テキストマイニングツール TTM（TinyTextMiner）の理念と使い方

松村真宏[1]（大阪大学）　三浦麻子（関西学院大学）　金明哲（同志社大学）

概要：テキストデータのような定性データは、大量のデータを分析することで安定した傾向を見いだせるが、人手で大量のテキストデータを分析することは現実的には難しい。しかし、テキストマイニングのツールの登場によって、大量のデータを統一的な視点・基準から少ない労力で分析することが可能になった。本発表では、「テキストマイニングを、“分かりやすく”、そして“タダ”で、行うことを可能にする」というコンセプトの元で開発したフリーソフト TTM（TinyTextMiner[2]）について述べる。

テキストマイニングは、テキスト情報から人々のニーズや不満を定量的に把握する手段として近年注目を集めている技術である。例えばインターネット上のクチコミ情報源から、製品やサービスに関する評判情報や、性別・年代・職業・家族構成といったユーザ属性を推定する手法が、製品開発やマーケティングの効果測定などに利用されている。また、質問紙調査の自由記述設問の分析にも適した手法である。

テキストマイニングは、テキストデータ特有の処理からなる「前処理」と、集計データに対する処理からなる「後処理」に大別できる。前処理では自然言語処理（形態素解析、構文解析など）によって語を切り出して集計データを作成し、後処理では統計解析（多変量解析、仮説検定など）やデータマイニング（分類器、予測器など）によって集計データを処理する。後処理については様々なソフトウェア（例えば SPSS や R など）が利用可能だが、前処理についての敷居が高いことがテキストマイニングに取り組むうえでの最大の障壁となっていた。

そこで、本稿の筆者のうち松村と三浦はテキストマイニングの前処理に特化したフリーソフトウェア TTM （TinyTextMiner）を開発し、TTM を利用したテキストマイニングの入門書を上梓した[3]。TTM はテキストデータを形態素解析器、構文解析器にかけて、その分析結果を読み込んで集計し、CSV ファイルを出力するシンプルなソフトウェアであるが、GUI、「タグ付きテキスト」の読み込み、キーワード・同義語・不要語の指定、品詞指定・閾値の設定、係り受け解析、英語テキスト対応、6 種類の出力ファイルといった機能を備えており、多様な分析目的に対応できるテキストマイニングの前処理を実現する。TTM の出力ファイルだけでもテキストマイニングの結果として利用できるが、統計解析ソフトウェアを併用することで、より高度なテキストマイニングが実現できる。

---

[1] 連絡先 matumura@econ.osaka-u.ac.jp

[2] TTM は http://mtmr.jp/ttm/ からダウンロード可能

[3]『人文・社会科学のためのテキストマイニング』（松村真宏・三浦麻子著、誠信書房 2009）