# テキスト・データの活用
## ―最適な行動につながる知見を導く―

**財務、経理、営業、製造、研究開発といった企業活動の現場では、日々情報が作られ蓄積されています。また誰もが参加できるソーシャル・メディアが発達し、お客様の声が企業外から直接届くようにもなってきました。**

**近年、これらの情報を蓄積するだけではなく、複雑化する事業環境の中でより優れた判断をするために有効活用しようと考えられるようになってきました。このためには、これらの情報から判断の基盤となる材料を獲得する必要があります。しかし情報の大半を占めるテキスト・データは、私たち人間が日常読み書きする言葉で書かれているため、コンピューターにとっては曖昧で解析が難しいという特徴があります。**

**本稿では、テキスト・データから次の行動につながる知見を得るために必要な技術、またそれらを用いたソリューションについて、IBM の取り組みを踏まえて解説します。**

### 1. はじめに

企業に存在するデータは増え続けています。そのデータの80%以上が非構造化データであり、この非構造化データの活用が、構造化データの活用に比べて十分ではないと言われています（**図 1**）。また、ビッグデータという言葉で呼ばれる企業内外の巨大なデータの集まりも、その大半が非構造化データであり、その活用が課題となっています。非構造化データは、ビデオやイメージなどのマルチメディア・データやテキスト・データから構成され、特にテキスト・データには多くの場合、企業経営の向上に役立つ情報が含まれています。

このような背景の下、近年、テキスト・データの活用はその重要性を増しています。本稿では、テキスト・データの活用にフォーカスし、テキスト・データを扱う上での課題と、その課題を解く技術、さらにその技術を利用したテキスト・データの活用ソリューション事例をご紹介します。

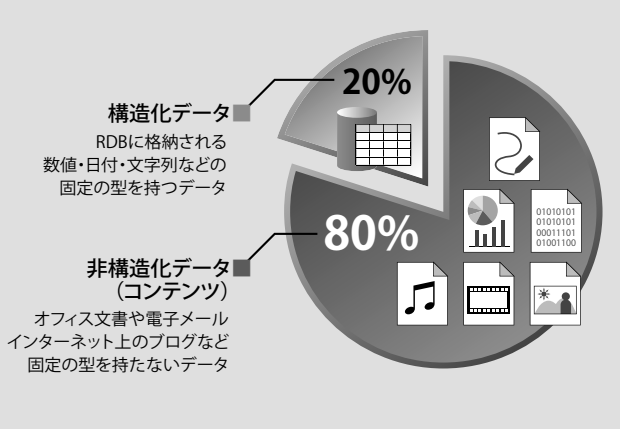


図 1. データの 80%が非構造化データ

### 2. テキスト・データの特徴

**図 2** は構造化データと非構造化データの具体的な例を示したものです。**図2**の左側のように代表的な構造化データであるリレーショナル・データベース（RDB）では、どの値がどのような意味を持つかがフィールド名などで定義されています。このため**図2**の左側のテーブルからは「IBM という名前の会社が 170 カ国以上で事業展開し、40 万人の従業員数を持つ」といったことが、コンピューターでも容易に取り出せます。

一方、**図2**の右側の非構造化データにも同じ内容のことが記されていますが、人間が日常読み書きに使う言葉で記されています。**図 2** では日本語ですが、話者によっては英語やフランス語などさまざまな言語が用いられます。このような言葉は自然言語と呼ばれています。

自然言語は、コンピューターが誤りなく解釈できるよう人工的に設計された Java や C++ などのプログラミング言語と異なり、語彙や文法が多様で、例外規則も多く解釈に曖昧さが残るという特徴があります。コンピューターにとってこのようなテキスト・データをただの文字列ではなく、意味のあるものとして解析するためには多くの技術的なチャレンジがあります。

**構造化データ**

**メタデータにより意味を定義**

例）RDBなどの表形式データ
―フィールド名により意味を表現

| | |
|---|---|
| 会社名称 | IBM |
| 事業展開 | 170カ国以上 |
| 従業員数 | 40万人 |
| 開発研究 | 基礎研究所 12カ所 |
| 製造施設 | 製造施設 24カ所 |

**非構造化データ**

**コンテキスト（文脈）により意味を定義**

例）テキスト・データ
―単語の前後関係で意味を表現

IBMでは、約40万人の社員が世界170カ国以上のお客様に製品やサービスをご提供しています。また、12カ所の研究所と24カ所の製造施設では、科学者や研究者が最先端の技術研究・開発に貢献しています。

