1B2-NFC3-10

テキストデータマイニングのための統合環境

Total Environment for Text Data Mining

砂山 渡[*1] Wataru Sunayama　高間 康史[*2] Yasufumi Takama　ダヌシカ ボレガラ[*3] Danushka Bollegala　西原 陽子[*4] Yoko Nishihara　徳永 秀和[*5] Hidekazu Tokunaga
串間 宗夫[*6] Muneo Kushima　松下 光範[*7] 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]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
[*4]東京大学大学院工学系研究科 Graduate School of Engineering, The University of Tokyo
[*5]香川高等専門学校 Kagawa National College of Technology
[*6]宮崎大学医学部附属病院医療情報部 Medical Informatics, University of Miyazaki Hospital
[*7]関西大学総合情報学部 Faculty of Informatics, Kansai 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

本チャレンジ (TETDM，テトディーエム) [*1]では，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，電子テキストを扱う多くのユーザの，創造的活動を支援するツールの提供を目指す．

テキストマイニングと呼ばれる研究には,「重要文抽出」「キーワード抽出」「トピック抽出」「テキストの一貫性評価」「複数文書要約」「テキストクラスタリング」などさまざまな課題があり，すでに多くの研究成果も世の中で発表されてきている．しかし，それぞれの技術を利用するためのシステムやツールは，各研究者が独自に構築することが多く，また論文用の試験的なシステムとなっていたりするため，実際に世の中で使われる技術はごく一部に限られてしまっている．

そこで，既存また将来の研究成果によるテキストマイニング技術を，１つのシステム内のモジュールとして扱うことができ，ユーザの選択したすべてのモジュールを連動して動作させられる環境を構築し，それを無償ツールとして公開することを目指す．これにより，以下の効果が見込まれる．

- 複数の技術を用いたいユーザの環境が整えられ，ニーズに応じたモジュールを選択した上で，分析作業に集中することができる．
- 試験的なものを含む多くのシステムやツールが集められるため，多くの技術の実用化や再利用が見込まれる．

また，テキストマイニング技術を開発する研究者の利点として，新しい技術の開発を促進できる次の効果が見込まれる．

- 関連技術を容易に収集することができ，開発技術と関連技術との比較検討や機能拡張が容易になる．
- 各研究者が研究成果を一つのモジュールとして配付することを意識できるため，研究の高いモチベーションの維持につながる．

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://www.sys.info.hiroshima-cu.ac.jp/people/sunayama/future/newfuture.html

2. TETDM チャレンジの課題と計画

TETDM では，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，作成した環境，および環境内で選択的に使用できるモジュールをダウンロードできる Web サイトを立ち上げることを目指す．以下で，TETDM でチャレンジする課題と，達成に向けての計画について述べる．

2.1 チャレンジ１：幅広い利用者と開発者の参入

ユーザと開発者の，利用と開発のしきいを可能な限り下げて，幅広い利用を見込める環境を構築する．具体的な数値目標として，100 万人のユーザと，1000 以上のモジュールが集められる環境を想定する．

これを達成するための必要条件として，ユーザ側の条件としては，ユーザが平易に使用できることと，また卑近なニーズに答えられることなどが挙げられ，開発者側の条件としては，客観的に効果が確認されていない手法，試験的な手法など完成度に依存しないモジュールが収集できることが挙げられる．

