

VMStudio におけるテキストマイニング

(株) 数理システム
科学技術部 齋藤宗香

2003 年 11 月

1 テキストマイニングについて

テキストマイニングについて昨今関心が高まっており、雑誌または Web のビジネスサイトなどで記事に取り上げられることが多くなってきました。テキストマイニングの有用性や用途についての解説はこれらの記事に譲るとしまして、ここではテキスト形式のデータを、どのようにして分析可能な形式にする変換するかについて考えていきます。

これまでもテキストデータを対象としたマイニングは行われていましたが、マイニングの前段階としてすべてのテキストに目を通して単語をピックアップしてからデータマイニングにかけるという、非常に労力を要するものでありました。この作業を終えたとしても、必ずしもそれが有用な結果をもたらしてくれるわけではありません。納得できるデータが得られるまで(もしかしたら得られないかもしれませんが)、この地道な作業を続けることとなります。データマイニングを行うアプリケーションは多く存在しているのですが、テキストデータをデータマイニングで解析する前までの手段が抜け落ちているです。

「テキストマイニング for VMStudio(仮称)」では、テキスト形式データを VMStudio で分析可能なデータ形式に変換するための機能を有しています。以降、テキストデータを VMStudio で分析するまでの手順と、この製品の機能・特徴について紹介します。

2 テキストマイニングの具体的な処理

2.1 テキストデータのテーブル化

テキストデータをデータマイニングツールで扱うには、テーブル化を行う必要があります。テーブル化とは、例えば次の文

我輩は猫である。名前はまだない。

をテーブル化すると以下ようになります。

ファイル ID	文章 ID	単語 ID	単語名	品詞
1	1	1	我輩	名詞
1	1	2	は	助詞
1	1	3	猫	名詞
1	1	4	で	助動詞
1	1	5	ある	助動詞
1	2	6	名前	名詞
1	2	7	は	助詞
1	2	8	まだ	副詞
1	2	9	ない	形容詞

表 1. テーブル化された文章の例

対象となるデータ量が不十分なので意味のある結果は得られませんが、このような形式であれば、データマイニングツール上での分析が可能です。

2.2 テキストマイニングの手順

テキストマイニングツールと呼ばれるものでは一般に、次のような処理を行います。

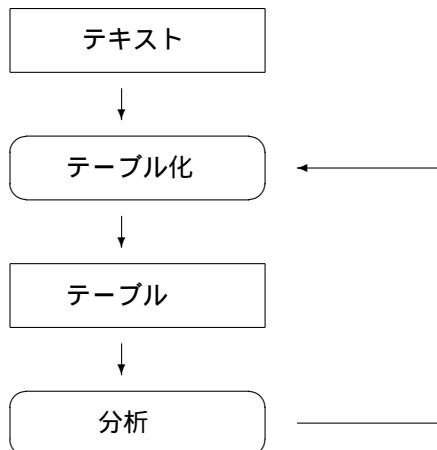


図 1. テキストマイニング処理の流れ

[テーブル化]

まず文章を単語に分けます。この処理を「形態素解析」といいます。英語などでは空白によって単語がはっきりと分割できますが、日本語では文の構成を理解した上でないと単語の分割ができません。そこで、日本語の形態素解析を行うツールが有償・無償を含めていくつか存在しますので、これらを使用します。これによって、先に挙げたようなテーブルを作成することができます。

[分析]

テーブル化したデータをデータマイニングアプリケーションで分析します。

[再度テーブル化]

テーブル化の処理において、即座に適切なテーブルができるわけではありません。テキストに新語などが多く形態素解析ツールでうまく分解できなかつたり、分析の対象となる単語を変更したくなつたりします。テキストマイニングでは、テーブルの作成 分析を何度も繰り返すことにより徐々に適切なテーブルを作り上げていきます。

2.3 テーブルの分析

テーブル化後のデータに対しては、以下の分析手法が利用できます。

- アソシエーション分析 (時系列分析・関連性ダイアグラムなど)
単語の共起の度合を算出します。
- 対応分析 (S-PLUS corresp 関数)
単語同士の類似度、あるいは単語に対するイメージを図示します。

3 テキストマイニングで使用するテキストの特徴

テキストマイニングで使用するテキストとして

- アンケートの自由記述欄の内容
- WEB の文書

などがあります。これらのテキストを形態素解析ツールを実行しても、すぐに期待通りの単語分解できるわけではありません。

これらのテキストには、若者言葉・略語など辞書に掲載されていない単語が多く含まれます。また、製品アンケートなどのテキストとなると、製品名・型番などが頻繁に出現することになります。形態素解析ツールでは独自に辞書を持っていますが、これらの単語をすべて収録することは事実上不可能です。収録されていない単語が多いと、形態素解析ツールは文章を単語単位に正しく分解できなくなります。

