2E5-NFC-01c-2

テキストマイニング共通語彙基盤の構築とツール実装への適用に関する検討

Developing an Integrated Text Mining Environment based on Conceptualization of Objects and Methods

阿部 秀尚 *1
Hidenao Abe

*1文教大学情報学部
Faculty of Information and Communications, Bunkyo University

In this study, a conceptualization of text mining objects and methods is constructed for multi-grain-sized text mining operations, which are observed in the TETDM project and other text mining practices. Since it is hard to learn text mining processes for novice users, the conceptualization of the text mining methods help to understand the processes, which carry out one or more text mining methods for a required result by the users. By employing both of the conceptualized hierarchies of the objects and the methods as the software repositories, I discuss a tool for combining the multi-grain-sized text mining methods for supporting the uses' text mining processes.

1. はじめに

インターネットをはじめとする種々の情報システムの利用機会の増大により，人が活動することによる電子データの蓄積が日々加速している．そこで，テキストに対する自然言語処理を適用し，取り出される構成要素や構成要素間の統計的な性質を利用して有用な情報を得る，テキストマイニングが電子的なテキストの収集可能化と共に広く求められるようになってきた．また，ビッグデータと呼ばれる現在のデータ集積の加速は，機械に取り付けられたセンサからのデータの蓄積の促進にも注目が集まっているが，その中心は人間がそのまま理解可能なテキストによるものである．

しかしながら，テキストマイニングは，特定の技術適用を指す言葉ではなく，多くの技術的課題解決方法の総称であり，そこで行われる適用 *1 は様々な処理操作を組み合わせて行われることが一般的である．このため，どのようなタスクが「テキストマイニング」にあたるのか，初学者が学習し，理解するまでには多くの経験と学習機会が必要であった．

本研究では，テキストマイニングにおけるより多くの処理操作 *2 を集積し，利用可能とする環境である TETDM[砂山 13] を基に，テキストマイニングにおけるメソッドとメソッドの操作対象であるオブジェクトの整理を行ってきた [阿部 14]．本稿では，テキストマイニングメソッドにおける入出力と参照にあたるオブジェクトをさらに整理し，オブジェクト階層として示す．さらに，多粒度のテキストマイニング関連メソッド群をメソッド階層として整理し，公開・利用するための基盤として Web サービスに基づく実装について，検討する．

2. テキストマイニングプロセスの類型

書籍などに示されるテキストマイニングプロセスは，図 1(a) に示すように，入力テキストを加工（前処理）し，規則性などを生成するマイニング処理を実行し，結果の可視化などにより評価を行う一連の工程として示される．しかし，実際のテキストマイニングの実行では，図 1(b) で示すように，それぞれの段階で試行錯誤が行われ，入力テキストから固有名を抽出するためのユーザ辞書の構築や，特徴語の選定などの洗練化が行われる．これらの洗練化の過程で用いられる改善策は多種多様であり，入力テキストと要求された結果に応じたマイニング処理の単純な選択と比して，多大な労力と専門知識が必要となる．あるいは，逐一単純なマイニング処理を組み合わせ，処理結果を確かめながら，テキストマイニング処理を進めるときにも，図中 (b) で示した繰り返し型のテキストマイニングプロセスとなる．

