

TETDM を用いたインタラクティブクラスタリングシステムの構築

Construction of Interactive Clustering System Using TETDM

北村 侑也^{*1}
Yuya Kitamura

高間 康史^{*1}
Yasufumi Takama

^{*1} 首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

This paper reports construction method of interactive clustering system using TETDM (Total Environment for Text Data Mining). TETDM has been proposed for developing various text mining systems. Although key concept of TETDM is that a user can combine various mining and visualization modules when analyzing text data, it is difficult to change combination of mining processing dynamically. This paper proposes a design pattern for TETDM, which enables dynamic combination of mining modules by introducing the concept of control panel. This paper shows the results of applying the proposed design pattern to develop an interactive clustering system.

1. はじめに

本稿ではテキストデータマイニング統合環境 TETDM (Total Environment for Text Data Mining) [砂山 14][砂山 13]を用いて構築したインタラクティブクラスタリングシステムについて報告する。現代ではビッグデータと呼ばれるような大規模で更新頻度の高い情報が現れ、これら进行处理する手法としてデータマイニングなどの計算機による支援が注目されている。特に、テキストデータは我々にとって最も身近な情報の一つであるため、多様なテキストマイニングシステムが開発されている。TETDM は、多様なテキストマイニングシステムの開発、普及を支援するために開発されているプラットフォームである。様々な開発者が作成したモジュールを組合せてマイニングを実行でき、複数のパネルに結果を出力可能であり拡張性に優れる。また、TETDM が統一したインタフェースを提供し、モジュール単位での機能変更が容易であるため、比較実験などの実施にも有効と考える。

しかし TETDM ではマイニング処理の組合せを動的に変更できないという課題があるため、本稿ではコントロールパネルの概念を取り入れた新しいデザインパターンを提案する。これにより、システム利用者はソースコードを変更することなく、マイニング処理の組合せを動的に変更することができる。

提案するデザインパターンを用いて、インタラクティブクラスタリングシステム[高間 13]を実装した。本稿では提案するデザインパターンを説明するとともに、このシステムを利用して行った比較実験の結果についても示す。

2. TETDM

TETDM はテキストデータマイニング環境の一つであり、多種多様なデータマイニングツールを統一的に扱うこと、ツールの開発・公開を支援することを目的としている。TETDM では入力テキストに対し形態素解析などの基本処理を施した結果を保持し、開発者はそれらを利用したモジュールを作成することができる。モジュールには TETDM のデータを利用して計算を行う処理モジュール、処理結果をパネルに表示する可視化モジュールの 2 種類があり、やりとりするデータが対応していれば動的に組み合わせを指定することができる。また、パネル間でも固有のメソッドを使うことで、連動を要請できる。しかし、処理モジュールの組

み合わせを変更したい場合にはソースコードを修正・追加する必要があり、動的な変更を行えないという問題点がある。

3. TETDM を用いたインタラクティブクラスタリングシステムの実装

3.1 デザインパターン

本稿で提案するデザインパターンを採用した、インタラクティブクラスタリングシステム[高間 13]の構成を図 1 に示す。クラスタリングシステムは、特徴量計算やクラスタリングアルゴリズムなどの処理から構成され、各処理について様々なアルゴリズムが提案されている。提案するデザインパターンではコントロールパネルがモジュール間のデータのやりとりなどを担当することで、利用するアルゴリズム(モジュール)を動的に変更可能としている。システムの処理の流れを以下に示す。

1. コントロールパネルで特徴量モジュールとクラスタリングモジュールを一つずつ選択し、特徴量モジュールに処理命令を送信する。
2. 特徴量モジュールが特徴量を計算し、コントロールパネルに渡す。
3. コントロールパネルはクラスタリングモジュールに特徴量など必要なデータとともに処理実行命令を送信する。
4. クラスタリングモジュールでは共通データセットを参照してクラスタリングを行い、結果をコントロールパネルに返す。
5. コントロールパネルは結果を受信モジュールに送る。
6. 受信モジュールは受け取ったデータを元に必要な処理を行った後、可視化モジュールにデータを渡す。
7. 可視化モジュールで結果を表示する。

実際にパネルにセットするのはコントロールパネルに関するモジュール、各受信モジュール及び可視化モジュールであり、各処理モジュールは実際にはセットせず、コントロールパネルから呼び出してデータをやりとりする。また共通データセットとしては、TETDM の全モジュールがアクセス可能な int 型の配列を用い、処理モジュールの ID や各種設定を格納する。

コントロールパネルに登録したい処理モジュールについてはソースコードと同じフォルダ内に、1 行目に処理のタイプ、2 行目にモジュール ID を記したテキストファイルを置いておく。コン