また、テキスト形式データにおいては書き手の自由度が大きいため、1つのものを意味するのに複数の単語が使われます。例えば、「パーソナルコンピュータ」は「PC」「パソコン」など、書き手によって様々です。これだけに留まらず、全角・半角の違い、最後の長音を入れるかななどにより、さらにバリエーションが増えてしまいます。分析する立場としては、これらの単語を同一に扱いたいところです。

4 VMStudio におけるテキストマイニング

テキストマイニングツールは VMStudio のアドオンの 1 つとして利用できます。

VMStudio にはユーザの用途に応じて、様々な分析手法が用意されております。テキストマイニングアドオンでは自由記述テキストデータからテーブルを作成する工程を担当し、分析については VMStudio にある手法を使用します。

また、テキスト形式のデータの特徴を考慮に入れ、テーブル化に十分な機能を実装しています。GUI の操作方法についても、ユーザに無駄なストレスがかかることのない GUI 設計を心掛けております。

4.1 入力

1 つまたは複数の自由形式テキストファイル¹を指定できます。

4.2 テキストマイニングアドオンの機能

テキストマイニングアドオンは、テキストデータからテーブルを作成する機能を有しています。

- chasen

形態素解析ツールとして chasen を使用しています。chasen についている辞書では不十分であるので、独自に辞書の拡充を施しております。

- 辞書

単語辞書機能として以下のものがあります。

[ユーザ辞書の作成]

chasen の出力結果で '未知語' と判定された単語が多いような文章では、形態素解析もうまくできておりません。'未知語' とは chasen の辞書に含まれていない単語で、これらの単語を辞書に登録し再度 chasen を実行することで形態素解析の精度を高めます。

[類義語 (単語グループ) の設定]

表記はことなるが同じ意味を持っている単語があります。またテキスト形式データでは誤字・脱字もあるでしょう。そのような単語をグループ化し、1 つの単語に置き換えます。

[抽出語・削除語の指定]

着目したい単語、または必要のない単語を指定し、それぞれ抽出・削除を行います。

¹現時点では、文字コードが Shift JIS であることを前提としています。他の文字コードについては今後対応を予定しております。

- 操作性のよいユーザインターフェース

テキストデータから分析可能なテーブルを作成するまでには、多くの単語の操作を必要とします。分析以上にこの過程に時間と手間がかかりますが、有効な結果を得るために最も重要な過程でもあります。

開発においては、この点を最大限に考慮に入れてユーザインターフェースの設計を行っております。ほとんどの操作をマウスのみで行うことができ、ストレスのない操作が可能です。

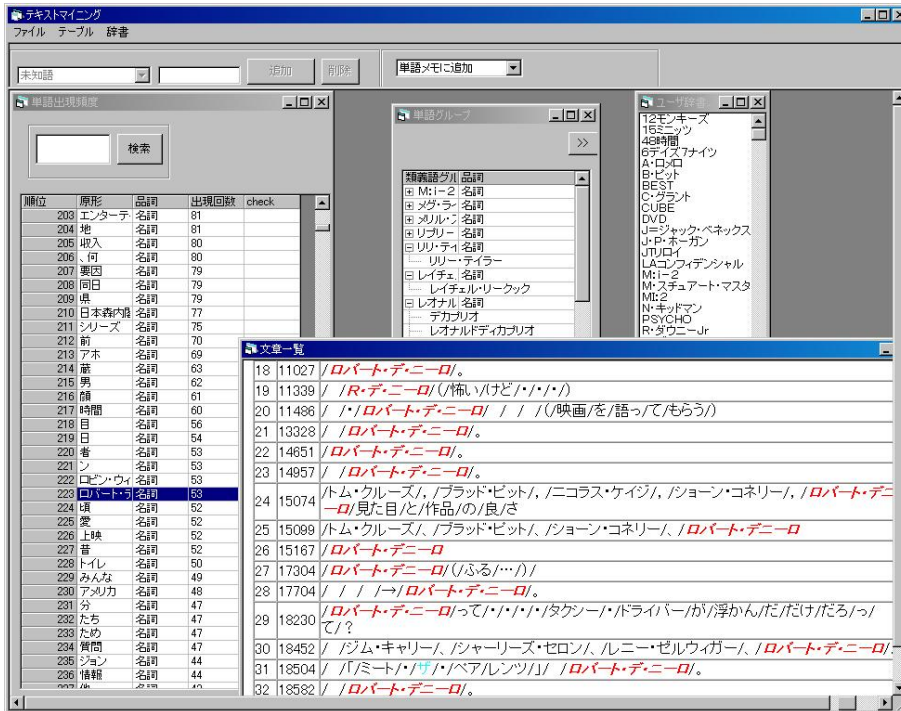


図 2. テキストマイニングアドオン GUI サンプル

5 今後の展開

VMStudio テキストマイニングアドオンの機能として、今後も機能の追加を検討しております。以下、現在実装を検討している機能の一部です。

- 単語より大きな単位の語句の登録
- 断定、否定、疑問、要望の判定
- ファイル・文章の分類

皆様からのご意見・ご希望等がございましたら、ぜひお聞かせ下さい。今後の開発の参考とさせていただきます。