膨大なテキスト情報が生み出す新しい価値
テキストマイニング技術を活用した生産性の向上

Creating New Value using Mass Text Data
Text Mining for Performance Improvement

尾﨑 和基[*1] Kazuki Ozaki
山縣 迅[*2] Shun Yamagata
錦 尚志[*3] Hisashi Nishiki
田口 孝[*4] Takashi Taguchi

製造時に発生する不適合の事象と原因，設備運転時のトラブルへの対処法，特許情報などのテキスト情報は，物理量等を表形式で表現し，構造化※が可能な数値情報と比較し，一般的に構造化されていないデータとして保有されているため，分析が進まず，活用が難しい状況にある。これらのテキスト情報を数値化・構造化できれば，数値情報に対する分析アプローチを活用することで，文章の分類や類似文章の検索が可能となり，新たに発生する事象に対する早期対応などテキスト情報の有効活用が期待できる。本報では，数値化・構造化を実現するテキストマイニング技術と適用事例について紹介する。

※構造化とは，大容量で複雑な事象を段階的に小さな区分に仕分けして取扱いを容易にすること

1. はじめに

地球環境問題や電力自由化などにより大きくなりつつあるエネルギー環境の変化に対応し，より効率的なエネルギー運用を可能とするため，当社ではソリューションの一つとしてデータ分析技術を活用したエネルギーデマンドの予測に取り組んでいる。ここで分析の対象となるデータは，工場設備に取り付けられた各種センサにて計測された物理量，カレンダ情報，気象情報そしてこれらを紐づける時間情報など数値情報で構成され，表形式で表現される構造化されたデータである。

他方，一般に，テキスト情報をはじめとする構造化されないデータは，企業が保有する全データの 80%を占めるといわれており，当社においても製造時における不適合事象への対策，効率的な設備運転のノウハウ，特許文献などの膨大なテキスト情報を保有しているが，構造化されていないために，テキストの相関を定量的に評価するなどといった分析が進まず，活用が難しい状況にあった。

当社では，このような構造化されていないテキスト情報に対し，後述するテキストマイニング技術の適用による数値化・構造化を行うことで，膨大な文章のクラスタリング（分類）や類似文章検索を実現し，収集したテキスト情報からの有用な情報抽出・活用を図っている。この取組みにより，不適合事象への対策や設備運転時に発生するトラブルへの対応方法の提示や文書確認作業の効率化などを可能とした。

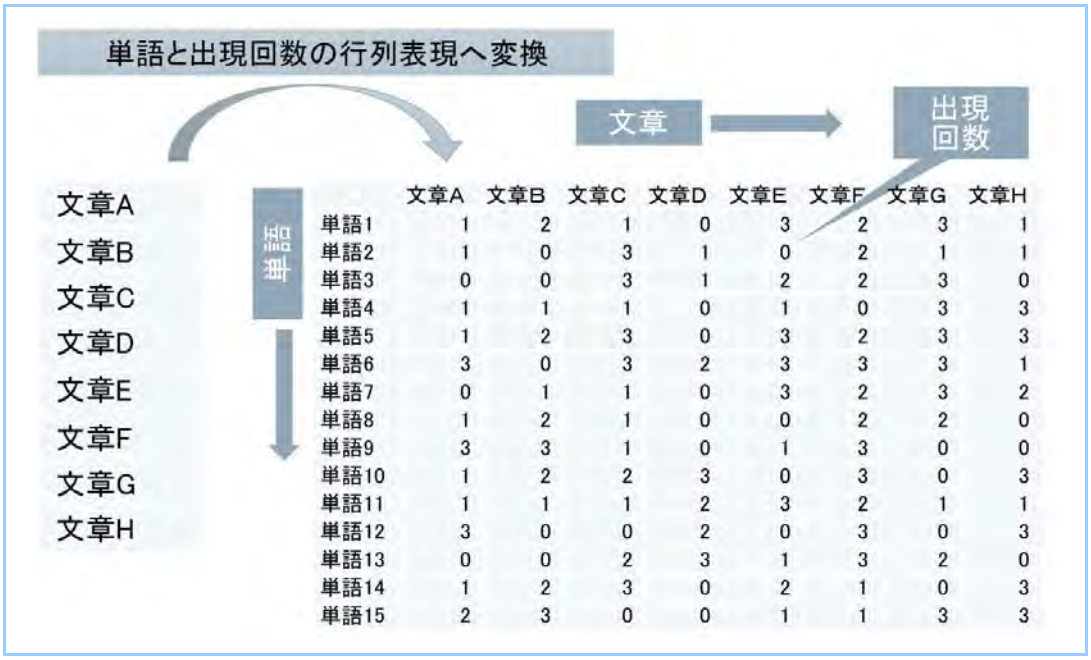
*1 ICT ソリューション本部 EPI 部
*2 パワードメイン パワー＆エネルギーソリューションビジネス(PESB)総括部 PESB 企画室 主席プロジェクト統括
*3 パワードメイン PESB 総括部 PESB 企画室 主席技師 *4 パワードメイン PESB 総括部 PESB 企画室