トロールパネルではこのファイルを実行時に読み込むことで、ソースコードを変更することなく、各手法を選択し利用できる。

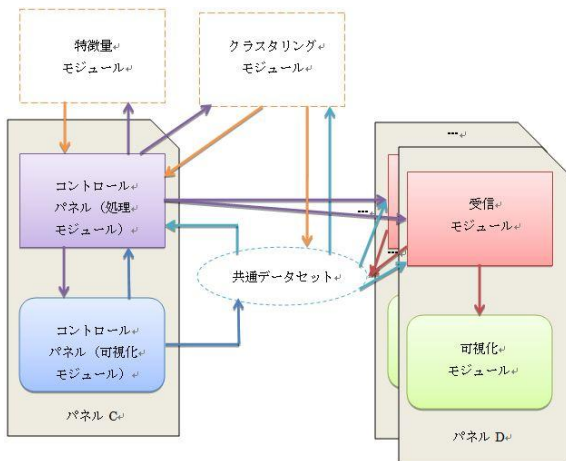


図 1: インタラクティブクラスタリングシステムの構成

3.2 インタラクティブクラスタリングシステム

TETDM を用いて実装したインタラクティブクラスタリングシステムのスクリーンショットを図 2 に示す。一番左のパネルがコントロールパネルであり、ユーザはここで特徴量モジュールとマイニングモジュールを選ぶ。その右隣から右に向かって順に、クラスタ表示用、クラスタ内の文書一覧表示用、単一文書の詳細表示用パネルである。ユーザはこれらを用いて文書の詳細を参考にしつつ、クラスタ表示用パネルでグルーピング操作を行い、制約を付与する。その後、制約を考慮して再クラスタリングを行い、新たな結果を表示する。なお、次節で述べる比較実験のために、オブジェクト対に対し、must-link を逐次追加する手法(逐次指定)も同様に実装している。この場合、制約は文書一覧表示用パネルで付与する。

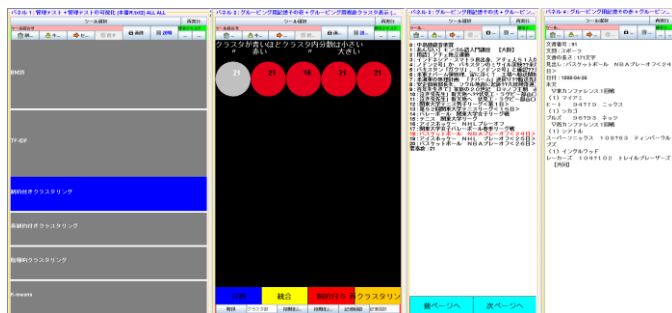


図 2: インタラクティブクラスタリングシステムのスクリーンショット

4. 評価実験

実装したシステムを用いて、制約一括生成手法と逐次指定手法の比較実験を行った。工学系大学・大学院生 20 名に協力してもらい、新聞記事をジャンルごとに分類する実験を行った。NTCIR プロジェクトで提供されている毎日新聞記事データから、芸術、家庭、経済、国際、スポーツの 5 ジャンルについて各 20 件、合計 100 件のデータセットを作成し、制約付きクラスタリングを反復的に実行してもらった。制限時間 30 分以内に、クラスタリング実行後の状態が(1)各ジャンルについて、過半数の記事が 1 クラスタに集中、(2)各クラスタについて、特定ジャンルの記事が過半数を占める、の両方を満たす場合にクラスタリング成功と

判断した。協力者を 10 名ずつの 2 グループに分け、それぞれ制約一括生成手法、逐次指定手法を用いて作業を行ってもらった。

一人あたり 3 回の試行を行ってもらった結果、合計 30 回の試行の内、成功回数は両手法とも 21 回であった。成功した場合の両手法の比較結果について表 1 に示す。作業時間が 30 分を超える場合もあったが、その場合は操作方法に関する質問があったためである。表ではそのような質問・回答に要した時間も含めている。また、制約組とは連結成分となる制約の集合を意味する。

表 1 制約付与手法に関する実験結果の比較

	一括生成	逐次指定
平均制約対数 (標準偏差)	60.14(3.51)	56.3(30.42)
平均制約組数 (標準偏差)	14.14(2.44)	8.7(5.58)
平均クラスタリング回数 (標準偏差)	3.47(2.63)	11.52(8.72)
平均作業時間	19'52"	13'4"

付与された制約対数について、平均では両手法にあまり差が見られないが、標準偏差では逐次指定手法が一括生成の約 10 倍となっている。これより、逐次指定手法では協力者によっては制約を過剰に付与するケースがあることがわかる。制約組については一括生成手法の方が多くなっており、逐次生成手法よりも大局的な視点から制約が付与される傾向にあると考える。

表 1 より、一括生成手法の方がクラスタリング回数は少ないものの、作業時間は長い傾向にあることがわかる。一括生成手法ではまとめて制約が作られるため、全ジャンルについてグルーピング操作を行ってから再クラスタリングを行う傾向が観察された。これに対し逐次指定手法では、少数の制約を付加した後、その有効性を確かめるために再クラスタリングを頻繁に行う傾向が観察された。これらの傾向が作業時間やクラスタリング回数の違いに影響しており、大規模データセットの場合にはクラスタリング回数が少なく済む一括生成手法の方が有利と考える。

5. 結論

本稿では、処理モジュールの組合せを動的に変更可能とする、TETDM の新たなデザインパターンを提案し、これを用いてインタラクティブクラスタリングシステムを実装した。また、構築したシステムを用いて制約付与方法に関する比較実験を行った結果についても示した。提案するデザインパターンは、TETDM を利用した多様なシステム開発への貢献が期待できる。

参考文献

- [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)
- [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)
- [高間 13] 高間, 三宅: グルーピング操作に基づくインタラクティブなクラスタリング対制約生成手法の考察, 第 27 回人工知能学会全国大会, 2F4-OS-04-3(2013)