TETDMを用いた汎用性を考慮したシステムの設計指針に関する基礎的検討

Consideration of Design Guide for Constructing General Purpose System using TETDM

梶並 知記[1*]　田代 航一[2]　利根川 拓馬[2]　北村 侑也[2]　高間 康史[2]
Tomoki Kajinami[1]　Koichi Tashiro[2]　Takuma Tonegawa[2]
Yuuya Kitamura[2]　Yasufumi Takama[2]

[1] 神奈川工科大学
[1] Kanagwa Institute of Technology
[2] 首都大学東京
[2] Tokyo Metropolitan University

Abstract: This paper considers a collaborative policy for combining tools, in development of system using TETDM. TETDM is an total environment for text data mining, can prepare for various mining tasks by combination of small mining tools. However, an useful guide in the design of system constructed with several small tools developed by different tool developers has not been considered. This paper describes a design guide adjusting user's purpose and system's specifications for constructing general purpose system, and shows an example of practice.

1　はじめに

本稿では，TETDMを用いたシステム開発における，ツール同士の連携方針について検討する．TETDMは，テキストデータマイニングのための統合環境であり，小規模なツール同士を連携させることで多様なタスクへ対応可能としている [6]．ツールの種類は,「マイニング処理ツール」と「可視化ツール」の２つに分類され，ユーザは任意のマイニング処理ツールと可視化ツールを１つずつ選択し，それらを１対１で組み合わせて使うことで，テキスト分析を行う．ここで，１対１の組み合わせは複数種類同時に使用することが可能で，TETDM上の複数枚のパネルそれぞれに，マイニング処理ツールと可視化ツールの組が１つずつ配置される．これにより，同時にさまざまな観点からテキスト分析を行うことが可能になっている．また，TETDMでは，ツール間の連動として，他のツールから出力されるデータを別のツールで利用する仕組みが用意されている．これにより，テキスト分析にとどまらず，複数のツールからなるテキストデータマイニングシステムを開発するプラットフォームとしてTETDMを活用することも，可能となっている．

*連絡先：神奈川工科大学情報学部情報工学科
〒 243-0292 神奈川県厚木市下荻野 1030
E-mail: kajinami@ic.kanagawa-it.ac.jp

しかしながら，複数の開発者が個別に開発した小規模ツール同士を連携してシステムを設計する指針については未検討である．本稿では，対話的なクラスタリング環境の構築を目的としたシステム開発において，目的優先と手段優先志向を摺り合せるシステム設計指針について述べる．TETDMの仕様を変更せず，仕様の制限がある中で文章を対象にした対話的なクラスタリング環境を構築する本研究の意義は，以下の３点である．

1. TETDMに備わっているツール連動の仕様から逸脱せず対話的なクラスタリング環境をTETDM上に実装する例を示すことで，TETDM上に新たなプラットフォームを構築するシステム開発に応用できる．
2. 可視化ツールとの組み合わせを想定しない複数のマイニング処理ツールを統合的に扱う手法を提案することで，マイニング処理ツールと可視化ツールを１対１対応させるTETDMの特徴を活かしつつ，TETDMを拡張する方向性を示す．
3. 対話的なクラスタリングのための，統一的なデータのやり取りを可能とすることで，クラスタリングに関連するツールを，同一環境上で比較し易くなる．

本稿では，複数のツールを連携し TETDM 上に対話的なクラスタリング環境を構築することを目指すが，特定のクラスタリング手法に特化したり，技術文書の分類や商品レビューの分類といった特定のタスクに特化するものではなく，汎用的なものである．そのため，汎用性を意識した，ツール連携の方略を検討する．

本稿の構成は以下のとおりである．2 節で，TETDM の応用に関する研究について述べ，本稿の位置づけを明確にする．3 節で，対話的なクラスタリング環境におけるツールの役割の同定や，ツール同士でやりとりするデータの内容を抽象化，データの型を定義する．4 節で，ツールを統合的に扱う管理パネル方式の提案を行い，5 節で，試験的なシステム実装例を示す．

