1H4-NFC-01a-4

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

Constructing Conceptualization of Text Mining Objects for Text Mining Methods in TETDM

阿部 秀尚 *1
Hidenao Abe

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

Many text mining tools such as TETDM have been developed without considering conceptualization of text mining related objects, which contains input text, given labels, output results and so forth. Users of text mining often face on some difficulties for understanding adequate processes for their input text. On the other hand, developers who want to implement their method for text mining need more concrete descriptions of input, output, and referencing objects for their developing methods. In this study, I considered some text books for text mining learners for constructing a conceptualization of text mining objects. Then, a design for implementing the objects into the TETDM is discussed.

1. はじめに

テキストマイニングをはじめとする統計的な自然言語によるテキストから知見を得るための手法は，1900 年代初頭から始まった統計的文体分析を端緒として様々な手法が開発されてきた．また，統計解析手法によるアンケート記述の分析など，テキストを対象とした統計解析事例も数多く存在する．さらに，コーパスに基づく統計的自然言語処理も近年の計算機技術の進歩により，より多くの処理手法が開発されてきている．このような背景の下，テキストマイニングを実行するツールやこれらのツールを用いる分析がより多くの現場で適用され，その有用性の認知が進んでいる．

ところが，テキストマイニングに関連する処理内容（以下，メソッド）を理解するためには，実際にテキストマイニングを実行し，経験を積むことが要求されている．これは，分析を実行する利用者に留まらず，開発者がメソッドの改良を行う際，どのような対象を扱うのかを把握するためにも困難が伴う．これらの困難は，テキストマイニングに関連する各メソッドが「何を」どう扱うのかが明示的でないため，テキストマイニング処理全般から詳細への共通理解が利用者と開発者の間で十分とられていないために生じている．同様の問題として，大規模かつ複雑なプロセスを伴うソフトウェア開発では，入出力などに関連するオブジェクトを整理し，ライブラリを整備してプロセスの組換えを柔軟に行うことの有効性が指摘されている [三輪 12].

そこで，本研究では，テキストマイニングツールが実行する種々のメソッドを複雑なソフトウェア部品と考え，実装されたツールの１つである TETDM を対象にメソッドの切り出しを考察してきた [阿部 13]. [阿部 13] における課題として，テキストマイニングメソッドが扱う対象物（以下，テキストマイニングオブジェクト）のより具体的な概念化が必要であることが明らかとなった．

本稿では，TETDM プロジェクトが提案する初歩的なテキスト加工技術を含むテキストマイニングプロセスにおいて，扱われるべきテキストマイニング関連オブジェクトの概念化を示す．さらに，ここで概念化したテキストマイニング関連オブジェクトについて，テキストマイニング環境である TETDM への実装について，その設計を示す．

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

2. テキストマイニングプロセス理解の課題

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 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を行い，結果の評価を行う一連の工程として示される．実際のテキストマイニングの実行では，図 1 中の点線で示すように，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．ところが，これらの洗練化工程の実際は，事例研究においても記されることは少なく，利用するテキストマイニングツールも一連の処理を一通りしか実行できないことが多い．このため，学習者から見たテキストマイニングプロセスは一通りの実行過程であり，初学者が反復的なテキストマイニングプロセスを理解し，各自で洗練化を実施する上での必要な支援が十分提供されているとは言い難い．

図 1: テキストマイニングの典型的なプロセス．

そのため，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，インタフェースと内部データの連動機構を用意した TETDM の開発を行っている．TETDM で実現されるテキストマイニングプロセスを図 2 に示す．本ツールでは，通常，１つの流れしか実現できないテキストマイニングプロセスの実行を各段階の処理結果の提示を受けつつ，パラメータの変更などを別のマイニング処理に反映させることが可能である．

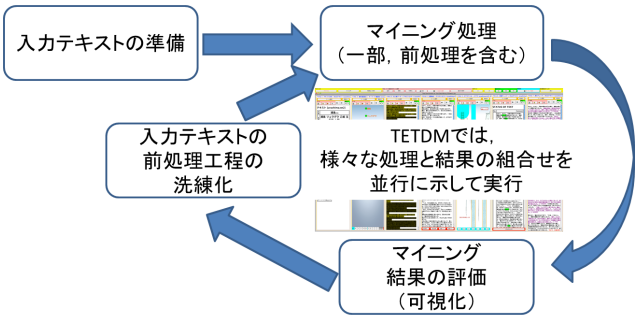


図 2: TETDM が前提とする反復的なテキストマイニングプロセス．

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．まず，テキストマイニングに関連した書籍 [那須川 06, 石田 12] 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 [砂山 13, Mahout] 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．次に，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．この結果を図 3 に示す．

