【第二版】

（株）NTT データ数理システム

Text Mining Studio グループ

目次

続きは読本を発送させていただきますので巻末の「(※) 当読本の発送をご希望の方は」をご覧ください。

1. テキストマイニングとは？

ここ数年来、「テキストマイニング」という言葉を頻繁に耳にするようになってきました。何かの役に立つようだけど、一体どういった使い方ができるものなのでしょう。そもそも「テキストマイニング」とは何？

「テキストマイニング」というからには、テキスト（文章）から何かをマイニング（発掘）するということは確かでしょう。ただ、よく似た「データマイニング」という言葉もあります。本章では、テキストマイニングとデータマイニングの関わり、そしてテキストマイニングで何ができるか、その守備範囲について解説します。

1.1.データマイニングって？

既に定着している「データマイニング」。データマイニングとは、大量のデータから役に立つ有益な情報を抽出するプロセス、そしてそのための技術を指します。

まず、データを集計し、それを様々な形式で可視化・グラフ化するという手順によりデータの傾向を把握するというステップが第一にあります。この集計は、ある程度のデータ量までであれば Microsoft Excel などを使うことで実現できますが、これはあくまでデータマイニングの「入口」で、単なる集計を超えるための分析手法をデータマイニングは秘めています。データマイニングの目指すところは、

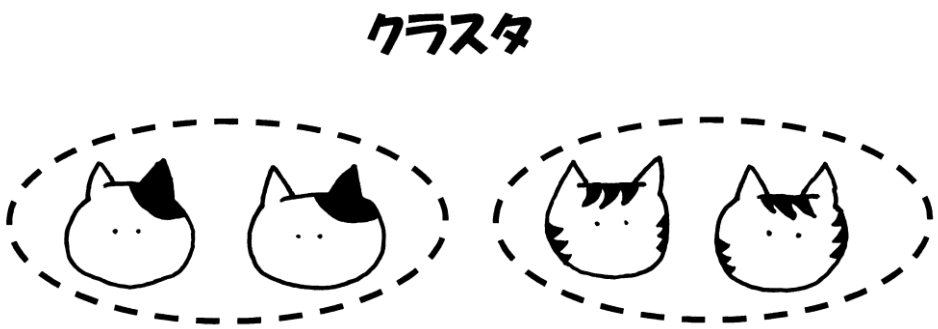
大量のデータから、まだ知られていない知識を、
前提条件のないところから獲得する

ことです。後程、「知識」とは具体的には何か、という点について**エラー! 参照元が見つかりません。**章で掘り下げていきますが、ここでは「前提条件を置かない」ということが一つのポイントです。一般的な従来型の統計解析手法では、データがどのように散らばっているか、例えば「正規分布を

仮定する」などとしてその分布を仮定したうえで臨むことが多いですが、データマイニングではその部分に関しては「まっさら」であり、存在するデータが全てなので､データに語らせようという態度を取ることが通常です。

データマイニングを行う際には、次の観点からの分析手法を使い分けていくことが重要です。

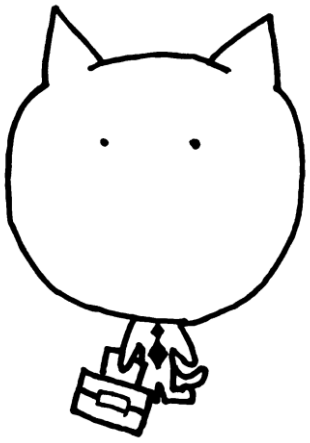
- データの中から関連性が深いとみなせる事項を自動的に抽出する
  ⇨ 関連性（アソシエーション）分析
- データを自動的にまとめ上げ、類似データのグループを作成する
  ⇨ クラスタ分析
- 既知のデータを学習し、未知なデータについての情報を予測する
  ⇨ 分類・判別分析



これらの分析は、Microsoft Excel 上ではなかなか行うことが困難です。これらの分析を容易に行えるようになっているものが、データマイニングツールとよばれるものです。（株）NTT データ数理システム（以下、数理システム）では、データマイニングツールとして **Visual Mining Studio**（**VMS**）を開発・販売しています。

近年、大量のデータを収集・処理するに適うだけのマシンの処理能力の向上、またそれらを分析して「他人が、同業他社が知らないことを知る」ことによるビジネス上の優位性獲得という視点から、データに語らせる、というデータマイニングの観点は非常に重要になってきています。データマイニングは、以下のような様々な分野での利用が進んでおり、成果を上げています。

- 製造業において
  - 品質管理・歩留り向上
- 流通・サービス業において
  - 購買パターン発見・顧客分類
    レコメンデーション
- 金融・保険業において
  - 信用リスク管理・顧客離脱分析
- 医療・バイオ産業において
  - 早期診断・遺伝子解析