これらを踏まえて，チャレンジ１に向けた課題として以下のものが挙げられる．

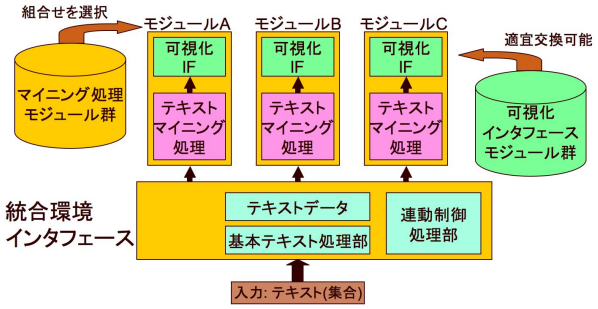

図 1: 統合環境構成図

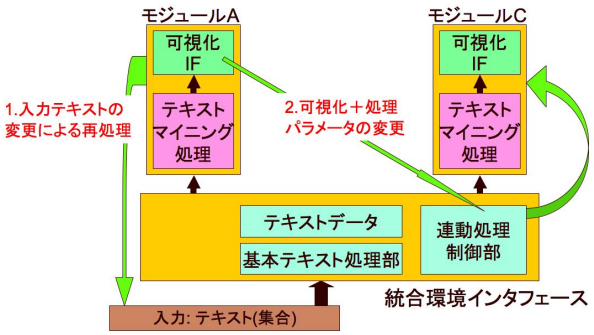

図 2: モジュール間相互インタラクション

a) ユーザの卑近なニーズに応えられること（レポートやメールなど自分の文章や，口コミ，ブログなど他人の文章をチェックできる）

b) ユーザの多様なニーズに応えられること（モジュールの種類が豊富で充実している）

c) ユーザが使用する際の手間が少ないこと（簡易で直感的な操作方法）

d) ユーザの興味を引けること（直感的な面白さを備え，好印象の噂が広がるようにする）

e) モジュール作成のしきいが低いこと（仕様の理解と作成が容易）

f) モジュール作成のための支援環境が整えられること（既存モジュールの拡張や再利用が容易）

g) モジュール公開のしきいが低いこと（バグの不安やメンテの責任，クオリティの低さに対する懸念など精神面での障壁の排除）

h) モジュールの積極的な利用が見込まれること（多くの潜在的なユーザがいること）

2.2 チャレンジ２：モジュール間での相互インタラクションの実現

図 1 の統合環境構成図に示すように，複数のテキストマイニング処理モジュールによる出力結果を，複数の可視化インタフェースモジュール上に同時に表示して，並列に並べて比較できるようにする．

特に可視化インタフェースモジュールは，出力インタフェースになると同時に，自モジュールや，他のモジュールへの入力インタフェースとしても機能させる（図 2）．単純に，入力テキストの変更により，各モジュールで再度処理を行った結果を表示させる機能（図 2 の 1.）に加えて，ある可視化インタフェースモジュールにおいて，「マウスによって選択されているデータが，他のモジュールのどこに出力されているかを，自動的に明示」したり，「他のテキストマイニング処理の実行と結果の表示」を相互に可能にする（図 2 の 2.）．

すなわち，チャレンジ２に向けた課題として以下のものが挙げられる．

a) 他のモジュールで選択されているデータに対応するデータを捉えるために，モジュール間で共通のデータ構造に基づく連動の仕組みを用意すること

b) 他のモジュールを操作するために，他に存在するモジュールの情報を共有するためのデータ構造を用意すること

2.3 チャレンジ３：知識創発のための基盤環境の構築

複数のモジュールを併用して，試行錯誤による結果の分析を行える環境の中で，当たり前の結果だけではなく，頻度が低くても価値の高いデータ，パターンや知識を発見できる環境を構築する．

すなわち，チャレンジ３に向けた課題として以下のものが挙げられる．

a) 精度や信頼度に依存しない多様なテキストマイニングモジュールが集められること

b) さまざまな角度からの分析が行える多様な可視化インタフェースモジュールが集められること

c) 可視化インタフェースモジュールにおいて，直感的にデータ間の関係を捉えられ，分析作業に没入できること

この a) と b) は主にチャレンジ１の，c) は主にチャレンジ２の達成が必要条件となり，それぞれ技術的な側面もさることながら，ユーザや開発者の精神面においても，利用や開発をサポートする環境づくりが望まれる．

2.4 TETDM チャレンジの計画

TETDM チャレンジの 5 年間の計画を以下に示す．

１年目：統合環境の仕様の策定
２年目：モジュールの基本仕様の策定
３年目：モジュール間インタラクションの仕様の策定
　　ダウンロードサイトの立ち上げ
４年目：モジュール開発者支援
５年目：知識創発に向けた利用者支援