図 2. 構造化データと非構造化データ

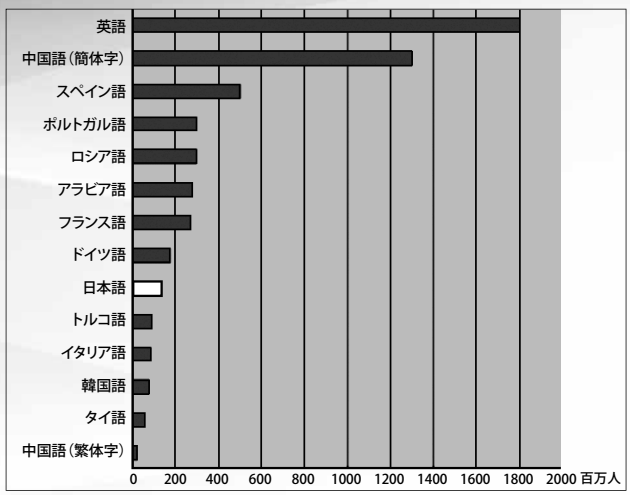


図 3. 世界の主な言語と話者の数　※ Wikipedia のデータをもとに作成

## 3. さまざまな言語

企業が事業をグローバル展開するようになると、入ってくる情報も日本語だけではなくなります。現在世界にはおよそ数千の言語があると言われており、英語をはじめ、中国語、スペイン語、ポルトガル語、ロシア語、アラビア語、ドイツ語と、日本語よりも話者の多い言語がいくつもあります（**図3**）。そして、これらの言語は、みなそれぞれ異なる文字と語彙、文法を持っています。

中でも、日本語は一般的に難しい言語だと考えられています。例えば表記の面では、ひらがな、カタカナ、漢字、アルファベット、数字など多くの字種が使い分けられます。漢字には歴史的な経緯から一つの字に多くの異体字があり、カタカナ語には長音や中点の有無といった揺れがあります。それぞれの単語は空白など明確な区切り文字を間に挟まずつなげて書かれるため、どこで分かち書きすればよいかは文脈によります。文法面では、動詞や形容詞、形容動詞には多くの活用形があり、語順が比較的自由で表現の省略が多用されます。また「田中」と書かれていたときにそれが人の姓なのか地名なのかといった、語の解釈が文脈に依存することもあります。

これらの違いを判別することは、コンピューターにとって容易なことではありません。こうした難しさは日本語に限ったことではなく、どの言語にもその言語特有の難しさもあります。例えば英語ならば、“'s” は “is” なのか “has” なのか “was” なのか、それとも所有を表すアポストロフィ “'s” なのかは文章中の使われ方に依存しており、すぐには判別できません。“book” が名詞なのか動詞なのかといった品詞の曖昧さもあります（**図4**）。また日本語同様に、語の解釈が文脈へ依存することがあります。例えば、“EPS” という語を聞いたときに何を思い浮かべるでしょうか。株価の動向に注意を払っている方なら “Earnings Per Share（一株あたり利益）” をまっ先に思い浮かべるでしょう（**図5**）。印刷業界の方ならば “Encapsulated PostScript（画像ファイルフォーマット）” かもしれません（**図6**）。綴りは平凡ですが

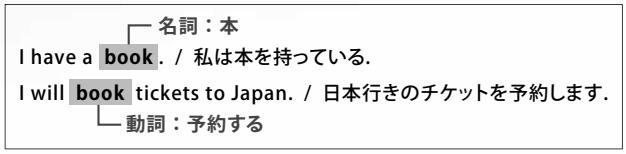


図 4. book = 本、予約する

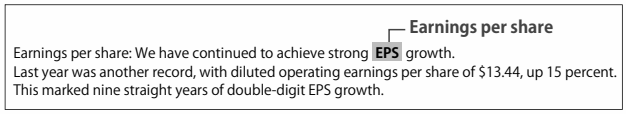


図 5. EPS = Earnings Per Share（一株あたり利益）

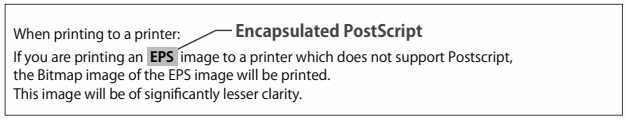


図 6. EPS = Encapsulated PostScript（画像ファイルフォーマット）

Wikipedia によるとこの語には、“External Power Supply（外部電源）” “Electric Power Steering（電動パワーステアリング）” “European Protected Species（EU 指定天然記念物）”、さらにはスタートレックに登場する “Electro-Plasma System” まで 30 余りもの意味があるとされています。そのどれが正しいかは情報が属する領域（金融、印刷、電気、機械、環境保護、メディアなど）によるため、一概には決められません。

