 03] Ye, Y. and Kishida, K.: Toward an Understanding of the Motivation Open Source Software Developers, Proceedings of the 25th International Conference on Software Engineering, pp.419 – 429 (2003).