# テキストマイニング ― 非構造データからの知見抽出技術 ―

**インフォメーション・オンデマンド（IOD）という新しいコンセプトの下で、企業におけるより効果的でタイムリーな意思決定を可能にするべく、情報資産の活用方法が大きく変化しつつあります。特に企業内情報の85%を占めるといわれる非構造データ（テキストや画像など）の有効活用がポイントです。テキストマイニングは、このような非構造データの中心となるテキスト情報から適切な情報を抽出し、それをさまざまな統計的処理や可視化手法によって知見といえるレベルに集約する技術です。現状ではまだ人手による知見の獲得を支援する段階ですが、コンタクトセンターからライフサイエンス研究機関などに至るまでの広い範囲で応用されるようになっています。**

**日本アイ・ビー・エム株式会社 東京基礎研究所では、1997年にテキストマイニングのプロジェクトをスタート。以来、多数のお客様プロジェクトを経てIBM TAKMIと呼ぶテキストマイニング・ツールを研究開発しており、学会や商業誌などの活動を通じてこの分野をリードしてきたと自負しています。2006年11月には、これまでの事例とテキストマイニングの成功の秘訣（ひけつ）を1冊の本にまとめることができました[1]。本稿では事例を中心に、テキストマイニング技術を紹介します。**

**日本アイ・ビー・エム株式会社**
**東京基礎研究所**
**Senior Technical Staff Member**
**武田 浩一** Koichi Takeda

［プロフィール］
1983年、日本IBM入社。以後、東京基礎研究所において自然言語処理やテキストマイニングの研究開発に従事。英日機械翻訳システム「インターネット翻訳の王様」、インターネット情報収集ツール「mySite Outliner」、テキストマイニング・ツール「IBM TAKMI」などの研究開発に貢献した。2002年には、約1,200万件のMEDLINE生医学文献抄録すべてを対象としたマイニングを実現した。現在は電子カルテなど医療情報のマイニングや新しいビジネスインテリジェンスの実現に取り組んでいる。

Article 2

## Text Mining

**– Technology for Extracting Insight from Unstructured Data –**

The practice of using enterprise information assets has been undergoing a significant change under the new concept of "information on demand," which aims to support efficient and timely decision-making by businesses. In particular, the effective use of unstructured data (such as text and images), which amounts to 85% of corporate information, appears to be crucial. "Text mining" has been proposed as a technology for extracting appropriate information from text, and aggregating it to the level of "insight" by applying a wide range of stochastic analysis and information visualization techniques. Although still at a stage where the technology assists in the human acquisition of insight, text mining technology has been widely deployed in many institutions such as contact centers and life sciences research organizations.

The Tokyo Research Laboratory in IBM Japan, Ltd. started its text mining research project in 1997, and developed the text mining tool called IBM TAKMI through a number of customer engagements. The laboratory takes pride in its leading role in active text mining research publications and recognition in various articles. A book on the past 10 years of text mining activities was recently published in November 2006, and the essence of such text mining technology – along with many specific use cases – is described in this article.

## ❶ テキストマイニングとは?

テキストマイニングは、1990年代にデータベースのような構造化データに対するマイニング技術を、大量のテキスト情報にも適用しようという発想から現れた技術です。当初はテキストを単純に相異なる単語の集まりと見なして、キーワードではなく内容に基づく類似文書の検索（概念検索）や、多数の文書ファイルを内容の似たもの同士でグループ化する（クラスタリング）ツールなどに利用されました。その後、テキストで表現された内容をより構文的・意味的に処理する技術を基に、FAQ（Frequently Asked Question：よくある質問）、製品の不具合に関する報告、お客様の苦情やお褒めの言葉などを抽出し、その増減傾向を把握できるようになったのです。

このようにテキストマイニングという用語は特定の技

術ではなく、大量のテキスト情報から有用な知見獲得を支援する技術の総称として使われています[2]。

テキストマイニングと情報検索とはどう違うのか、というご質問をよくいただきます。情報検索ではユーザーが事前に情報要求を検索条件として表現し、システムがその検索条件に該当する文書や最も関連のある文書を提示します。求める情報がそこに含まれていれば情報検索は成功したといえるでしょう。検索条件に該当する文書がたとえ1件でも1万件でも、求める情報を見つけられる限り（例えば検索結果の最上位に求める文書が提示されれば）問題になりません。