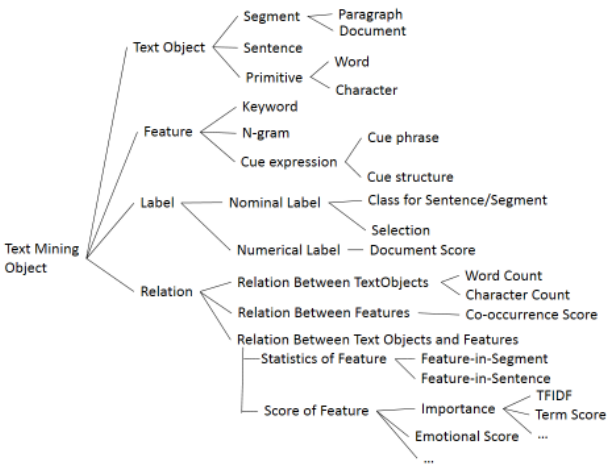


図 3: 構築したテキストオブジェクトの概念階層．

3.1 入力に関連するオブジェクト群

TETDM において，入力テキストはセグメント，文，単語から成ることが想定され，それぞれの間での包含関係から成ると定義されている．以上の認識は，これまで TETDM の開発者が暗黙のうちに獲得したものであり，これを明示的な概念階層構造として定義したものが図 3 中の Text Object の部分である．処理ツールおよび可視化ツールは，これらの入力テキストにあるオブジェクトを統合環境から入力オブジェクトとして取得し，それぞれの処理を行う．

入力テキストに存在するオブジェクトを階層化した概念構造の各葉接点では，Implemented-as という属性を与え，クラス名を指定することで実装との対応付けができる．

表 1: 処理ツールの出力データ型とそれに関する説明テキスト（README.txt）中の名詞の出現頻度．

順位	String 語	String 出現頻度	int[] 語	int[] 出現頻度	double[] 語	double[] 出現頻度
1	テキスト	15	リスト	8	値	5
2	単語	8	キーワード	7	評価	5
3	データ	7	指定	6	単語	4
4	列	7	番号	6	数値	1
5	辞書	5	単語	6	関連	1
6	文字	5	色	5	データ	1
7	入力	5	表示	5	リスト	1
8	結果	4	ID	5	主題	1
9	基準	4	主題	4	結論	1
10	評価	4	列	4	度	1

順位	String[] 語	String[] 出現頻度	int 語	int 出現頻度
1	集合	4	数	4
2	単語	2	単語	2
3	ラベル	1	上	1
4	ノード	1	ノード	1
5	エリア	1	エリア	1
6			リスト	1
7			チェック	1
8			地図	1
9			スコア	1
10			結果	1

表 2: 可視化ツールの入力データ型とそれに関する説明テキスト（README.txt）中の名詞の出現頻度．

順位	String 語	String 出現頻度	int[] 語	int[] 出現頻度	double[] 語	double[] 出現頻度
1	表示	6	数値	6	集合	6
2	テキスト	6	キーワード	6	ため	3
3	ボックス	3	情報	5	ボックス	3
4	ラベル	3	色	5	値	3
5	集合	3	テキスト	5	評価	3
6	チェック	3	指定	5	キーワード	3
7	軸	2	ID	4	チェック	3
8	色	2	単語	3	数値	2
9	単語	2	列	1	データ	2
10	関連	1	配列	1	列	1

順位	String[] 語	String[] 出現頻度	int 語	int 出現頻度
1	表示	4	数	6
2	テキスト	2	チェック	3
3	区切り	2	集合	3
4	結合	2	ボックス	3
5	列	1	ノード	1
6	ラベル	1	クラスタ	1
7	ノード	1	エリア	1
8	エリア	1		
9	キー	1		
10	フレーズ	1		

3.2 処理ツールの出力・参照に関連するオブジェクト群

TETDM における処理ツールからの出力はこれまで，boolean 型，int 型，double 型，String 型について，それぞれの値，一次元配列，二次元配列として定義されている．ところが，処理ツールの出力と可視化ツールの入力について，これらのデータ型とツールに添付された説明文の特徴語との間での顕著な対応は，表 1 および表 2 に示すように見られなかった．

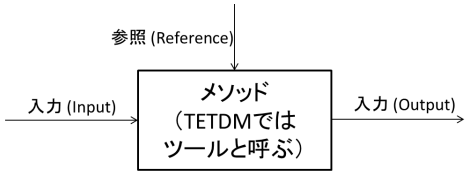
これは，それぞれのツールにおいて，語彙上の共通理解が無く，入出力のデータ型をそれぞれの概念で捉え，利用していることに起因すると考えられる．このため，実装される基本のデータ型よりも概念化の度合いの高いクラスとして，これらの入出力を表現する必要がある．図 3 中の Feature[*1] 以下のオブジェクトは，処理ツールからの出力であり，可視化ツールへの入力・参照となる．例えば，各文を特徴づけるための素性を作成するメソッドがあったとき，これらは Text Object 同士の関連を示す値である Word Count などを利用して，特徴

