テキスト分析をリードする **TMStudio** 3 つの進化

～ 予兆発見、自動分類、そして WEB ソリューション～

株式会社 数理システム 岩本 圭介

1. はじめに

(株)数理システム のテキストマイニングツール Text Mining Studio （TMStudio） は、お蔭様で現在 240 サイトを超えるユーザ様にご利用いただいている製品となりました。その利用の分野も大きな広がりを見せており、使い勝手 と 自由度の高い分析 とを両立させるという開発コンセプトを受け入れていただいた結果であると感謝しております。

2012 年 1 月リリース予定の TMStudio 次期バージョン 4.1 では、特に大きな機能追加項目として、次の 3 つを実現いたします。

- **予兆発見 … 時系列分析機能を強化！**
  出現頻度の上昇・下降といった傾向を把握し、急変動しているものを自動的に抽出することで、テキストからの予兆発見が可能となります。
- **テキスト自動分類 … テキストの自動分類機能を全面刷新！**
  全テキストの分類状況を一覧できる GUI を新規に搭載します。また、意味的なまとまりに着目して分類を行うクラスタリングの新手法を利用可能にします。
- **Web ソリューション … Web 公開機能対応！**
  データマイニングツール Visual Mining Studio (VMStudio) の、分析プロジェクト WEB 公開機能に対応します。VMStudio と同時にご利用いただくことで、WEB から TMStudio の機能が手軽に利用可能になります。

本稿では、これら 3 つの新機能を中心に、TMStudio 次期バージョン 4.1 の新機能について解説いたします。

2. 予兆発見

日時の情報が付随したテキストデータから、時間軸に沿ってことばの出現頻度がどのような傾向を示しているか、上昇傾向・下降傾向といったトレンドを把握することは非常に重要です。例えば、新製品のニュースリリースに伴ってその製品に関する口コミがどのように波及していったかを WEB 上のテキストから把握する、また蓄積された特許文書から技術的な動向がどのように推移しているのかを理解する、といった応用の場面があります。

TMStudio 現バージョン 4.0.1 には、時系列分析の機能として、特に頻度の高いことばの時系列推移を図示すること、また全体でのバラつきが大きいことばを抽出することが可能ですが、新バージョン 4.1 では時系列分析の機能を強化し、これらに加えて、時系列に沿って上昇傾向にあることば、下降傾向にあることばといったものを明示的に抽出できるようになります。

また、上昇傾向にあることばの中でも、その上昇度合が激しいことば、また最近出現回数が上昇してきたことばについては特に注意が必要です。例えば、コールセンターにおいてユーザからの問合せが日々蓄積されているような場面においては、至近において急に発生してきた話題をとらえるということは、新規クレームの予兆発見・早期把握を意味し、これは極めて有用な情報となり得ます。

新バージョン 4.1 では、出現頻度の上昇・下降傾向をとらえる際に「それがどの程度の変動だったか」「その変動がいつ起こったか」というポイントをとらえ、ことば毎の時系列データに対して、最近起こった変動でそれが急であるほど高い値を与えるような指標値付けを行います。一般にテキストに含まれることばの数は非常に多数であり、時系列に沿ったことばの頻度の推移データもその数だけ存在することになりますが、この「指標の値が高いことば」をその指標値の順に図示することにより、着目すべき時系列変動をしていることばを一目で明らかにします（図 1）。