一方、テキストマイニングでは文書集合全体から得られる知見を扱います。例えば「テキストマイニング」について知りたいときに、Ronen FeldmanやMarti Hearstといった人名がよく出てくるとか、最近は評判分析が盛んに議論されている、といった情報が獲得できます。

従って、仮に情報検索が目的でも、求める文書を見つけるまでに、その分野についてのかなりの背景知識や特徴的な情報を得ることができます。極端な場合には、検索結果に含まれる文書を全然読まなくてもよいかもしれません。これがテキストマイニングによる知見獲得の効果です。システムで扱う文書量が多くなれば、一般的にそれに比例して獲得できる知見も豊富になるのが大きな利点です。

## ❷ コンタクトセンターでの応用事例

前章で述べたテキストマイニングの技術が最も普及しているのがコンタクトセンターです。

従来は、人手によって、月次または週次にお客様から寄せられた電話やメールの内容を分析し、定期報告や早期問題発見の業務を行っていました。しかしながら、年間100万件といった規模になると、人手ではその数パーセントをサンプル処理するのがやっとであり、問い合わせ数やお客様IDのような定型情報は全数が把握できても、応対に関するほとんどのテキスト情報が死蔵されることになりました。

この状況を解決したのがテキストマイニングです。コンタクトセンターで扱われる製品の名前や分類（ソフトウェア、ハードウェアやその細分類など）、問題表現（動かない、止まる、起動しないなど）といったテキストに現れる表現を辞書化し、テキストマイニングをカスタマイズします。これによりテキスト情報からFAQ候補の抽出、製品／期間別の問題数とその増減傾向、問い合わせ内容／件数と担当者の応答時間の関係などが把握できるようになりました。

社内のPCヘルプセンターの事例[3]でいえば、1998年にテキストマイニングを検討し、辞書やオントロジーと呼ぶ分類体系の準備を2週間ほど行った後で、試用期間に入りました。通常コンタクトセンターで扱う内容は分野が限定されていることもあり、数千語程度の最頻出単語を抽出・登録することで、大半の用例をカバーできます。

**図1**はテキストマイニング・システムの画面例で、指定した期間で話題になったソフトウェアに関する情報を表示しているところです。このシステムはIBM TAKMIと呼ばれ、東京基礎研究所（以下、TRL）が研究開発したものです。IP（Intellectual Property：知的資産）アセットというソフトウェア資産として多数のお客様に導入されています。

社内では日米の2カ所のPCヘルプセンターに導入され、FAQの作成や、製品出荷後の初期不良の早期発見などに活用されてきました。特に日本の場合は、Webサイトでのセルフヘルプ型のFAQ提供に連動させることで、コンタクトセンターへの問い合わせの負荷を分散させるとともに、Webサイトのお客様満足度を高めることに成功しました。テキストマイニングを効果的に業務プロセスに取り入れることができるようになったのです。

上記のようなコンタクトセンターの事例は、インバウン


**図1. テキストマイニング・システムの画面例**

ドと呼ばれるお客様からの問い合わせ処理が中心ですが、逆にアウトバウンドと呼ばれるお客様へのセールスマーケティング活動でもテキストマイニングが利用されています。具体的には、お客様へのコンタクト記録を分析して、成約に至るパターンを抽出することができました。

興味深いことに、お客様へのお礼を含む表現を中心に分析すると、担当者のスキルレベルが高いグループでは短期的に集中したコンタクトで成約に至るパターンが見つかり、それほどスキルレベルが高くないグループでは継続的なコンタクトで成約に至るパターンが見つかりました。これは直観に合う結果だと思われますが、経験的に認知されてきたベストプラクティスがテキストマイニングで検証されたといってよいでしょう。

