

テキストマイニングの世界へようこそ

～ TMS for VMStudio と Text Mining Studio ～

株式会社 数理システム データマイニング室 岩本 圭介

1 はじめに

テキストマイニングとは、テキストを分析して意味ある情報や知識を発見するプロセスであると位置づけられている。しかし、人手でテキストを読解して印象を掴んだり主張を取り出したりするという行為は非常に客観性を欠くものであり、データの多さゆえ大抵の場合は多大な労力を要するものである。そのためテキストデータを自動的にマイニングの対象となる形式へと変換するツールが必要であり、TMS for VMStudio はそういった要請に応えるべく開発された。

TMS for VMStudio は汎用データマイニングツール Visual Mining Studio (VMStudio) のアドオンモジュールとして提供され、2004年11月にバージョン1.1がリリースされた。これはテキストデータ、すなわち日本語の自然文を自動的に単語単位に分割してVMStudioで分析可能な形式へと変換する機能を持つ。これにより、テキストを含むデータ対してもVMStudioの諸機能を用いて解析を行うことが可能となる。

アドオンモジュール TMS for VMStudio を用いて本格的な解析を行う場合はVMStudio本体との連携が必須であるが、現在、テキストマイニング機能に特化したパッケージソフトウェア Text Mining Studio の開発が進められており、これは分析結果がボタン一つで得られるような使い勝手の良さを備えたものとなる。

したがって、我々の提案するテキストマイニングのソリューションには次のものがある。

- ・ 汎用データマイニングツール Visual Mining Studio + テキストマイニングアドオン TMS for VMStudio
- ・ テキストマイニングツール Text Mining Studio (2004年末リリース予定)

本稿では TMS for VMStudio の機能・特徴を紹介する。また、新製品 Text Mining Studio の特色についても述べる。

2 TMS for VMStudio

2.1 TMS for VMStudio の特徴

TMS for VMStudio は、次の特徴を備えている。

単語連結モードを備えた分かち書き機能

文章の分割、すなわち分かち書きを行う際、一般の形態素解析ツールを用いて単純に単語単位で分割すると、その分割単位が細かすぎるために、「使える」データにするためのクリーニング作業の手間が膨大になる。しかし、単語連結モードの適用により文節単位で文章をとらえることが可能となり、さらに助詞や記号等のストップワードが自動的に削除されるので、有益な結果を迅速に得ることができる(表1)。また、高精度の構文解析機能により分かち書きと同時に単語間の修飾関係(係り-受け)を抽出することができる。

見出し語	原形	品詞
買った	買う	動詞
て	て	助詞
から	から	助詞
10	10	名詞
ヶ月	ヶ月	名詞
経つ	経つ	動詞
て	てる	動詞
なかつ	ない	助動詞
た	た	助動詞
ので	ので	助詞

単語連結モードを使わない

見出し語	原形	品詞	述語属性	関係子
買ってから	買う	動詞	なし	状況
10ヶ月	10ヶ月	名詞	なし	状況
経ってなかった	経つ	名詞	neg	理由

単語連結モードを適用

「述語属性」「関係子」といった情報が付加され
記述のニュアンスを保持することができる

表1. 「買ってから10ヶ月経ってなかった」という文章を分かち書きした場合の比較

使い勝手を考えた 3 種類の辞書

未知の単語を登録するユーザ辞書、同一視したい言葉を登録する類義語辞書に加え、単語連結モードの結果をコントロールすることができる分割辞書の計 3 種類の辞書を搭載しており、それらへの登録は解析結果のテーブルからマウス操作で簡単に行うことができる。また、未知語を自動抽出する登録支援機能を搭載している。

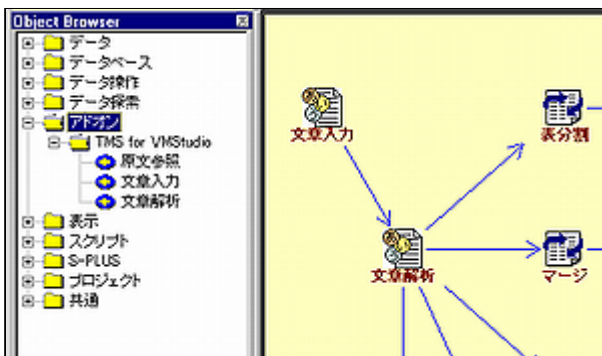
単語・係り受け頻度解析機能

単語や係り受けの出現頻度表をボタン 1 つで表示することができる。これによりテキスト全体のイメージを把握する助けとすることができる。頻度表は各種フィルタ設定を備えている。

VMStudio の持つマイニング機能の利用

解析結果に対して、VMStudio が備えるあらゆるデータ解析機能を適用することができる。

2.2 TMS for VMStudio によるマイニングの流れ



TMS for VMStudio をインストールすると、VMStudio のオブジェクトブラウザに「TMS for VMStudio」フォルダが作成され、その中に

- 文章入力
- 文章解析
- 原文参照

の 3 種類のアイコンが追加される。これらを VMStudio のプロジェクトボード上に配置することで (図1)、次のような手順をもってマイニング作業を行うことができる。

図1. TMS for VMStudio アイコン

- 『文章入力』アイコン で データ属性を指定してファイルをインポートする
- ↓
- 『文章解析』アイコン で 文章を文節・単語単位に分割する
また、頻度解析を行って文章の大まかな傾向を掴む
- ↓
- 文節・単語単位に分割したデータを元に、VMStudio のマイニング機能を用いて解析を行う
- ↓
- 『原文参照』アイコン で 解析の結果得られたキーワードの原文中での表現を確認する

『原文参照』アイコンは、バージョン 1.1 より追加され、キーワードやデータ属性による条件の指定を受け、該当する原文を検索して一覧表示を行う機能を提供する。

2.3 ファイルのインポート：文章入力アイコン



『文章入力』アイコンの処理画面で入力ファイルの指定を行う (図2)。入力ファイルの形式として想定するのは区切り文字で区切られた表形式のファイル、もしくはフォーマットを特に持たないテキストファイルである。

ここで、テキスト列以外のデータはテキストに付随する「属性」とであるとみなされ、このような属性の違いを考慮した解析を行うことも可能である。図2 は、アンケート結果のテキストに「年齢」と「性別」の 2 つの属性が存在している例であり、ダイアログ下部の『プレビュー』欄にその様子が示されている。

図2. 文章入力画面

2.3 分かち書きと頻度解析：文章解析アイコン



図3. 文章解析画面

『文章解析』アイコンではテキストの分かち書き・辞書の登録と管理・頻度表の作成といった主要な処理を行う(図3)。分かち書きされたテキストと作成された単語・係り受け頻度表はテーブルエリアに表示され、これらは文章解析アイコンの出力テーブルとなり VMStudio の他の解析機能で用いることができる。

2.4 原文参照アイコン



図4. 原文参照画面

『原文参照』アイコンを用いることにより、指定したキーワードを含んでいる原文をテーブル形式・レポート形式で表示させることができる。また、データ属性によって結果のフィルタリングを行うことができる。

解析の結果得られたキーワード群について、それがどのような形・どのような文脈でオリジナルデータの中に登場していたのか確認する際に威力を発揮する。