2. テキストマイニング技術を活用した分析手順

テキストマイニング技術の活用により，テキスト情報を数値情報へ変換し，構造化されていない膨大なテキスト情報を構造化し，クラスタリング（分類）や検索を行う手法について4つのステップで紹介する。ここでの手法は汎用化されており，次章で述べるテキストマイニングの実証事例において共通である。

ステップ1：テキスト情報の数値化

形態素解析により文章を単語に分割し，その出現回数を算出することで，テキスト情報から数値情報への変換を行う。文章を単語と出現回数で構成される表形式の表現とすることで，計算機での処理が容易となる（図1）。



	文章A	文章B	文章C	文章D	文章E	文章F	文章G	文章H
単語1	1	2	1	0	3	2	3	1
単語2	1	0	3	1	0	2	1	1
単語3	0	0	3	1	2	2	3	0
単語4	1	1	1	0	1	0	3	3
単語5	1	2	3	0	3	2	3	3
単語6	3	0	3	2	3	3	3	1
単語7	0	1	1	0	3	2	3	2
単語8	1	2	1	0	0	2	2	0
単語9	3	3	1	0	1	3	0	0
単語10	1	2	2	3	0	3	0	3
単語11	1	1	1	2	3	2	1	1
単語12	3	0	0	2	0	3	0	3
単語13	0	0	2	3	1	3	2	0
単語14	1	2	3	0	2	1	0	3
単語15	2	3	0	0	1	1	3	3

図1　テキスト情報の数値化

ステップ2：テキスト情報からの潜在的な意味情報抽出

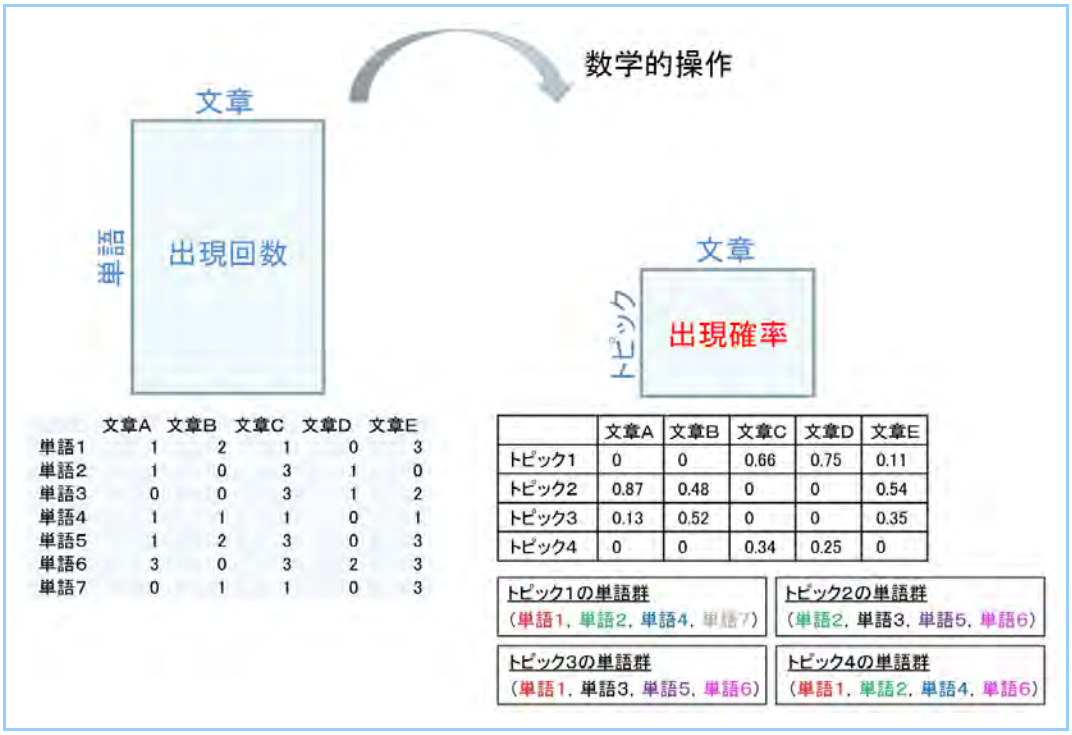
上述の単語と出現回数の行列に対し，数学的操作（確率・統計的操作）を適用することで，文章を単語と頻度の行列ではなく，単語群から成る潜在的な意味情報（トピック）とその出現確率で表現可能とする変換モデルの構築を行う（図2）。

ステップ3：文章のクラスタリング（分類）

ステップ2の操作にて，文章ごとに得られるトピックの出現確率ベクトルを対象に，ベクトル間距離を用いた階層クラスタリングによる文章分類を行う（図3）。

ステップ4：文章の検索

任意の入力文章に対してステップ1及びステップ2の操作を行うことで，検索対象の文章と同様にトピックの出現確率ベクトルを算出し，検索対象となる文章群の出現確率ベクトルとの比較から算出される類似度により，類似文章の検索・抽出を行う（図4）。