*1 ここでは，選別された特徴語などを意味し，bag-of-words を構築するための素性となるものを表している．

語 (keyword) や n-gram，手掛かり表現 (cue expression) の有無などが素性として選定され，出力される [*2]．同時に，各素性の重みを評価指標として計算する処理ツール [阿部 12] などが，素性として選定されたオブジェクトの重みを計算する．TETDM では，連動機構を通じてこれらのオブジェクトへの処理結果を並行する処理ツールと可視化ツールの組み合わせにより，反映させることが可能である．

4. テキストマイニングメソッドの概念化

以上の考察により，テキストマイニングメソッドを図 4 のようにとらえる．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，[阿部 13] 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	可視化ツール
Module ID: 整数型の値	Module ID: 整数型の値
Implemented-As: 文字列型の値	Implemented-As: 文字列型の値
Input: Text Mining Objectで定義	Input: Text Mining Objectで定義
Reference: Text Mining Objectで定義	Reference: Text Mining Objectで定義
Output: Text Mining Objectで定義	Output: Output Objectで定義

図 4: テキストマイニングメソッドの定義の改善結果．

5. TETDM におけるテキストマイニングオブジェクトの実装についての検討

本節では，図 3 に整理したテキストマイニング関連オブジェクトの実装について，検討を加える．図 5 に各テキストマイニング関連オブジェクトに関連したクラスを具体化したクラス図を示す．また，概念階層に基づくクラスのうち，実際の入力テキストに対応した TextData 型の関連を示す．

まず，TextData 型については，図 5 中の記述を基に設計し，TETDM に実装する．さらに，これらのクラスを用いて，マイニング処理ツールの出力結果を扱い，可視化ツールへの入力・参照となるクラスの設計と実装を行う．

6. おわりに

本稿では， TETDM プロジェクトが提案する初歩的なテキスト加工技術を含むテキストマイニングプロセスにおいて，扱われるべきテキストマイニング関連オブジェクトの概念化の結果を示した．さらに，ここで概念化したテキストマイニング関連オブジェクトについて，テキストマイニング環境である TETDM への実装について，その設計を示した．

今後は，実際のツールへの実装を行い，テキストマイニングプロセスの実行によって検証を行う．また，テキストマイニングオブジェクトおよびメソッドの定義を充実させることにより，ツール開発者の支援方法について検討を行っていく．

*2 実装されるデータ型との対応は，各オブジェクトの属性として記述する．

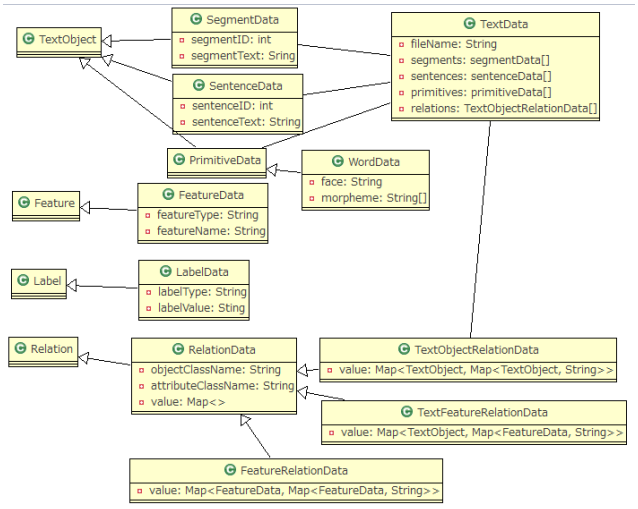


図 5: 構築したテキストオブジェクトの概念階層に対応したクラス階層．

参考文献

[三輪 12] 三輪一郎：”RC2E ”(リポジトリ中心の CASE 環境) 普及の価値と課題，第 8 回情報システム学会全国大会，P030 (2012)

[阿部 13] 阿部秀尚：TETDM モジュール構成に基づくテキストマイニングメソッドの概念化に関する一考察，2013 年度人工知能学会全国大会（第 27 回），3B3-NFC-01a-2 (2013)

[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)

[那須川 06] 那須川哲哉：テキストマイニングを使う技術／作る技術，東京電機大学出版局 (2006)

[石田 12] 石田基広ら：コーパスとテキストマイニング，共立出版 (2012)

[Mahout] Apache Mahout: `https://mahout.apache.org/`.

[阿部 12] 阿部秀尚：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)