アウトバウンドのコンタクト記録の分析からは、ほかにも成約機会を示唆するお客様情報の抽出や、特定の商品への関心の抽出が可能です。データベースに事前に登録されたお客様のプロファイル情報は陳腐化しやすく、なかなか新規データのキャプチャーができる業務プロセスが存在しないものですが、コンタクト記録のテキストマイニングにより解決できるようになったのです。テキストマイニングを利用したマーケット調査については、ほかにも上田らの著書[4]で詳しく紹介されています。

## ❸ 評判分析とナレッジマネジメント

インターネット上の掲示板やブログでの口コミ情報が消費者の購買行動に大きな影響を与えるようになっています。このような情報源の大きな特徴は、良い／悪い、好き／嫌いといった主観的な情報表現が多く含まれていることです。

TRLでは、ある商品やブランドに対する評判情報をマイニングする技術を開発しました。

評判情報処理の困難さは、好評／不評という大きな二つの分類ですらなかなか容易ではないことです。例を挙げると「デザインがよい」と「デザインがよいとはいえない」のように、最初の7文字がまったく同じで、正反対の内容を表現することがあります。つまり単純な文字列や局所的な「～が＋よい」のようなパターンにマッチさせる手法では不十分なことが分かります。さらに「デザインよりも機能がよい」といった比較表現では「デザインはあまりよくない」という、マイルドな不評情報が含まれているといえます。

一般に好評／不評は「よい」のような述語だけで決定できることはまれで、「値段が高い」と「操作性が高い」のように主語や目的語を含めた句表現や、映画における「泣ける」が好評を意味するように、分野依存性を考慮する必要があります。TRLでは、このような分野依存性、構文レベルの好評／不評表現を抽出する手法に加えて、大量のテキストから文脈上の特性を利用した好評／不評表現候補を抽出する手法も開発しています。これにより、人手による評判分析用辞書やパターンの作成に要する作業をかなり軽減することが可能になりました。

評判分析の考え方を広げると、ナレッジマネジメントへの応用につながります。評判の対象を社内の知的資産や特定の事実にフォーカスすることで、企業内で有用な知的資本の特定や再利用を支援できますし、さらに人を対象としたエキスパート検索にも適用可能です。TRLでは2006年に行われたInnovation Jam（全世界のIBM社員がネットワークを通じ、イノベーションについて語り合う場）のデータを分析しています。このような意見交換の情報源から話題の盛り上がりや、同意／反論といった情報を抽出することで、大規模な社内コミュニケーションの概要をリアルタイムに把握することができました。

もう一つ重要な応用として、特許文書や製造業における技術文書のマイニングがあります。特定の技術内容についての他社とのポートフォリオの比較や、新規技術開発に関する先行技術の調査、過去の研究開発からの知見獲得などにテキストマイニングが有望視されています。

## ❹ ヘルスケア・ライフサイエンス分野の応用事例

最後にヘルスケア・ライフサイエンス分野でのテキストマイニング応用事例についてご紹介します。

米国国立医学図書館（NLM）では、1950年代以降に発表された約1,500万件の生医学文献抄録と、MeSH（Medical Subject Headings）と呼ぶ分類体系およびUMLS（Unified Medical Language System）と呼ぶ専門用語の体系を公開しています。

ヒトゲノム・プロジェクトの完了に伴い、ゲノム情報の

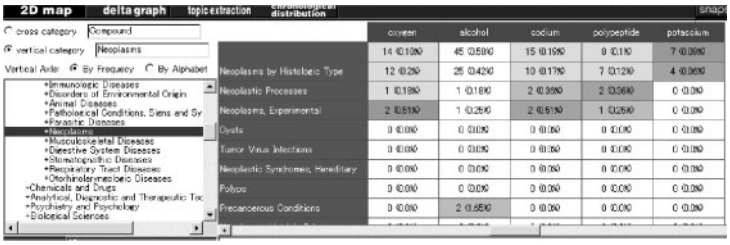


**図2.　MedTAKMIによる分析画面例**

データベースが整備されていますが、論文に含まれる膨大な情報との関連付けは、人手によるキュレーションという作業に依存しているため、年間数十万件の論文が新規に発表される状況に対応できません。これらのリソースは世界中の研究機関で共有されており、テキストマイニングによる知見獲得ができれば、極めて応用範囲の広いソリューションが提供できることになります。