以下で，各計画の詳細について述べる．

2.4.1 １年目：統合環境の仕様の策定

図 1 の「統合環境インタフェース」部分の仕様を策定する．すなわち，利用可能なモジュール群の中から，選択的にモジュールを選んで利用できる枠組み，また，入力されたテキストデータを保持するためのデータ構造を定める．データ構造は，多くのモジュールが利用する可能性の高いデータ（テキスト内の単語の出現情報，品詞情報や頻度情報など）についての情報ももつことを想定しており，図 1 の「基本テキスト処理部」において，それらのデータを作成する．これは主に，チャレンジ１の c) に関係する．

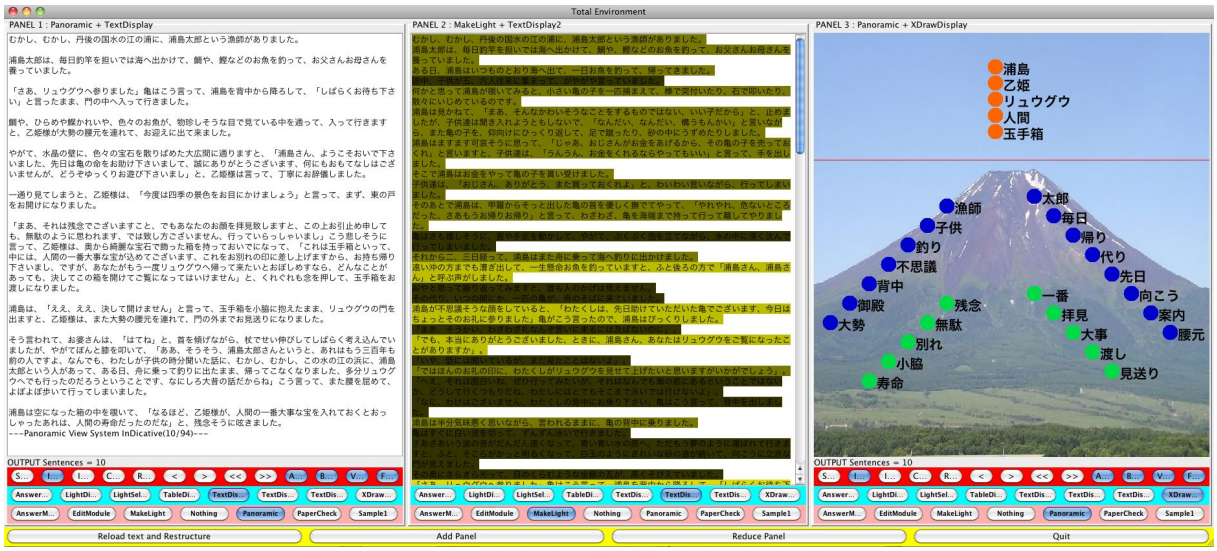

図 3: サンプル環境画面

2.4.2 2年目：モジュールの基本仕様の策定

モジュール間のインタラクション部分を除いた仕様を決定する．各モジュールの入出力に関わる仕様，ならびに複数のモジュールを並列に動作させるにあたって，統合環境とモジュール間でデータの受け渡しを行うためのインタフェース（データ構造や関数）を定義する．この仕様の策定は，チャレンジ 1 の a)，b)，e) と主に関わっており，ユーザのニーズや開発のしやすさに応じた仕様を策定する必要がある．

2.4.3 3年目：モジュール間インタラクションの仕様の策定

統合環境内で利用される各モジュールは，それぞれが独立に動作するだけでなく，あるモジュール内での操作が，他のモジュールにも反映される仕組みを導入する（チャレンジ 2）．そのため，各モジュールが他のモジュールにアクセスするための方法と枠組みを策定する．この仕組みが実現されることは，チャレンジ 1 の d) や，チャレンジ 3 の b) と c) にも関わる．

2.4.4 3年目：ダウンロードサイトの立ち上げ

統合環境と各種モジュールをダウンロードできるサイトを立ち上げる．Web サーバに環境とモジュールのアプリケーションを置き，ダウンロード環境を構築する．また研究者が，各自が作成したモジュールをアップロードできる CGI を実装する．これは，チャレンジ 1 の h) や，チャレンジ 3 の a)，b) に関わる．

