3B3-NFC-01a-2

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

Considering Conceptualization for Text Mining Methods based on TETDM Modules

阿部 秀尚[*1]
Hidenao Abe

[*1]文教大学
Bunkyo University

In many text mining tools such as TETDM, they contain potential difficulties on user-developer interaction because of lack of common understandings about text mining methods. Users of text mining often face on some difficulties for understanding how the presented methods process their input text. However, they know what to do for their owned text. On the other hand, developers who want to implement their method for text mining need more concrete guide line with more convenient explanations for users. In this study, I consider about conceptualization of text mining methods for understanding them with common concept. Based on the implemented method as TETDM modules and their short summaries, I enumerate their input, output, and referenced objects on these text mining methods. Then, a database for retrieving adequate methods depended on users' demands is described.

1. はじめに

テキストマイニングツールのように，多くの処理手法が列挙され利用可能なツールにおいて，ノービスユーザーが入力と要求に沿うように処理手法を適切に適用するためには困難が伴う．適切な処理手法が提供されているにも関わらず，これを利用できないことは大きな損失である．一方，テキストマイニングの手法について新たな手法を開発し，利用促進をしようとするとき，TETDM をはじめとする統合型テキストマイニングツールは有効な手段である．ところが，適切な利用場面を全て想定することは開発者にとって困難であり，一定の基準に基づいて提供する処理手法の説明を記述することが求められている．以上の問題は，大規模なソフトウェア開発やライブラリの利用において重要な研究課題であり，利用者，開発者双方の理解を促進するための知識基盤を構築することが必要とされている [三輪 12].

本研究では，TETDM でのモジュールとして作成されてきたマイニング処理および可視化について，各モジュールの入出力に注目し，テキストマイニングメソッドとしての明示的な概念化について考察する．TETDM では，テキストマイニングやプログラミングに関しての初学者でもモジュール作成の提案に参画できることを目指している．このため，テキストマイニング中の個別の処理内容よりも処理の入出力に着目し，一連の処理が可能となるよう支援するための情報資源が必要だと考えられる．また，開発に不慣れな開発者の支援策として，利用者と開発者双方で用いることができるテキストマイニングメソッドの概念体系について示し，議論する．

2. テキストマイニングメッソドリポジトリの整備による利用者・開発者支援

テキストマイニングでは，自然言語処理に基づく処理の他，情報の可視化を含めたテキストからの有用情報獲得が求められている．そのため，種々のテキストマイニング事例からユーザは，各自が所有するテキストへの適用にふさわしいテキストマイニング手法を選定し，実行する必要がある．しかしながら，テキストマイニングにおける処理は多岐にわたり，理解を促進する可視化処理も加わると，さらに選定が困難となる．また，テキストマイニング手法の開発者が開発したテキストマイニングにおける各処理が，どのような入力テキストに対し有効であり，どのような処理をし，どのような結果をもたらすかを非統一なテキストとして提供してきた．このように，テキストマイニングツールの開発を効率的に進めるためには，手法の開発者による記述を統一して理解する基盤となる語彙情報が重要な役割を果たす．

統合型のテキストマイニングツールのように，大規模なシステム開発や複雑なソフトウェアの開発では，ソフトウェア部品の動作内容に応じて，開発者の負担を軽減する支援が求められている．三輪らは，情報システム開発におけるソフトウェア・リポジトリの活用を提唱し，実際の開発現場での上流工程への適用を試みようとしている [三輪 12]．また，阿部は，PSM を考慮したメソッド定義に基づいてデータマイニングで用いられるマイニング手法のメソッドを同定し，構成的メタ学習によるデータマイニングアプリケーションの生成手法を開発してきた [阿部 04]．以上のような知見は，テキストマイニングの処理過程についても適用可能であると考えられ，開発者と利用者双方がテキストマイニングにおける処理単位であるテキストマイニングメソッドの概念を共有することが理解の促進に繋がると考えられる．

開発されたテキストマイニングメソッドは，それぞれのメソッドを規定し，整理することで各メソッドの情報を記述することでリポジトリとして有効に機能するようになる．また，複雑な処理や評価基準を伴うタスクについて，多岐にわたる処理手法を集積し，公開することは，当該タスクを扱う分野の発展に寄与する重要な活動の 1 つと言える．例えば，CAIDA[CAIDA] では，インターネット上の主にネットワーク構造に着目して有用な情報を抽出するタスクに関わる処理プログラムの集積と公開を 1 つの活動として行っている．CAIDA の Web サイトでは，40 以上の処理プログラムが公開され[*1]，一部はタクソ