この分野でTRLは、2002年8月にセレスター・レキシコ・サイエンシズ株式会社とともに、IBM TAKMI for Biomedical Documents（以下、MedTAKMI）というソリューションを開発し、MEDLINE全文献抄録のマイニングを実現しました[5]。**図2**は、MedTAKMIの分析画面例です。文献中に現れる新生物（neoplasms）と化合物の記述から、高い相関を示す組み合わせをハイライト表示しています。

最近ではマイクロアレイと呼ばれる遺伝子発現情報の解析器具が普及しており、大量の遺伝子発現を測定することができるようになりました。テキストマイニングを利用して、測定結果に含まれる遺伝子群の機能特定にかかわる記述を効率的に関連付けることで、マイクロアレイ分析の効果的な支援が可能です。東北化学薬品株式会社は、MedTAKMIを利用してマイクロアレイ分析に対するこのような付加価値サービスを提供しています。

## ❺ おわりに

本稿では、テキストマイニングの概念と主な応用事例について説明しました。

IBMではこのようなテキストマイニングの中心となるテキストからの情報抽出アーキテクチャーを、UIMA（Unstructured Information Management Architecture）という非常に柔軟性の高いコンポーネントとして実装しています[7]。UIMAは企業向けサーチエンジンであるIBM OmniFind™ Enterprise Editionにおいて、テキスト情報に対する索引作成用に組み込まれていますが、オープンソースとしても公開されています。これにより、例えば大学のような研究機関が独自のTAE（Text Analysis Engine：情報抽出エンジン）をUIMA向けに公開したり、ISV（Independent Software Vendor）が、特定の業務用途にUIMA向けのTAEを開発することが可能になっています。実際に米国の防衛高等研究計画局（DARPA）が支援するGALEという多言語情報検索／翻訳を実現する政府プロジェクトではUIMAが採用されており、要素技術の相互運用性を高める上で大きな貢献をしています。

最近のBI（Business Intelligence）においてもテキストマイニングが重要な役割を果たしています[6]。今まではデータベース中のいわゆる構造化データのみを利用してBIツールが構築されていましたが、今後はテキストを中心とした非構造データの扱いが進むと予想されます。構造化データでは失われている豊富な背景情報が、非構造データによって補完されますし、特に法令順守の観点で典拠情報（provenance）に関連付けた業務プロセスの支援などは有力な応用分野になってくるでしょう。また、テキストマイニングを利用して非構造データから必要な情報を選択的に抽出し、構造化データとしてBIツールを強化することや、情報検索をより高度化することが一般的になるでしょう。

**[参考文献]**

**[1] 那須川哲哉：テキストマイニングを使う技術／作る技術，東京電機大学出版局, ISBN4-501-54220-9（2006）.**

**[2] 那須川哲哉, 河野浩之, 有村博紀：“テキストマイニング基盤技術，”人工知能学会誌 Vol.16, No.2, pp.201-211（2001.3）.**

**[3] 那須川哲哉, 諸橋正幸, 長野徹：“テキストマイニング－ 膨大な文書データの自動分析による知識発見－，” 情報処理, Vol.40, No.4, pp358-364（1999.4）.**

**[4] 上田隆穂 , 戸谷圭子, 黒岩祥太, 豊田裕貴：テキストマイニングによるマーケティング調査, 講談社, ISBN 4-061-55757-2（2005.11）.**

**[5]“MEDLINE情報を研究開発に有効活用するセレスター・レキシコ・サイエンシズ，” http://www.ibm.com/jp/solutions/lifesciences/solutions/column/no3/**

**[6] 武田浩一：“ビジネス・インテリジェンスと人工知能技術，”情報処理, Vol.47, No.7, pp.723-728（2006.7）.**

**[7] D. Ferrucci, and A.Larry：“Building an example application with the Unstructured Information Management Architecture，” *IBM Systems Journal*, Vol.43, No.3, pp.455-475（2004）.**