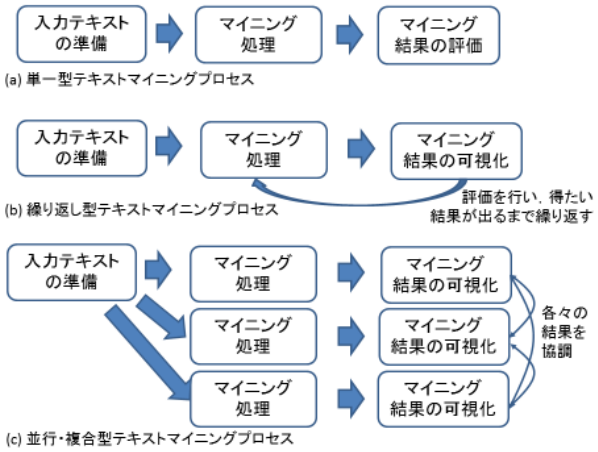


図 1: テキストマイニングの典型的なプロセス．

さらに，TETDM プロジェクトでは，初学者が簡単な処理からマイニングまで，複数の処理を並行して結果を見ながら実行できるよう TETDM と呼ぶ統合環境を提供している．TETDM では，それぞれの処理結果（可視化ツールと呼ばれる）における操作を連動させる機能が提供されるため，単に複数のマイニング処理結果を並べるだけではなく，互いに強調した処理を実現している [利根川 15]．TETDM で実現される並行・複合型のテキストマイニングプロセスを図 1(c) に示す．

連絡先: 阿部秀尚，文教大学情報学部情報システム学科，〒253-8550 神奈川県茅ヶ崎市行谷 1100，0467-53-2111，hidenao@shonan.bunkyo.ac.jp

*1 本研究ではタスクと呼ぶ．
*2 本研究ではメソッドと呼ぶ．

3. テキストマイニングオブジェクトの概念化

本節では，テキストマイニングオブジェクトの同定と概念化を示す．先行研究 [阿部 14] では，テキストマイニングに関連した書籍に記述された複数のテキストマイニング事例およびテキストマイニングツールの入力，テキストデータを扱うマイニング手法の実装を基にテキストマイニングオブジェクトの同定を行った．これにより，各オブジェクトの間に is-a 関係を定義し，テキストマイニングオブジェクトの概念階層を構築した（図 2）．

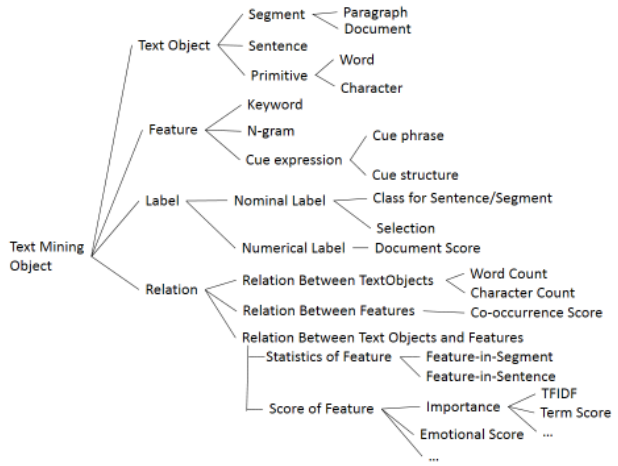


図 2: テキストオブジェクトの概念階層 [阿部 14].

以上のテキストオブジェクトは，テキストマイニングにおける個々の処理の入出力，および参照の概念を表しており，実装時には基本データ型などへの変換が必要となる．そのため，実装では，各オブジェクトに対応したクラス定義などとともに，setter/getter として各データ型への入出力メソッドを必要とする．

さらに，可視化ツールの出力は，評価対象となるオブジェクトを出力していると考える．現在，TETDM で利用可能な可視化ツールの出力から，図 3 に示すように出力オブジェクトの類型に基づく体系化を行った．

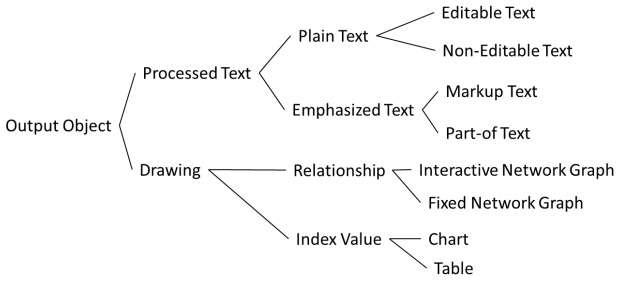
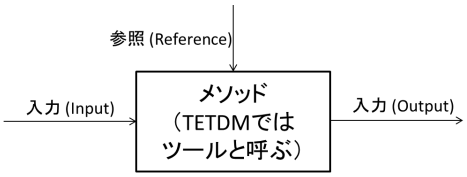


図 3: TETDM の可視化ツール類型に基づく出力オブジェクト群.