2　関連研究

2.1　TETDM の活用や拡張

TETDM を用いることで，ユーザはさまざまなツールから，システム上可能な範囲で任意の組み合わせを選択して，テキスト分析処理の結果を得ることができる．実践的な活用例として，医療現場でのカルテ分析がある [7]．また，R といった既存の分析ソフトウェアと連携し，TETDM を拡張する研究もおこなわれている [8]．TETDM の拡張に関する研究として，マイニング処理ツールと可視化ツールの組み合わせをユーザが能動的に選択する必要がある TETDM の特徴に着目しているものがある．TETDM のコアとなるプログラム部分もオープンソースであることを活かしてツールの組み合わせ作業の支援が行われている [2][4]．

本稿では，テキスト分析を行う特定の現場を想定したものではなく，また，既存のソフトウェアとの密な連携を目指すものではない．本稿は TETDM の拡張に関する研究であるが，TETDM のコアとなるプログラム部分には触れず，マイニング処理ツールと可視化ツールの 2 種類のツールを実装する枠組み，TETDM の仕様に従いツール同士を連携させる枠組みの中で，新たなプラットフォームを構築するものである．

2.2　対話的なクラスタリング

対話的なクラスタリングは，ユーザの要求に応じたクラスタリング結果を出力するための方法で，ユーザによるクラスタリングに必要なパラメータ，制約の入力を支援する [3]．ユーザは，自身の意図や背景知識を考慮したクラスタリングへの制約付与を行い，クラスタリングした結果とのインタラクションを繰り返し，望みのクラスタリング結果を得る．対話的なクラスタリングは，文書の分類に応用されている [5]．

本稿では，対話的なクラスタリング環境の構築を目指すが，ツールの組み合わせによってさまざまな視点からテキストデータを眺め，インタラクティブに分析する TETDM と，異なるクラスタリング結果を並列に眺め，そこからユーザの意図や背景知識に応じて，反復的にクラスタリングを行う対話的なクラスタリング環境には親和性があると考える．

3　クラスタリングのためのツール連動

本稿では，ユーザが複数のクラスタリング結果を見比べることができ，また使用するクラスタリング手法，各種パラメータの設定が動的に行える環境の構築を想定したシステム設計の方略を考える．また，クラスタリング結果（可視化）としてユーザが見たい情報は，クラスタ集合，それに含まれるクラスタ，クラスタに分類されている文書，文書内の単語の 4 種類であると想定する．3.1 節で，クラスタリングの流れを 3 段階にわけ，マイニング処理ツールや可視化ツールとの対応について述べる．3.2 節で，対話的なクラスタリングのための，ツール間連動で用いるデータ型について述べる．3.3 節で，複数人からなるシステム開発の中で実際に行った，ツールの分類作業，ツール間連動の整合性確認作業について述べる．

3.1　クラスタリングの流れ

クラスタリングの実行手順を大きく 3 段階に分けると，以下のようになる．可視化処理段階と TETDM の可視化ツールは自然と適合するが，前処理とクラスタリング処理は，ともにマイニング処理ツールとして実装する．

前処理 クラスタリングする文書のベクトル化・特徴量の算出する段階

クラスタリング処理 任意のクラスタリング手法によりクラスタリングする段階

可視化処理 選択したクラスタリング手法に応じた/ユーザの意図に応じた可視化手法によって，クラスタリング結果を出力/フィルタリングする段階

複数のツールを組み合わせてクラスタリングシステム全体を構成するため，最低でも各段階 1 つずつのツールを連結することで，クラスタリングが一通り完了できることになる．

3.2　クラスタリングに必要なデータ型

本節では，ツールの役割分担を考える際，システム設計の際に採用されるデータの流れに着目する考え方 [1] を参考に，ツール間でやりとりするデータに具体性を持たせて検討を行う．