2.4.5 4年目：モジュール開発者支援

モジュール開発者支援として，チャレンジ 1 の f)，g)，ならびにチャレンジ 3 の a)，b) に関わる支援の枠組みを検討する．すなわち，モジュールを容易に作成できるように，サンプルモジュールを充実させたり，既存のモジュールの再利用を促す環境を構築する．また，モジュールの公開に当たって，開発者がその内容についてのクレームや，保守の責任などの懸念が発生しないような枠組み，(例えば匿名の開発者を許すことなど) を検討し，積極的なモジュールの公開を促す．

2.4.6 5年目：知識創発に向けた利用者支援

知識創発に向けた利用者支援として，チャレンジ 1 の a)，b)，d)，ならびに，チャレンジ 3 の c) に関わる支援を行う．すなわち，情報推薦の既存技術を応用するなどにより，ユーザの多様なニーズを満たすモジュールをスムーズに選択できる枠組みの検討，ならびに，効果的な使用例と使用結果などのサンプルを積極的に公開する．

3. テキストデータ分析のための統合環境

図 3 に現在策定中のサンプル環境[*2]を示す．環境はすべて Java 言語で記述されている．ユーザは任意の数のパネルを横に並列に並べることができる．画面下部のボタンで，各パネルで用いる「テキストマイニングモジュール」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すると，それぞれの処理結果が表示される．必要に応じて，各テキストマイニング処理を行うボタンを押すとその処理結果により，表示が更新される．

現在はこの環境並びにモジュールの仕様を策定しており，仕様の策定とモジュールの試作の繰り返し，汎用性並びに使い勝手の面で，多くに人に利用してもらえるよう調整を行っている．現試作段階では，モジュールの追加と削除は，各モジュールのファイルを指定するフォルダに追加，削除することで行うことができる．

モジュールは，テキストマイニングに関わるモジュールでなくても追加が可能であり，情報アクセスに関わるモジュールとの連携により，インターネットにアクセスして得られたテキスト集合をそのまま環境の入力テキストデータとして用いる連携や，学習支援モジュール等との連携により，学習結果のテキストデータを即座に分析して，得意／不得意分野の傾向を知るなど，幅広い分野間での連携も視野に入れている．

特に多くのユーザの利用が見込まれる卑近な使用例としては，メールの作成支援として，自分の作成したメールの誤字や脱字，敬語のチェック，読んだ人が受ける感情の推定結果などを表示することや，Web ページ検索の結果のテキスト集合や，興味あるブログやつぶやき [twitter] のテキスト集合から，キーワードや関連語情報，分類結果と各分類の要約を提供することなどによる，情報へのアクセス支援と，新たな興味への発想の手がかりを提供することが考えられる．

*2 環境の試用や，モジュールの作成を試されたい方は，TETDM ホームページ (P.1 脚注 1 に記載) から，ダウンロードおよびモジュールの作成方法を参照してください．

4. TETDM チャレンジの位置づけ

本章では，関連する技術との比較から TETDM の位置づけを明確にする．

4.1 商用テキストマイニングソフトウェア

DIAMining，Text Mining Studio，TRUE TELLER，Text Mining for Clementine など [DIAMining, Mining Studio, TELLER, Clementine]．はいずれも商用のソフトウェアとして，確立された技術をもとにした信頼度の高いマイニング結果を出力する．しかし信頼度が高い結果は，既知の一般的知識を多く含み，発見的な知識を得ることは難しい．また有償のソフトウェアであるため，誰もが気軽に用いることはできない．TETDM では，必ずしも信頼度が高くないが何らかの特徴があり，応用の可能性を秘めたデータを，多様なモジュールの組合せとデータとのインタラクションを通じて発見できる環境の構築を目指す．

4.2 テキストマイニングの統合環境

テキストマイニングのための統合的ツール GATE [GATE] は，研究の実用化のためのベンチマークテストデータの設定など，研究の再利用を意識し，専属の開発チームによるリリースが行われているが，精度が重視されるモジュール構成で，多様なモジュールを集められる枠組みにはなっていない．