例えば英語では、慣用表現や不規則変化動詞（例： go（現在形）－ went（過去形）－ gone（過去分詞形））といった文法上の例外規則が多々あります。また、そもそもアルファベットで書かれたテキストが英語であるとも限りません。英語で使われるアルファベットの多くは、ドイツ語やフランス語などの他のヨーロッパ系諸言語でも使われており、綴りが同じ語も存在します。ドイツ語では書き手が動詞や形容詞、名詞を組み合わせて複合語を作ることができます。読み手は複合語を分解してみないと元の表現が何であったか分かりませんが、そのためにはいくつもある分解候補の中から文法上どれが最も適しているかを決めなければなりません。

アラビア語やヘブライ語といった中東の言語では、左から右へ書く語（アルファベットで綴られた外来語や数字）と、右から左に書く語（伝統的なアラビア文字やヘブライ文字で綴られた語）が一つの文の中で混在しており、取り扱いに注意が必要です（**図7**）。一見奇妙な書き方に思えますが、日本語も戦前は横書きを右から左へ書くケースと、左から右へするケースが混在しており、統一されていませんでした。

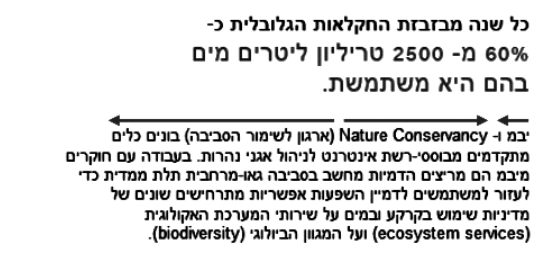


図 7. 一つの文の中で双方向に綴られるヘブライ語

## 4. 求められる要素技術

このように文字、語彙、文法が多種多様なテキスト・データをコンピューターで解析するためには、以下のような要素技術が必要になります（**図 8**）。


図 8. テキスト・データを解析するための要素技術（自然言語処理）

① テキストが何語で書かれているかを推定する「言語識別」
② 推定された言語の文法に基づいて、テキストを単語に切り分ける「単語識別」や、見つかった単語ごとに文中での役割（品詞）を求める「品詞推定」
③ 単語の並びから人の姓名や組織名、地名といったひと続きで意味を持つまとまりを見つける「固有表現抽出」
④ どの語がどの語を修飾しているかを見つける「係り受け解析」
⑤ 楽しい、つまらないといった表現（フレーズ）を見つけ出す「感情表現抽出」など

これらの要素技術は一般に自然言語処理（Natural Language Processing）と呼ばれ、テキスト・データを解析する上で不可欠な要素技術です。IBM では特に、①②③④については研究開発で使う共通の基礎コンポーネントとして、多言語に対応した LanguageWare と呼ばれるライブラリーを内製しています。LanguageWare では企業向けソフトウェアに求められる高パフォーマンス・省メモリーといった特性を満たしつつ、言語間で同じ処理ができる部分は共通化してパフォーマンスの最適化とメンテナンス・コストの低減を図っています。

## 5. テキスト・データを解析するためのプラットフォーム

テキスト・データを解析するためには、**図 8** に示すような要素技術が必要ですが、これらがばらばらのままでは役に立たず、それぞれのコンポーネントをモジュール化し、コンテナ上で一連の解析フローとして連携させる仕組みが必要になります。そのようなテキスト・データを解析するためのプラットフォームとして、本稿では、Unstructured Information Management Architecture [1]（以下 UIMA）、System-T [2]、General Architecture for Text Engineering [3]（以下 GATE）の 3 つをご紹介します。