ここでは，前処理段階，クラスタリング処理段階，可視化段階の間にどのようなデータが必要であるか検討する．できるだけ複雑にならず，なおかつユーザが必要とする要素（クラスタ集合，それに含まれるクラスタ，クラスタに分類されている文書，文書内の単語）を表現するのに十分なデータ型である必要がある．なお，TETDM の仕様に従い，ユーザからシステムに入力するものはテキスト形式の文書ファイルとする．入力文書は単一ファイルとは限らず，複数の文書ファイルにも対応できる．また，TETDM の標準的な機能により，文書を段落や文章，単語に分割する操作は完了しており，文書内の文章数や単語数などは特定の変数に格納され，また特定の単語などを，配列の要素数（ID）を指定することで一意に定めることができることを前提としている．したがって，本稿では，ツール間で具体的にやりとりするデータの内容を文書ベクトルリスト，クラスタ文書リスト，クラスタ単語リストの 3 つとし，TETDM で用意されている，ツール連動用のデータ型に対応させる．文書ベクトルリストは，文書と単語の 2 次行列で定義する．中身は，任意の特徴量（TF-IDF など）によって計算された各単語の重みとなる．クラスタ文書リストは，クラスタを行，クラスタに含まれる文書を列とする 2 次行列で定義する．クラスタ単語リストは，クラスタを行，クラスタに含まれる単語を列とする 2 次行列で定義する．

表 1 は，文書ベクトルリスト，クラスタ文書リスト，クラスタ単語リストについて，TETDM で用意されているデータ型との対応を示している．クラスタ文書リストの部分に，boolean と double の 2 つの型があるが，列数が全文書数ありクラスタに含まれている文書を 1，含まれていない文書を 0 とする 2 値表現を行う場合と，あるクラスタに含まれている文書 ID を配列として格納する場合の両方に対応するためである．

表 1: 具体的なデータとデータの型.

データ	型
文書ベクトルリスト	double[][]
クラスタ文書リスト	boolean[][]，int[][]
クラスタ単語リスト	double[][]

3.3　実際の設計方略

3.1 節と 3.2 節で述べた，段階分類とデータの定義に基づき，本稿で実際に行ったシステム設計方略は以下のとおりである．

1. ツール名と入出力データの内容と処理内容を記載するカードを用意
2. 複数の開発者（プロジェクトメンバ）による，カードへの記載
3. ツール同士の入出力データのマッチングを精査
4. ツールの入出力データ再検討や，ツールの分割や統合

1 つのツールを 1 枚のカードで表現し，前処理，クラスタリング，可視化の 3 段階に分類されたツールをつなぐために，データ入出力の整合性をとる流れである．データ入出力の整合性がとれない場合は，処理内容と入出力データの関係が適切かどうか，またツールの処理内容を分割または統合可能かどうか検討する．なお，前処理，クラスタリング，可視化のいずれかに当てはめるのが難しいツール，特定のクラスタリング手法に依存するツールに関しては，別途オプションカテゴリとする．

上記方略の（1）と（2）が，開発プロジェクトの目的を考慮した目的優先の志向に対応し，(3）と（4）が，TETDM の仕様から実現可能な手段を考慮した手段優先の志向に対応する．すなわち，開発者やユーザの考える,「実現したいこと」の「入出力データが何か」検討し，TETDM のツール連動の仕組みに適合するようなデータの流れになるよう，調整していく．

表 2 に，具体的に出されたツール案の一部を示す．前述したクラスタリングの段階ごとに，ツールを分類している．括弧内のものは，オプションカテゴリのものである．また，本研究は教育機関で実施しており，著者らの一部（工学系学生，大学院生）のクラスタリング手法に関する学習も兼ねている．したがって，ここで既存のクラスタリング手法のすべてを列挙することは目指していない．

表 2: クラスタリングの段階とツール群.