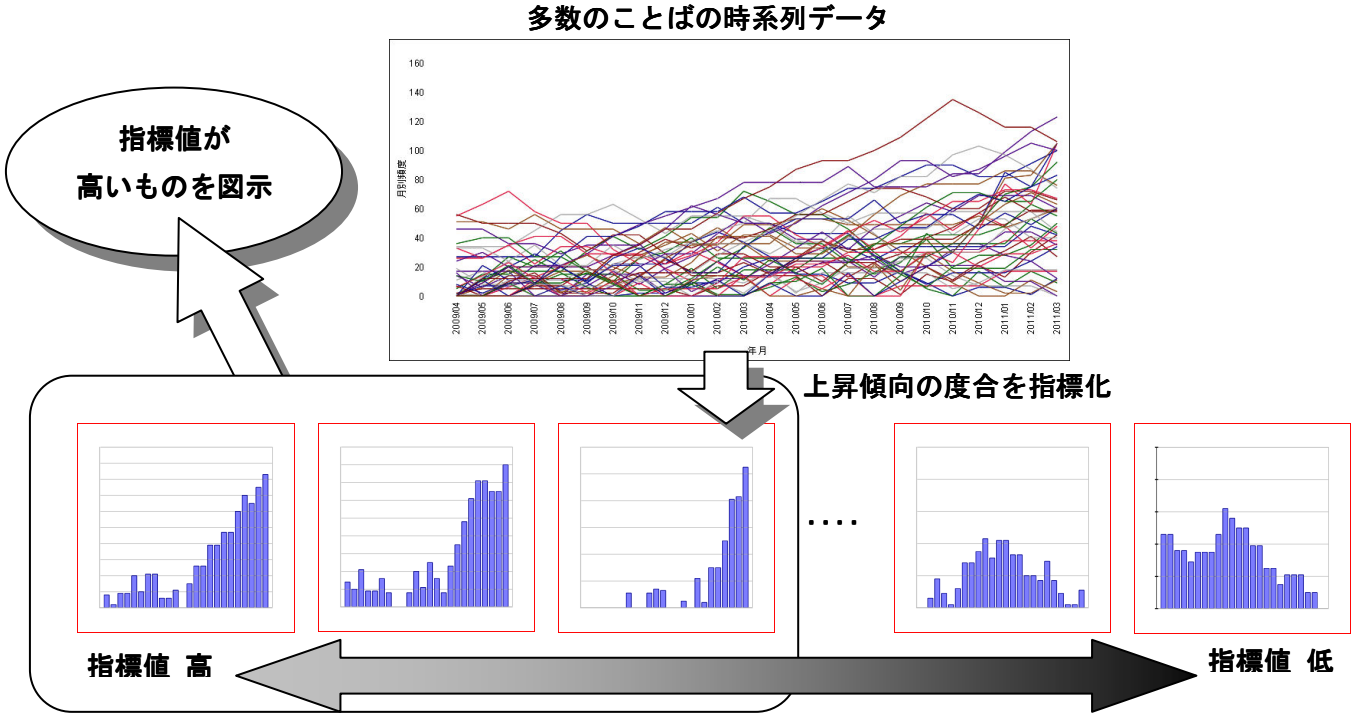

図 1 上昇傾向の度合による 指標値 を用いた抽出

3. テキスト自動分類

テキストから自動的に類似したもの同士のグループを作成すること、すなわちテキストをクラスタリングすることは、まず内容把握の大きな助けになります。どんな話題が出現しているか、またそれぞれの話題においてどんなことばがキーワードになっているか、といった点を見ることにより、意見・発言の固まりを把握することができます。

新バージョン 4.1 では、クラスタリング機能を提供する分析「文章分類」のユーザインタフェースを刷新いたします。まずテキスト全体のクラスタリング状況を**サマリ画面**によって確認し、ここからどのデータがどのクラスタに分割されたか、というオリジナルのデータの情報をドリルダウンによって即座に表示させることが可能となります（図 2)。これにより、スムーズにクラスタリング結果を把握することができます。