ユーザとシステム間のインタラクションを想定した統合環境に LanguageWare [Language] がある．入力テキストの言語の推定，単語の抽出，品詞の推定，単語の正規化，固有表現抽出などのテキスト分析を行うことができる．また，UIMA（Unstructured Information Management Architecture）[Ferrucci 04] と呼ばれる基準に基づいてコンポーネントを作成しているため，この基準に基づく他のコンポーネントとの組合せも可能となっている．LanguageWare は，主にビジネスの現場にいる人をユーザとして想定しており，利用にはテキストマイニングの知識や経験が求められる．本チャレンジでは，学生や主婦など，PC は利用するがテキストマイニングという言葉を知らないユーザも想定しており，単純かつ直感的に用いられ，利用価値がある環境の構築を目指す．

種々の言語のデータを，さまざまなモジュールで扱えるようにするためのミドルウェアアーキテクチャとして，Heart of Gold [Heart] がある．コーパスへの自動かつ多次元のアノテーション，XML ベースでのモジュール同士の結合などを行える．これは主に，言語データの中から共通パターンを見つけ出すことなどを念頭においているが，本チャレンジでは希少価値のあるパターンを発見できる環境づくりを目指す．

U-Compare[狩野 08] は処理の実行順序がプログラマブルで，相互依存するコンポーネント（モジュール）を入出力定義により自動的に組み合わせられる．また，UIMA にも準拠している．U-Compare においても複数コンポーネントの結果を視覚化して同時に並列表示するが，本研究では，複数の可視化モジュール間でのインタラクションを可能にし，ユーザの集中的な試行錯誤を促せる環境の構築を目指す．

既存のテキストマイニングシステムにおいては，信頼度の高いモジュールが多く，ヒューリスティックな手法（発見的手法）や思いつきによるアドホックな手法に基づくモジュールが比較的少ない．客観的な評価を得たシステムをモジュールとして収集することはもちろん望ましいが，モジュール作成のしきいも大きく上がるため，モジュール収集の範囲が狭まったり，新技術の速報性にも欠ける可能性がある．本チャレンジでは，信頼性のしきい値を下げることによって，多様な技術のモジュールを集められる枠組みを構築し，ユーザの思考を活性化し，新たな知識の発見を促せる環境づくりを目指す．

5. 結論

TETDM チャレンジでは，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ユーザの環境が整えられ，ニーズに応じたモジュールを選択し，集中的して作業を行えること，またテキストマイニング技術の開発に際して，他の技術との連携を意識しつつ本環境に統合することで，多くの研究が認知，実用化されること，が期待できる．

TETDM チャレンジが目指す環境は，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，開発者と利用者の双方が意欲的に活動できる大きなコミュニティが形成されることを期待している．

参考文献

[Clementine] Text Mining for Clementine（http://www.spss.co.jp/software/modeler_ta/）

[DIAMining] DIAMining（http://www.mdis.co.jp/products/diamining/）

[言語グリッド] 言語グリッド（http://langrid.nict.go.jp/jp/）

[GATE] GATE（http://gate.ac.uk/）

[Heart] Heart of Gold（http://heartofgold.dfki.de/）

[狩野 08] 狩野芳伸，辻井 潤一： UIMA を基盤とする相互運用性の向上と自動組み合わせ比較－国際共同プロジェクト U-Compare，情報処理学会自然言語処理研究会報告，Vol.2008, No.67, pp. 37 – 42, (2008).

[Kimani 03] S. Kimani, S. Lodi, T. Catarci, G. Santucci and C. Sartori: VidaMine:A Visual Data Mining Environment, Journal of Visual Languages and Computing, Vol.15, No.1, pp.37 – 67, (2004).

[Language] LanguageWare（http://www.ibm.com/software/jstart/languageware）

[Mining Studio] Text Mining Studio（http://www.msi.co.jp/tmstudio/）

[orange] orange (http://www.ailab.si/orange/)

[R-Project] R-Project（http://www.r-project.org/）

[TELLER] TRUE TELLER（http://www.trueteller.net/）

[twitter] twitter（http://twitter.com/）

[Ferrucci 04] Ferrucci, D. and Lally, A. : UIMA: an architectural approach to unstructured information processing in the corporate research environment, Natural Language Engineering, Vol.10, No.3-4, pp.327 – 348, (2004)

[Weka] Weka（http://www.cs.waikato.ac.nz/ml/weka/）