段階	ツール
前処理	TF-IDF 計算, BM25 計算
クラスタリング	K-means，階層的クラスタリング，制約付き階層的クラスタリング，(重心計算，距離計算)
可視化	ネットワーク型図，階層構造図

4　管理パネル方式によるツール管理

本節では，複数のマイニング処理ツールの管理を行う管理パネルを TETDM 上に構築し，1 つのパネルを利用してツールの組み合わせを変更する，管理パネルモデルを提案する．便宜上，ここでは TETDM で採用されている基本的なツールの管理を「基本方式」，管理パネルモデルによるツールの管理を「管理パネル方式」を呼ぶ．

4.1　基本方式の問題点

図 1 に，基本方式に基づく，クラスタリング環境を示す．基本方式では，マイニング処理ツール同士の連携（データのやり取り）が許されているものの，1 枚のパネルにマイニング処理ツールと可視化ツールを 1 対 1 で組み合わせて配置する．ツール開発者の視点では，マイニング処理ツールを開発する際に，必ずなんらかの可視化ツールとセットで使われることを想定しておかなくてはならない．TETDM で用意されている，マイニング処理ツール同士の連動機能を用いて，他のマイニング処理ツールでのみ使われるデータを出力するツールの作成も可能であるが，原則的に，TETDM ではマイニング処理ツールと可視化ツールを 1 対 1 対応させる設計方針となっている．このことは，ツール利用者（ユーザ）側においても，問題となる．ツールの組み合わせを指定する，どのツールとどのツールが組み合わせ可能なのか事前に知っておく，またはツールに付属する説明文を熟読して調べる必要がある．そのため，ツール選択の改善を試みた研究もなされている [2][4].

対話的なクラスタリング環境を構築する際，ユーザにとって重要なことは，どのパネルにどのような (名前・機能の) ツールを組み合わせるかより，クラスタリングに必要なパラメータ（前処理段階）をいかに与えるか，また実行したいクラスタリング手法が選択できるかといった点である．したがって，TETDM 上に実装されているクラスタリング関連ツールを統合的に扱う，インタフェースの必要性が生じる．

4.2　管理パネルモデル

図 2 に，管理パネルモデルによるクラスタリング環境の概要を示す．本稿で提案する管理パネルモデルは，クラスタリングのために TETDM 上に仮想的な統合環境を構築するものである．パネルへマイニング処理ツールや可視化ツールを配置する仕様や，定義されているツール間のデータ連動に用いるメソッドやデータ型など，そのまま使用している．TETDM のコアとなるプログラム部分に手を加えることはない．