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@a.email.ne.jp

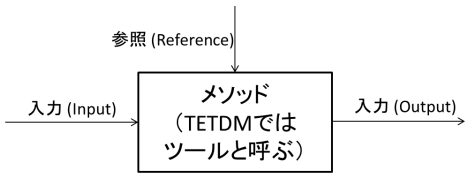
*1 http://www.caida.org/tools/

ノミーに従って整理され，処理プログラムを検索することが可能[*2]となっている．以上のように，ユーザ・開発者（処理手法の研究者を含む）に対して，整理されたメソッドの記述を提供する基盤は，双方がより深く対象とするタスクを理解するために有用であると考えられる．さらに，テキストマイニングメソッドのテキストマイニングによる同定手法や自然言語処理での情報抽出手法 [Mooney05] などを適用し，テキストマイニングにおけるメソッドを定義していくことで，より多くのテキストマイニングメソッドの開発が可能となると考えられる．

3. TETDM に基づくテキストマイニングメソッドの概念化

2. 章では，メソッドリポジトリのソフトウェア開発における有用性について述べたが，ここでは，実際のテキストマイニングツールを取り上げ，リポジトリ構築について考察する．TETDM[*3]のメソッドは，テキストの統計的データを利用して処理を行う「（マイニング）処理ツール」と処理結果の可視化を行う「可視化ツール」に分けて開発が進められている．TETDM プロジェクトでは，同時にユーザ，開発者双方の支援方法の開発も統合環境の一部として考えられており [砂山 13]，テキストマイニングメソッドリポジトリの構築はユーザと開発者が相互に各メソッドを理解するための共通基盤であると言える．

以下の考察では，テキストマイニングメソッドを図 1 のように捉え，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，定義する．



処理ツール	可視化ツール
Module ID: 整数値	Module ID: 整数値
Implemented-As: 文字列	Implemented-As: 文字列
Input: Text Objectで定義	Input: Processed Text Objectで定義
Reference: Text Objectで定義	Reference: Text Objectで定義
Output: Processed Text Objectで定義	Output: Output Objectで定義

図 1: TETDM におけるテキストマイニングメソッドの定義.

3.1 TETDM におけるオブジェクトの定義

TETDM では，TextData 型のクラスを入力テキストに対して実体化するため，以下の処理を自動的に実行する．

1. セグメント（段落，文章）の切り分け
2. セグメントに含まれる文の識別
3. 形態素解析による単語の切り出し
4. セグメント内の単語出現，文中の単語出現に基づく索引付け

*2 http://www.caida.org/tools/taxonomy/
*3 TETDM プロジェクトで作成されているツールの名称を意味する．

このように，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一例が図 2 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，あるいは参照データとして取得し，それぞれの処理を行う．

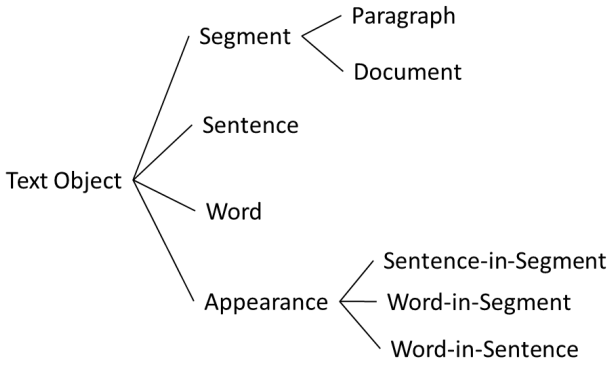


図 2: TETDM におけるマイニング処理ツールへの入力，および各ツールへ参照として提供されるテキストオブジェクトの概念階層（http://tetdm.jp/pukiwiki/index.php?テキストデータ を基に作成）.