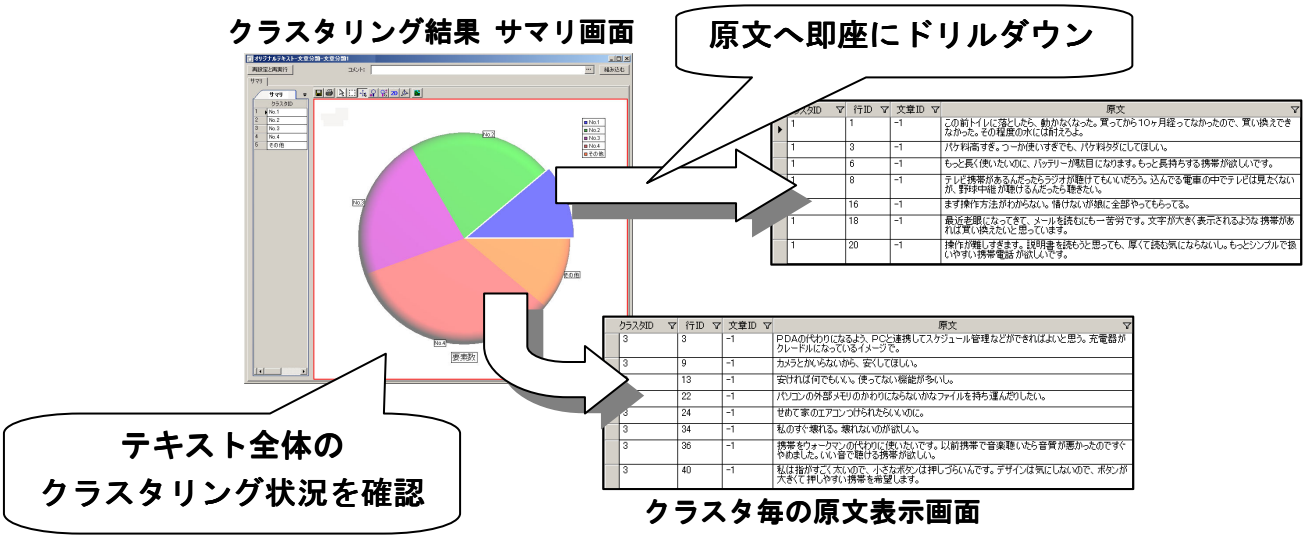

図 2 サマリ画面 から 対応する 元データ の確認

また、クラスタリングの手法として、文章間の頻出する共通単語を元にしてクラスタをまとめ上げる従来手法に加えて、話題の間の背景で出現することばの存在も考慮し、意味的なまとまりをより強く抽出する新手法を利用可能にします。これにより、例えば一見製品 A と製品 B との問合せがそれぞれ異なるジャンルとして存在しているように見えるものの、クラスタリング新手法を用いると実は両者は使われる状況が似ているためにまとめ上げられるなど、意外な発見を導くための材料としてクラスタリングの結果を利用できます。

更に、成功したクラスタリングの結果を用いて新規のデータを判別し、それらをクラスタに振り分けることも可能となります。新規データの傾向把握、また既存データとそれがどの程度かけ離れているか、といった情報を読み取ることができます。

4. WEB ソリューション

数理システムの データ解析・マイニング・最適化・シミュレーション といった各種ソリューションを連携させ、より実りのある結果を導くべく、プラットフォーム Visual Analytics Platfrom (VAP) 上で数理システムの各種ツールが統合的に利用可能になります。この VAP とデータマイニングツール VMStudio は密接な関係にあります。VAP には WEB ベースの分析プラットフォームとしての側面もあり、VMStudio で作成した分析フローを、VAP 上で動く WEB アプリケーションとして公開することができるようになっています。

更に、TMStudio は、VMStudio とシームレスに連携し VMStudio の一機能として TMStudio 自体をそのまま利用できるようになっています。次期バージョン 4.1 より、TMStudio がこの VMStudio の分析 WEB 公開機能に対応いたします。これによって、VMStudio と TMStudio を同時に用いることで、TMStudio の分析機能を WEB アプリケーションとして手軽に利用することができるようになります（図 3)。

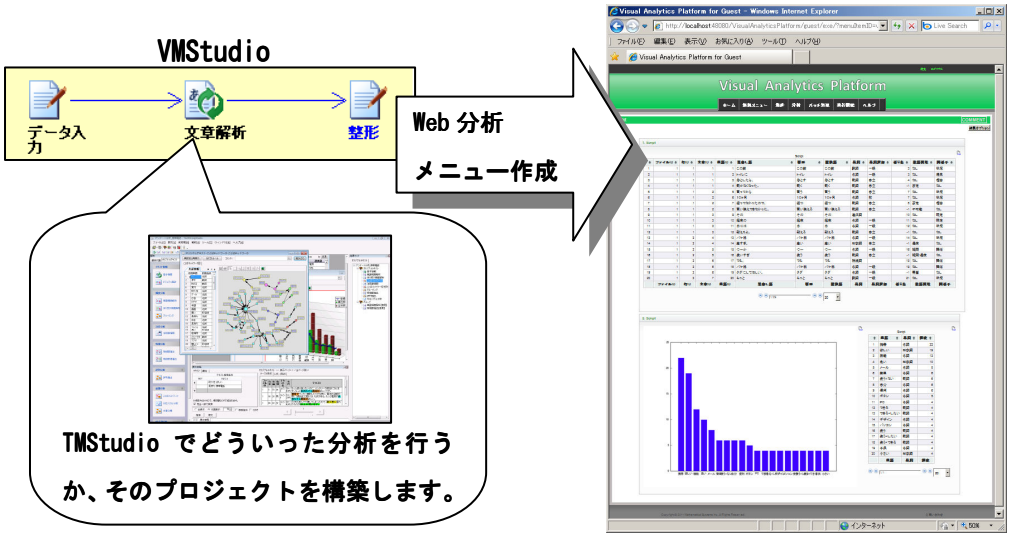

図 3 VAP からの TMStudio の利用

5. 今後の Text Mining Studio

TMStudio の次期バージョン 4.1 は、2011 年 1 月にリリース予定です。本稿の内容のほか、カテゴリチェック表出力対応等のバッチ実行時の利便性向上、分割辞書の分割自由度向上、属性加工のユーザビリティ向上、その他不具合対応が盛り込まれます。

今後は、より一層手軽なテキストマイニング利用の形態をご提供させていただくべく、分析メニューを厳選・集約しダイレクトに有意義なアウトプットに繋げられるようなツール Text Mining eXpress（仮称）を、VAP 上の WEB アプリケーションとして提供させていただく予定です。

今後の 数理システム のテキストマイニングソリューションに、是非ご期待ください。