### ● UIMA

UIMA は 4 章にあげたような解析エンジンを連携させるためのプラットフォームとして代表的なソフトウェア・フレームワークです。一般にテキスト・データの解析では求められる機能や対象の言語が多岐にわたるため、一つの企業や学術機関が提供する解析エンジンだけでは機能が不足することがあります。その場合強みのある解析エンジンを相互に連携させるための、何らかのオープンな仕組みが必要になります（**図 9**）。

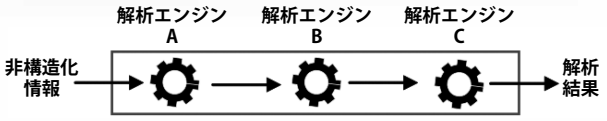


図 9. UIMA

UIMA は、このために IBM ワトソン研究所により提唱されたフレームワークで、①解析エンジン間でやり取りするデータの形式を共通化し、②個々の解析エンジンをモジュール化し、③配布のための標準的なパッケージ形式を定めています。これにより解析エンジンのモジュール性や、解析エンジン間の相互運用性が確保できる仕組みになっています。

UIMA は 2006 年より Apache ソフトウェア財団傘下のオープンソース・プロジェクトとして実装が提供され、誰でも利用することができます。また解析エンジン間の相互運用性が将来にわたって特定の企業や団体に依存しないようにするために、仕様については 2009 年に国際標準規格化団体 OASIS から承認を得ています。IBM では 4 章で述べた LanguageWare や、6 章で述べる IBM Watson [4]（以下、Watson）、IBM Content Analytics with Enterprise Search [5]（以下 ICA）などが、UIMA に準拠した解析エンジンを用いています。これにより IBM 内製の解析エンジンとサードパーティー製の解析エンジンを連携させて全体の解析機能を拡充するといったことが可能になっています。

### ● System-T

System-T は IBM アルマデン研究所により研究開発されたテキスト解析のためのツールキットです。System-T では Annotation Query Language（AQL）と呼ばれる独自の照会言語を使って、プレーン・テキストから照会条件に合致する単語や固有表現を取り出すことができます。内部では要素技術として LanguageWare を用いて単語やその品詞を識別しています。AQL は SQL に似た命令セットを持ち、「create view」「extract」「select」「output」といった構文を使って照会を行います。これにより SQL に親しんだ開発者にとって初学時の学習障壁が下がることを期待しています（**図 10**）。

System-T は、Apache Hadoop を基礎とする IBM InfoSphere BigInsights（以下、BigInsights）に同梱されており、Map-Reduce 方式で照会を分散バッチ処理することができます。これにより Twitter などのソーシャル・ネットワーキング・サービス（以下、SNS）やブログなど日々大量に生成されるテキスト・データの中から特定の語が使われた頻度をカウントしたり、流行の語を見つけ出すといったことが素早く行えます。

```
create view PhoneCandidate as
extract
        regexes /\d{3}-\d{3}-\d{4}/
            on D.text as num
from Document D;

output view PhoneCandidate;
```

図 10. プレーン・テキストに含まれる電話番号を照会する AQL

### ● GATE

GATE は UIMA とよく比較されるフレームワークです。自然言語処理コンポーネントとその周辺ツール、統合開発環境から成り、GNU Lesser General Public License（LGPL）に基づき配布されるオープンソースです。イギリスの大学による研究プロジェクトのため、どちらかというとヨーロッパの学術機関で採用されることが多いプラットフォームです。UIMA と異なり仕様が標準化されていませんが、10 数年にわたり研究開発が続けられてきた老舗のプロジェクトです。

## 6. テキスト・データから知見を導く

4 章、5 章で解説した技術を利用してテキストから取り出された単語や品詞、固有表現、感情表現などは、個々のドキュメントの特徴を表します。自然言語処理では一つのドキュメントから数万個の特徴が抽出されることも珍しくありません。これらの特徴の中には、重要なものとそうでないもの、あるいは他のドキュメントと突き合わせると関連が見つかるものなどが入り混じっています。

この玉石混交の特徴群を基にして、人間が実際に意思決定するのに役立つ知見を取り出すためには、特徴の頻度や偏差、相関の計算、主成分分析、確信度分析といった統計解析の手法を組み合わせることが有効です（**図 11**）。そのような例として本稿では、WatsonとIBM Content Analytics with Enterprise Search（以下 ICA）をご紹介しますが、その他に BigInsights や SPSS 製品にもこれらの技術が活用されています。