4. テキストマイニングメソッドの概念化

以上のテキストマイニング関連オブジェクトは，テキストマイニングメソッドを図 4 のようにとらえることによるものである．それぞれのメソッドに入出力と参照に関するオブジェクトをプロパティ値として与え，実装された各ツールへの実装と対応するよう段階的に詳細化し，概念化を定義する．なお，可視化ツールの出力は，利用者に提示される可視化内容の類型を表すため，図 2 および図 3 において定義したオブジェクト階層を用いる．



処理ツール
Module ID: 整数型の値
Implemented-As: 文字列型の値
Input： Text Mining Objectで定義
Reference: Text Mining Objectで定義
Output: Text Mining Objectで定義
Pre-method: 前提となるツールのModule IDの値*
Post-method: 後続のツールのModule IDの値*

可視化ツール
Module ID: 整数型の値
Implemented-As: 文字列型の値
Input: Text Mining Objectで定義
Reference: Text Mining Objectで定義
Output: Output Objectで定義
Pre-method: 前提となるツールのModule IDの値*

*:接続条件が無い場合は空値

図 4: テキストマイニングメソッドの定義.

各ツールにあたるテキストマイニング関連メソッドについて，現在の TETDM では系統分けなどの区別が行われていないが，明らかにメソッド間で処理対象の粒度が異なるツールが登録可能な仕組みとなっている．例えば，阿部が [阿部 12] において示した複数の単語重要度の計量化指標を算出するツールに対し，個別に計量化指標を算出するツールの作成 [砂山 14] も可能である．このため，テキストマイニングでは，マイニング処理の目的に応じて，粒度が異なるメソッドが考案可能であると言える．そこで，メソッドに関しては，入出力・参照において利用するオブジェクトの種類，および数の差異などの観点から，メソッド概念の階層化による体系化を行っていく必要がある．

5. セマンティック Web サービス上での実装についての検討

現在，TETDM は Java で実装された統合環境処理部を中心に構成されている．また，マイニング処理および可視化のためのツール実装も，統合環境の実装言語である Java で行うこととなっている．しかしながら，これまで述べてきた共通語彙にあたるテキストマイニング関連オブジェクトを利用し，公開を行っていく上では，単一のプログラミング言語による実装を行うことは制約が強い．

データマイニングツールについても，同様に強すぎる実装への依存性が指摘され，既存のデータマイニングツールである Orange にサービス指向の実行機構である Web サービスを適用した先行研究 [Podpečan11] をはじめ，多くのツールの開発が進められている．Web サービスは，実装言語に依存しない意味（セマンティクス）づけを行い，プログラムからの呼び出しを可能にする方法であり，オブジェクト記述と機能呼び出しのそれぞれに重きを置いた 2 つのアプローチが存在する．オブジェクトの記述に重きを置いたものは， WADL(Web

Application Description Language) 等を用いて，WebAPI と呼ばれる REST 型の Web サービスである．SOAP 型の Web サービスでは，オブジェクト設計者が意図した処理メソッドを提供できるが，利用者が自由に処理メソッドを追加することは困難である．

一方，機械可読なセマンティクスを持つ処理メソッドの記述は， OWL-S[*3]，および WSDL(Web Service Description Language) に基づく記述による，処理の遠隔呼び出しを指向した SOAP 型の Web サービスである．SOAP 型の Web サービスでは，実装とは独立した入出力を実現するため，処理メソッドへの入出力として文字列など基本データ型のみがやり取り可能である．このため，実装に依存せず，セマンティクスを付与したオブジェクトを入出力するためには，REST 型の Web サービスの利用などを考えなければならない．