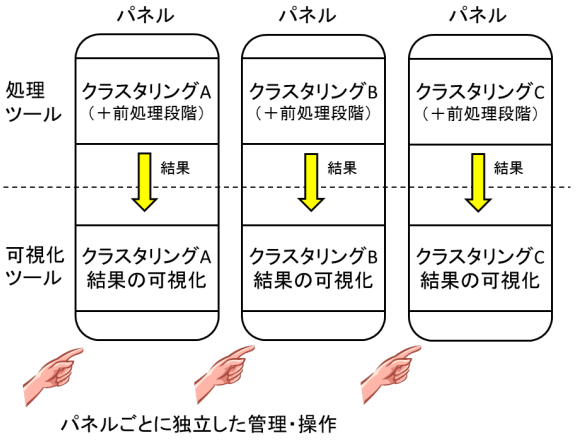

図 1: 基本方式によるクラスタリング．

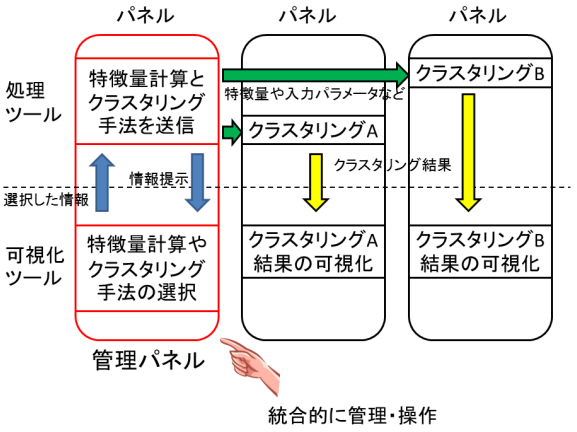

図 2: 管理パネル方式によるクラスタリング．

管理パネルは，TETDM の仕様に従い，マイニング処理ツールと可視化ツールが 1 対 1 対応して，TETDM のパネル上に配置される．管理パネルが行うことは，文字通り，クラスタリングに関連するツール群の管理であり，ユーザに提示する情報は，どのような前処理が可能か，どのようなクラスタリング手法が利用可能かの 2 点である．ユーザは，管理パネル上で，自身が利用したい前処理方法，クラスタリング手法を選択する．管理パネルのマイニング処理ツールは，ユーザが選択した前処理とクラスタリング処理の組み合わせに応じて，関連するマイニング処理ツールを動作させる．この際，前処理とクラスタリング処理の組み合わせを変えた，異なる処理を並列に実行可能である．

管理パネル以外のパネルには，管理パネルで選択した情報を引き継ぐマイニング処理モジュールを配置する．これにより，ユーザからは特定の前処理やクラス

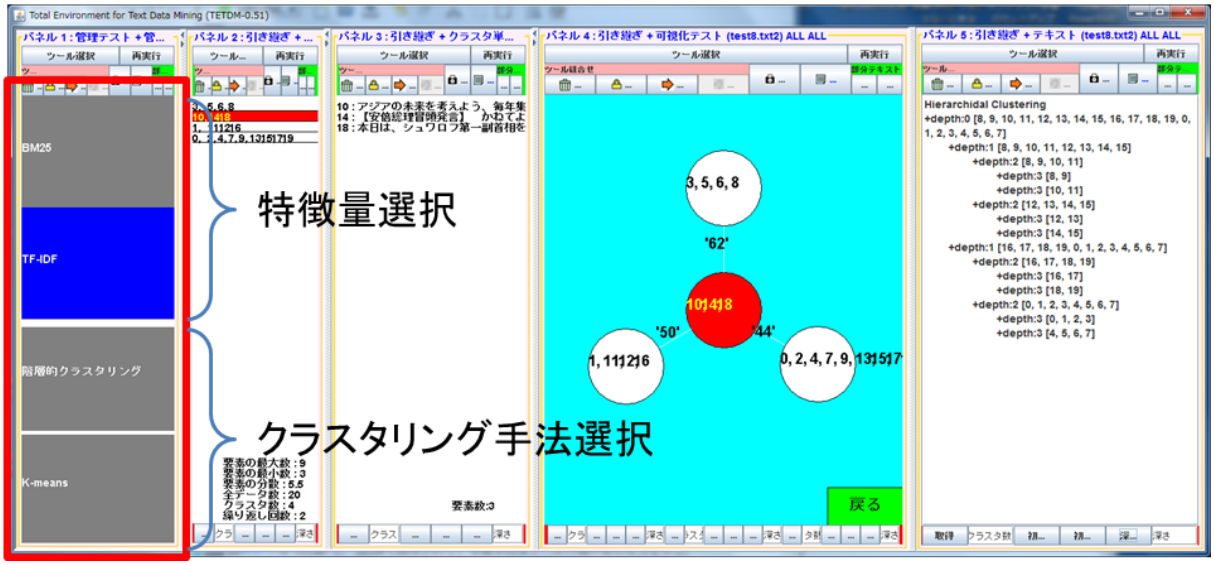

図 3: 管理パネルモデルに基づくクラスタリング環境.

タリング手法を個別に選択する指定する処理が隠されることになる．ユーザが管理パネル以外のパネルで自らの要求に応じて選択するのは，どのような情報が見たいか，すなわち可視化手法の選択だけである．