### ● IBM Watson

Watson は 2011 年に米国の著名な長寿クイズ番組「Jeopardy!」で人間のチャンピオンに挑戦し勝利した質問応答システムです。この番組で出されるクイズは、単純に百科辞典をキーワード検索すれば答えが分かるようなものではありません。凝った表現、語呂合わせ、言葉遊びなどがたくさん使われた問題文から手掛かりを見つけ、答えを推測して当てる問題（**図 12**）であるため、コンピューターで解くのは非常に難しいのです。

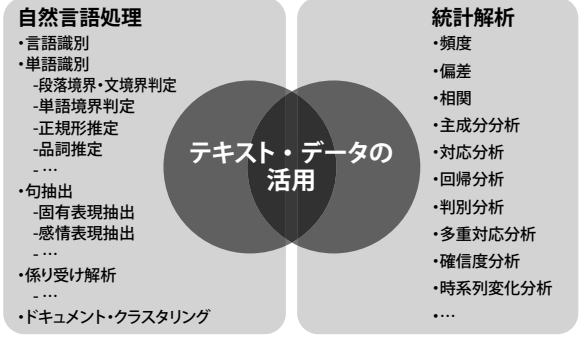


図 11. テキスト・データの活用＝自然言語処理＋統計解析

Watson は、IBM の基礎研究部門が中長期的な視野に立ち、技術的なブレークスルーを目指すグランド・チャレンジとして 4 年をかけて研究開発が行われました。

| | |
|---|---|
| 1 | **カテゴリー：Dialing for Dialects（方言について答えよう）**<br>問題文：While Maltese borrows many words from Italian, it developed from a dialect of this Semitic language<br>（マルタ語はイタリア語から多くの語彙を借りているが、それはこのセム語系の方言から発展した）<br>答え：Arabic（アラビア語） |
| 2 | **カテゴリー：Alternate Meanings（2つの意味を持つ単語）**<br>問題文：4-letter word from the iron fitting on the hoof of a horse or a card-dealing box in a casino<br>（馬のひづめに付ける金具、またはカジノでカードを入れる箱を表す4文字の語）<br>答え：Shoe |

図 12.「Jeopardy!」で Watson が正答したクイズ

Watson では**図 12** のような質問に答えるために、2 億ページに相当する膨大なコンテンツが、2,880 個の POWER プロセッサー・コアと UIMA に準拠する数百個の解析エンジンにより、さまざまな角度から解析されました（**図 13**）。自然言語処理を用いた解析の結果、答えの仮説はいくつも見つかります。その中からどれが最も適切か判断するため、統計モデルに基づく確信度の分析が用いられました。Watson で実現された質問応答技術は、例えばヘルスケアの分野で医療従事者がさまざまな医療情報を基に治療する際の意思決定支援に応用ができると考えられています。

### ● IBM Content Analytics with Enterprise Search (ICA)

ICA は、テキスト・データの山から役立つ情報を見つけ出すために使われる検索、テキスト・マイニング・ソフトウェアです。

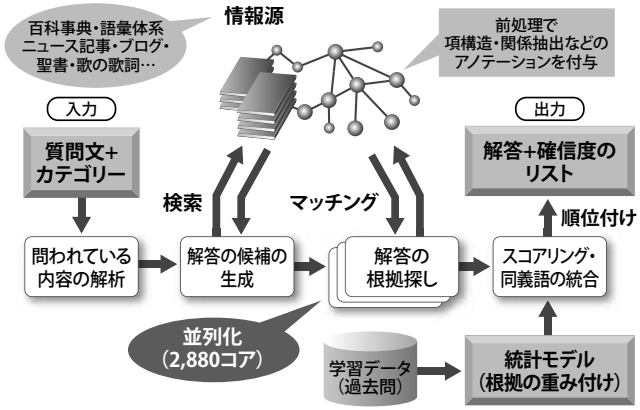


図 13. Watson の解析フロー

要素技術として LanguageWare を用いており、検索は 20 言語、マイニングは 15 言語に対応しています。ICA ではユーザーが大量のデータをさまざまな角度から分析して意思決定につながる知見を導けるよう、テキスト・データの処理結果を統計解析に基づく 7 つの切り口（分析ビュー）で提供しています（**図 14**）。