1.2.テキストマイニングって？

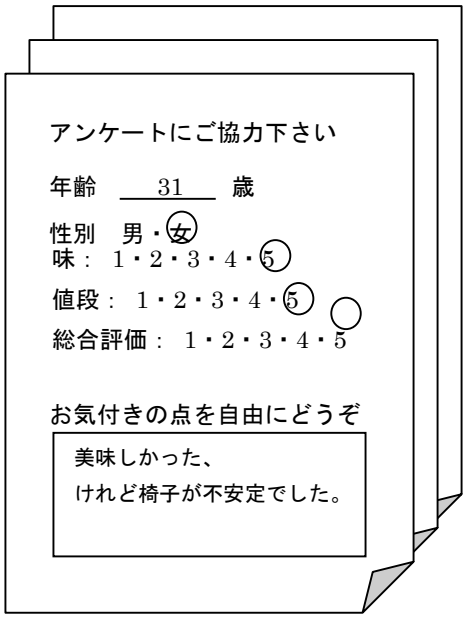
前章で述べたようなデータマイニングの手法を利用しつつ、テキストを分析して情報を発掘していくものが、テキストマイニングです。データマイニングで相手にするようなデータに加えて、テキスト・文章の情報も分析対象とすることにどんなメリットがあるのでしょうか。

まず、テキストデータの大きな性質として、

テキストの中には、
事前に内容を仮定することのできないような情報がある

と言えます。

右図のようなアンケートデータを考えてみましょう。飲食店の座席によく置かれているようなものをイメージしています。



まず、「年齢」の項目には何らかの数値が入っていてしかるべきです。「味」「値段」「総合評価」はお客様が 1～5 のポイント付けを行うもので、これらの回答としては 1～5 の数値が入っていてしかるべきです。これら

の情報は、いわば「数値データ」であり直接的にはデータマイニングの分析対象となるものです。また、「性別」は男性・女性のどちらかであり、回答者が男性なのか女性なのかということを表すラベルのようなものです。こういったラベルの性質を持つ情報は「カテゴリーデータ」とよばれ、やはりデータマイニングの分析対象です。なので、自由回答以外の部分はデータマイニングで処理を行うことが可能です。もし、他のお客様からのアンケートでも大体同じようなポイント付けが行われていれば、おそらく評判の良いお店である、という分析結果が得られることでしょう。

しかし、一番下にある自由記述の部分を見てみると、「美味しかった、けれど椅子が不安定でした」との記述があります。これも飲食店の居心地という点では非常に重要な意見でしょう。椅子がガタガタするせいで落ち着いて食事ができないのであれば、再度来店してもらえる可能性は低くなることは予想できます。もしこの点に気付かなければ、アンケートの評判は良いのにリピーターが増えない、それはなぜだ、と店舗の経営者が思いあぐねることとなります。

この「椅子の座り心地」といった情報は、ポイント付けによる数値データやカテゴリーデータだけでは知りえない情報です。かといって、「椅子の座り心地はどうでしたか？」というようなアンケート項目をわざわざ揃えていくことも無理があります。人によっては、内装が気に入らなかったり、テーブルが低すぎたり、店舗の入り口が分かりづらかったり、喫煙席の煙が流れてきて不快だったり、本当に様々な感想を持つことでしょう。これらをすべて選択式のアンケートとして列挙することはキリがありませんし、いくら増やしたところでアンケート実施者の想像力の範囲内でしか情報を集められないことには変わりません。また、いたずらに項目を増やしたとしても、それはお客様の回答意欲を大いに削ぐことになるだけでしょう。

したがって、定型ではない情報、予期していなかった観点の情報が得られるという点で、テキストを分析すること自体の大きな意味が特に生じてい

る、と言えます。

このようにテキストの情報から何らかの情報抽出を行うことは、伝統的には「人が読んで理解する」という方法が至極当たり前に行われてきました。しかし、この方法ではある程度の量を扱おうとすると手間がかかり、費用対効果の面で大きな問題がありました。また、人の主観が入り込む余地が多く、「分析」に求められる客観性を確保することが困難であるという問題点もあります。

しかし、コンピュータでテキスト・自然言語を扱うための自然言語処理技術が実用的なレベルに達したことにより、テキスト分析・テキストマイニングは既に利用可能な道具として皆様のお手元にお届けできる時代となりました。数理システムでは、テキストマイニングツールとして **Text Mining Studio** を開発・販売し、皆様のご要望に応えております。

特に、テキストマイニングに求められる自然言語処理技術としては、テキストから適切な単語の切り出しを行う「形態素解析」、文章の構造を把握し意味の繋がりを発見する「構文解析」といった要素が大いに進展してきました。しかし、これらの精度を上げるため、また結果を補完するための辞書作成技術が現在大きな課題となっています。特に、類義処理を行うための辞書作成、意味的なまとまりを作り出すための辞書作成といった場面で、皆様それぞれが分野・対象領域の異なるデータをお持ちの状況の中、テキストマイニングツールがいかにこれを支援することができるか、という点の問題解決を図るべく数理システムも取り組んでまいります。

＜続きは読本をお読みください（※）＞

（※）当読本の発送をご希望の方は

tmstudio-info@msi.co.jp まで下記を明記の上、ご連絡下さい。

・メールのタイトル「テキストマイニング読本発送希望」

・ご所属先名（企業名・大学名）

・部署名

・お名前

・電話番号

・ご相談事項などがあれば、何でもお書きのうえご送信ください。

お問い合わせ先
株式会社NTTデータ数理システム
Text Mining Studio担当
〒160-0016東京都新宿区信濃町35番地信濃町煉瓦館1階
TEL:03-3358-6681
FAX:03-3358-1727
E-Mail:tmstudio-info@msi.co.jp
URL:http://www.msi.co.jp/tmstudio/

2014/6

文章　イラスト：**TMS** チーム