これら Web サービスの各アプローチの特徴を踏まえ，TETDM のように自由に開発者が開発した処理メソッドを追加していくためには，テキストマイニング関連オブジェクトおよび各処理メソッドの記述の共通語彙のとして，メソッドの実装とは独立したセマンティクスを提供することが必要であると考える．このため，3. 章に示したテキストマイニング関連オブジェクトおよびメソッドの体系化を踏まえ，2. に示した各種テキストマイニングプロセスが実行可能な環境を Web サービスにより実装する際は，図 5 に示すような構成を考えていく必要がある．

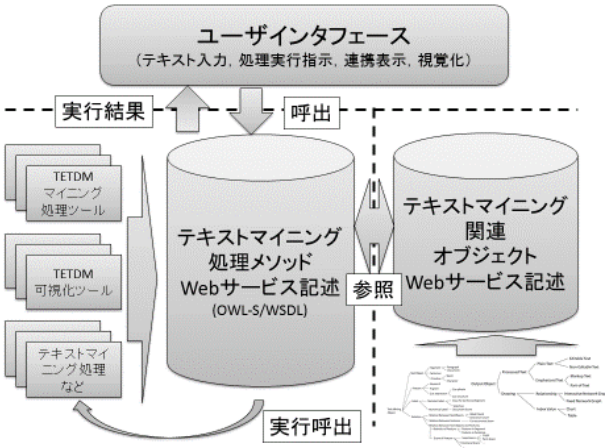


図 5: Web サービスを利用した統合型テキストマイニング環境の概観.

本枠組みでは，開発者は以下のように開発を行うことを想定している．

1. 処理や可視化内容を考案
2. テキストマイニング関連オブジェクト共通語彙による入出力，および参照を決定
3. 任意のプログラミング言語によって，処理メソッド（マイニング処理，可視化など）を実装
4. Web サービス化を行う記述を作成 [*4]

一方，これまで TETDM の実装として蓄積された知見に基づき，任意の実装によるインタフェースは，以下の機能を最低限として，利用者に提供することが求められる．

- テキストの入出力機能
- 処理メソッドの実行指示
- 可視化結果の表示
- 表示連動機能（テキストマイニング関連オブジェクトのインスタンス取得による）

6. おわりに

本稿では，テキストマイニングメソッドにおける入出力と参照にあたるオブジェクトをさらに整理し，オブジェクト階層として示した．さらに，TETDM プロジェクトにおいてこれまで開発されてきた多粒度のテキストマイニング関連メソッド群をメソッドとして整理し，公開・利用するための基盤として Web サービスに基づく実装について，検討した．なお，詳細な実装については，Web ページ [*5] を通じて順次公開を行っていく．

参考文献

[砂山 13] 砂山，高間，西原，徳永，串間，阿部，梶並：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)

[阿部 14] 阿部：TETDM におけるテキストマイニング関連オブジェクトの整理と実装，2014 年度人工知能学会全国大会（第 28 回），1H4-NFC-01a-4 (2014)

[利根川 15] 利根川，高間：協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムの提案，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 (第 9 回)，SIG-AM-09-02 (2015)

[阿部 12] 阿部：テキストマイニングにおける語句計量化指標群の利用に関する一考察，2012 年度人工知能学会全国大会（第 26 回），3K2-NFC-3-2 (2012)

[砂山 14] 砂山，高間，西原，梶並，串間，徳永：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)

[Podpečan11] V. Podpečan, M. Zemenova, and N. Lavrač: Orange4WS Environment for Service-Oriented Data Mining, The Computer Journal, doi: 10.1093/comjnl/bxr077, (2011)

*3 http://www.w3.org/Submission/OWL-S/.
*4 手順 3 と 4 は順不同.
*5 http://abe-lab.jp/tools/