図 14. ICA の分析ビュー

① ファセット分析：品詞、固有表現といったファセットごとに含まれる語の出現頻度、相関値を見るためのビュー
② 時系列分析：時系列方向でデータの出現頻度の推移を分析するためのビュー
③ 偏差分析：あるファセット内で、時系列方向に激しい増減を示す語を検出するためのビュー
④ トレンド分析：ある語が時系列方向で顕著な増加を示したときに検出するためのビュー
⑤ ファセット・ペア分析：ある 2 つのファセットのペアの中で、高い相関を持つ語を検出するためのビュー
⑥ コネクション分析 : ある 2 つのカテゴリー間の相関関係をネットワーク・グラフで視覚的に表示するビュー
⑦ 感情表現分析：肯定表現・否定表現などの感情表現を読み取るためのビュー

これらの切り口を組み合わせてデータを絞り込んでいくことで、大量の生データを見ているだけでは気づかなかった事実にスポットをあて、サービス品質の改善、不正行為の検知、意思決定の最適化など、次のビジネス施策につながる知見を導けるようになります。

またこのような用途のソフトウェアでは、テキストを解析した結果が膨大な量のレコードになるため、ICA ではシステムの応答時間を短くし、ユーザビリティーを向上させるため、特にボトルネックになる部分については独自に開発したアルゴリズムに基づく索引を採用しています。これにより数百万～数千万のドキュメントを高速にマイニングすることが可能になっています。

## 7. テキスト・データを活用したソリューション

本章では前述の技術、製品を用いてテキスト・データから次の行動につながる知見を導くソリューションとして、IBM の取り組み事例を4つご紹介します。

### ●ヘルスケア

医療機関では医師のメモや問診票、退院時病歴要約といった診療情報、事務情報が大量に蓄積されています。このテキスト・データの中に埋もれている情報間の関連性を検索、調査、把握できれば、より高度な診断や治療につなげられると考えられます。このためには医療情報に含まれる傾向、パターン、偏差を特定し、予想される結果の確度を割り出すソリューションが役立ちます（**図 15**）[6]。

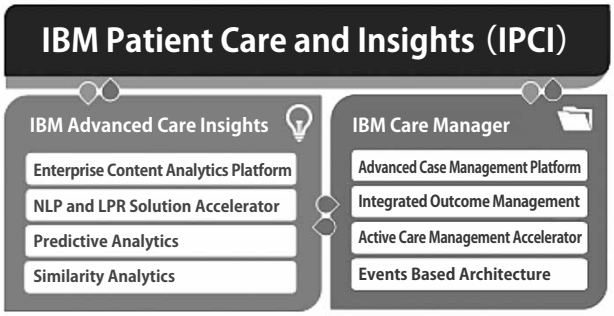


図 15. 医療情報を解析するソリューション「IBM Patient Care and Insights」

### ● SNS の分析

Twitter や mixi などの SNS には、毎日たくさんの人がメッセージを書き込みます。これら SNS 上の情報の中には、例えば市場における株価の変動と関連性が高いものもあると考えられます。人手で SNS 上のテキスト・データをすべて精査することはほぼ不可能ですが、システムを使って情報を加工すれば、これまでにない新しい観点での投資情報としてお客様に提供できる可能性があります（**図 16**）[7]。

### ●「お客様の声」分析

多くの企業では、電子メール、コールセンター・ログ、Web で行う調査から得た数千のコメントなど、日々多くの「お客様の

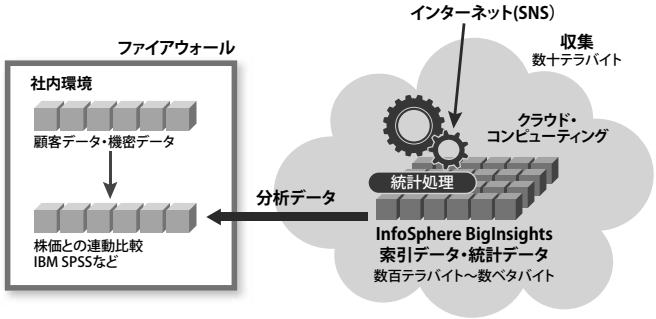


図 16. ICA と BigInsights を用いた SNS 情報分析システム

出展：PROVISION 72 号

声」が集められています。これら「お客様の声」を継続的に分析することで、企業は顧客満足度を高め、顧客ロイヤリティーを改善していけると考えられます。しかしこれらのテキスト・データを人手で解析した場合、「量が膨大で時間とコストがかかる」「解析担当者のスキルによって結果にばらつきが出る」といった問題が生じます。テキスト・データを分析するプロセスの正確性とスピードを改善できれば、サービス品質の改善など具体的な行動につながる洞察を素早く導けるようになります [8]。