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-by という属性を与え，属性値として TextData 型でのメンバ変数名を与えることで，実装との対応付けができると考えられる．

また，TETDM の可視化ツールでは，処理ツールで行った様々な形式の結果を入力テキストや処理結果と組み合わせて効果的に表示することを可能としている．これは，様々な研究者がテキストマイニング手法を個別に開発してきた検証を含めたテキストマイニングの結果表示を区分することで，様々な形式での結果表示が可能にしようとするものである．このため，図 3 に示すように各種のオブジェクトの形式として，ユーザのテキストからの有用な情報の獲得を支援する必要があると考えられる．

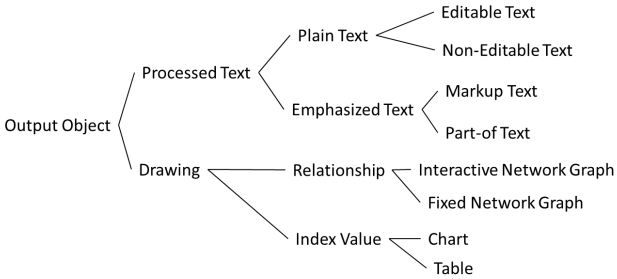


図 3: TETDM における可視化ツールの出力オブジェクトの概念階層（http://tetdm.jp/pukiwiki/index.php?可視化ツール を基に作成）.

一方，マイニング処理によって処理された結果は，ツールでの実装毎，様々な形式で出力されてきた．TETDM では，こ

れらのマイニング処理からの出力を以下のデータ型を用いて，視覚化モジュールに入力している[*4]

- 値（スカラ）
  - int
  - double
  - boolean
  - String
- 一次元配列（リスト）
  - int[]
  - double[]
  - boolean[]
  - String[]
- 二次元配列（マトリクス）
  - int[][]
  - double[][]
  - boolean[][]
  - String[][]

3.2 TETDM におけるテキストマイニングメソッド

TETDM における各処理ツールおよび各可視化ツールは，必ずマイニング処理モジュール（MiningModule）か視覚化モジュール（VisualizationModule）のいずれかの継承を求め，サブクラスの継承を非推奨としている．このため，クラス階層を示すと，それぞれ１つスーパークラスにこれを継承した全てのサブクラスが配置される一階層の構造となってしまう．しかし，図 1 に対応するそれぞれのトップレベルのメソッドを図 2，図 3 に示すに加えて，処理済みテキストオブジェクトの階層構造を参照することで，中間層となる仮想的な抽象メソッドを定義することが可能である．現在，処理済みテキストの階層構造構築を行っており，これと合わせて，処理ツール・可視化ツールとして実装された各テキストマイニング関連メソッドと葉節点が対応する階層構造を構築することを想定している．図 4 に処理ツールの実装に対応される処理メソッドの階層構造の構想される例を示す．

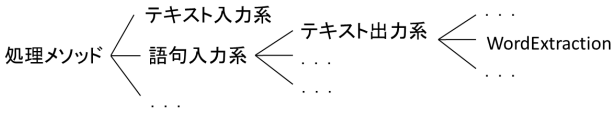


図 4: TETDM における処理メソッドの入出力オブジェクトを基準とする場合の階層構築例.

3.3 メソッド定義に基づく利用者・開発者支援

新たなモジュールの開発者向けには，入出力および参照の各オブジェクトで記述される最も詳細なレベルでのオブジェクトを指定して，その他の記述項目と共に登録を行うインタフェースを用意する．一方，ユーザや開発者に移行する途上の開発者に対しては，各メソッドの記述レベルを上げ，オブジェクトのラベルや概要説明を中心に情報提供とともにモジュールを提供するインタフェースの作成が考えられる．

*4 現在，マイニング処理ツールから共通の認識が得られるよう議論が進められている．

4. おわりに

本稿では，TETDM のモジュール作成に関するガイドラインと実装されたメソッド群を基にテキストマイニングメソッドの概念化について考察した．今後は，テキストマイニングに関する手法を紹介する研究論文・書籍等に対してテキストマイニング手法を適用し，更にメソッドの同定およびオブジェクトの洗練を行っていくことが課題である．

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 04] Abe, H., Yamaguchi, T.: Constructive Meta-learning with Machine Learning Method Repositories. In Proc. of IEA/AIE 2004, pp.502-511 (2004)

[CAIDA] The Cooperative Association for Internet Data Analysis: http://www.caida.org/home/

[Mooney05] Mooney, R. J. and Bunescu, R.: Mining knowledge from text using information extraction, SIGKDD Explor. Newsl, Vol.7, No.1 (2005)

[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)