管理パネル方式では，ユーザは，パネルごとにマイニング処理ツールと可視化ツールの組み合わせに悩む必要がなくなる．ユーザは，管理パネル 1 つで複数のツールをまとめて取り扱うことができるようになった上で，クラスタリング結果を出力する可視化ツールを配置したパネルに現れる，ボタンや入力フォームなどを利用し，出力のフィルタリング，結果に対するフィードバックを行うこともできる．3.1 節で述べたオプションカテゴリのツールは，特定のクラスタリング手法との結びつきが強いマイニング処理ツールとなるため，そのクラスタリング手法の結果を出力するパネルからパラメータ入力を受け付ける形で実装することが望ましい．管理パネル方式を採用することで，TETDM の特徴でもあるパネルごとのインタラクション機能を維持したまま，管理パネルで複数のツールを統合する，汎用的な対話的クラスタリング環境が構築できることになる．基本方式ではパネルの独立性が高いのに対して，管理パネル方式では，複数のパネルにまたがって共通する処理を行うマイニング処理ツールを統合管理する．

5　クラスタリング環境の実装

3 節で述べた連動ルールと，4 節で述べた管理パネルモデルに従った，試験的なクラスタリング環境を TETDM 上に構築する．

図 3 に，クラスタリング環境の実行例を示す．図中の，左端のパネル（図中赤枠で囲った）が，管理パネルである．特徴量選択とクラスタリング手法の選択が可能である．現在のクラスタリング環境では，特徴量を TFIDF と BM25 から選択でき，クラスタリング手法を K-means と階層的クラスタリング（最近隣法）から選択できる．クラスタリング結果は，管理パネルの右側に並んでいるパネルに表示されており，テキストや図形を用いて結果を提示している．

6　おわりに

本稿では，TETDM を用いたシステム開発における，ツール同士の連携方針について検討した．対話的なクラスタリング環境の構築を目的としたシステム開発を想定し，ユーザの要求と TETDM の仕様に関する摺り合せを行った．データの流れに着目し，クラスタリングの段階ごとに必要なデータの型と，TETDM で定義されているツールの連動に関する仕様を対応させた．また，対話的なクラスタリング環境に適する，パラメータの指定やクラスタリング手法の選択を支援する，管理パネルモデルを提案した．本稿では，2 種類の特徴量計算手法の指定と，クラスタリング手法の指定が可能な試験的なシステムの実装を行った．

今後，ユーザによる制約の指定，ユーザからのフィードバックに応じた処理を可能にするほか，前処理のツール，クラスタリング手法のツールを増やし，クラスタリング環境を充実させる．

本稿で提案する枠組みにより構築する環境は，クラスタリングによる文書分析を行いたいユーザだけでなく，新規のクラスタリング手法や可視化手法などを他

の手法と比較して評価を行いたい研究者にも有益であると考える.

参考文献

[1] トム・デマルコ（著）, 高梨智弘, 黒田純一郎 (監訳) : 構造化分析とシステム仕様—目指すシステムを明確にするモデル化技法—, 日経 BP 出版センター (1994)

[2] 中垣内李菜, 川本佳代, 砂山渡 : 統合環境 TETDM を用いたテキストマイニングにおける初心者のためのツール選択支援, 第 27 回人工知能学会全国大会, 3B3-NFC-01a-1 (2013)

[3] 中村朋健, 上土井陽子, 若林真一, 吉田典可 : クラスタリング結果の特徴抽出を用いる高次元データの対話的クラスタリング, 情報処理学会論文誌 : データベース, Vol.47, No.SIG 19（TOD 32）, pp.28–41 (2006)

[4] 大塚直也, 松下光範 : テキスト分析における試行錯誤の支援に向けて—TETDM のインタフェースに関する一考察—, 第 2 回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, SIG-AM-02-10, pp. 56–61 (2012)

[5] 佐藤祐介, 岩山真 : 半教師有りクラスタリングを適用した対話型文書分類技術の提案, 情報処理学会研究報告, Vol. 2009-DBS-148, No. 7, pp.1–6 (2009)

[6] 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)

[7] 谷恵里香, 砂山渡 : 電子カルテにおける新人とベテランの特徴比較支援システム, 第 3 回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, SIG-AM-03-07, pp. 37–43 (2013)

[8] 徳永秀和:R によるテキストマイニング用 TETDM モジュール開発, 第 27 回人工知能学会全国大会, 3B3-NFC-01b-2 (2013)