### ●品質モニタリング

車社会の米国では、米国運輸省道路交通安全局（National Highway Traffic Safety Administration：以下、NHTSA）が自動車の安全性に関する調査を実施しデータを公開しています。この中には、自動車ユーザーが記入した 80 万件以上のテキスト・データがあります。これらのデータを分析すれば、自動車関連の不具合を早期に発見し、開発・製造における品質改善につなげられると考えられます。このためには、増え続ける膨大なテキスト・データから車種と部品とにおける問題発生率を確認し、発生率が高いものについては不具合と相関が高いものが何かを見つけ出し、実際の問題レポートの中身を検証する、というステップを効率的に繰り返せるシステムが有効になります（**図 17**）[9]。

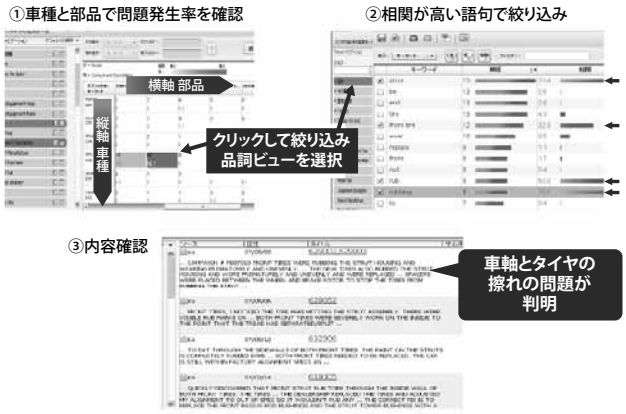


図 17. ICA を用いて NHTSA が公開するテキスト・データを分析する例

## 8. おわりに

テキスト・データを活用したソリューションは、本稿で紹介した例以外にも、不正請求の検知、保険の未払い防止、犯罪情報分析などたくさんあります。また、他のアナリティクス技術を組み合わせることで、テキスト・データをさらに有効に活用できる場合もあります。日々増え続けるテキスト・データをタイムリーに分析し、そこで得た知見を迅速に行動につなげていくことが、ビジネスの差別化のために、今後よりいっそう重要さを増していくと考えられます。

[参考文献]

[1] Apache UIMA Project、http://uima.apache.org

[2] Y. Li、F.R. Reiss、L. Chiticariu 、“SystemT: a declarative information extraction system”、Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Tech-nologies: Systems Demonstrations、pp. 109–114、2011

[3] General Architecture for Text Engineering：http://gate.ac.uk/

[4] 武田浩一、金山博、質問応答システム Watson が示す未来ー質問応答システムがもたらす情報処理の新たな世界ー PROVISION No.70.（2011）

[5] IBM: IBM Content Analytics with Enterprise Search：http://www.ibm.com/software/jp/data/search/index.html

[6] IBM: IBM Patient Care and Insights：http://www.ibm.com/software/ecm/patient-care

[7] IBM：インターネット上の膨大なデータを収集・分析し、株価との関連性に基づいた新サービスの提供を模索、PROVISION No.72、(2012)

[8] IBM: 導入事例　Hertz Corporation が IBM と Mindshare Technologies のソリューションを活用し、アナリティクスに基づくより深い洞察を獲得：http://www.ibm.com/software/jp/data/casestudies/search/hertz.html

日本アイ・ビー・エム株式会社
ソフトウェア開発研究所
技術理事

**濱田 誠司**

Seiji Hamada

［プロフィール］

1986 年、日本 IBM 入社。複数のソフトウェア製品開発をリードし、現在はエンタープライズ・サーチ製品、テキスト・アナリティクス製品である IBM Content Analytics with Enterprise Search の開発の技術的な総責任者を務める。

日本アイ・ビー・エム株式会社
ソフトウェア開発研究所
スタッフ・ソフトウェア・エンジニア

**中山 章弘**

Akihiro Nakayama

［プロフィール］

2001 年、日本 IBM 入社。以後、ソフトウェア開発研究所で自然言語処理コンポーネント LanguageWare の開発、ソリューション支援に従事。2010 年より IBM Content Analytics with Enterprise Search 開発プロジェクトに参画。