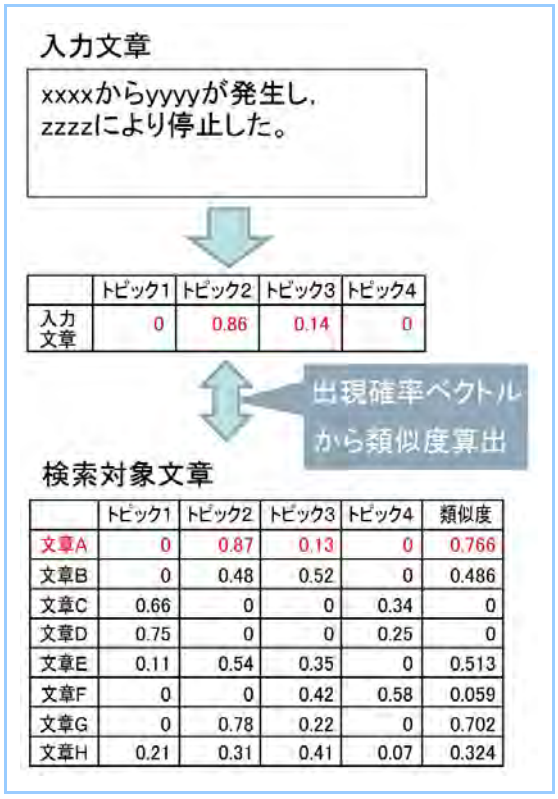


	文章A	文章B	文章C	文章D	文章E
単語1	1	2	1	0	3
単語2	1	0	3	1	0
単語3	0	0	3	1	2
単語4	1	1	1	0	1
単語5	1	2	3	0	3
単語6	3	0	3	2	3
単語7	0	1	1	0	3

	文章A	文章B	文章C	文章D	文章E
トピック1	0	0	0.66	0.75	0.11
トピック2	0.87	0.48	0	0	0.54
トピック3	0.13	0.52	0	0	0.35
トピック4	0	0	0.34	0.25	0

図2　潜在的な意味情報の抽出

図3　階層クラスタリングの例



	トピック1	トピック2	トピック3	トピック4
入力文章	0	0.86	0.14	0

	トピック1	トピック2	トピック3	トピック4	類似度
文章A	0	0.87	0.13	0	0.766
文章B	0	0.48	0.52	0	0.486
文章C	0.66	0	0	0.34	0
文章D	0.75	0	0	0.25	0
文章E	0.11	0.54	0.35	0	0.513
文章F	0	0	0.42	0.58	0.059
文章G	0	0.78	0.22	0	0.702
文章H	0.21	0.31	0.41	0.07	0.324

図4　入力文章に対する類似度算出

3. テキストマイニングの実証事例

本章では，テキストマイニングを活用した3つの実証事例を紹介する。

(1) 不適合情報管理におけるカテゴリ提示についての実証例

品質保証部門では，不適合発生時にその内容・原因・対策などを登録するシステムを用いて過去の不適合情報の管理を行い，これらの情報を活用した再発防止活動に取り組んでいる。この活動は，製品の品質のみならず，コスト，納期，安全，顧客満足にも大きく関わることから，必要不可欠なものである。

システムへの不適合情報の登録は各部門で行われるが，事象の表現方法が属人的で統一されていないなどの理由により，必ずしもこれらの情報は再発防止活動に活用しやすく整理されていない。

そこで，過去の不適合情報に対してテキストマイニング手法を活用することで，テキスト表現の異なる類似事象を自動的にクラスタリングし，それぞれのクラスタを対策に関連付けられる不適合原因に自動分類するモデルを作成した。本モデルを活用することで，従来属人化していたテキスト表現を容易に分類できるようになり，再発防止活動のレベルアップが図られた。

(2) 大型設備試運転時のトラブル対応についての実証例

大型設備・プラントの運転，特に立ち上げ時の試運転でトラブルが発生し，対応に時間を要すると，運転開始時期の遅延など損失が大きい。これに対し，試運転部門では，過去に発生したトラブルについてフォルトツリー解析(Fault Tree Analysis: FTA)を実施し，その内容・原因・対策などを“戦訓集”としてまとめ，トラブル対応や若手技術者の育成に活用している。

しかしながら，その内容は多岐にわたり，複雑な技術用語を含むテキスト情報で構成されることから，従来の整理方法では，トラブルが発生した際の適切な戦訓集の選定にベテラン技術者の経験とノウハウが必要となり，時間を要する作業となっている。

そこで，過去の戦訓集にまとめられたトラブル内容や原因，対策をいったん分解した上でデータベース化し，各情報の関連性に対してテキストマイニング手法(2章のステップ1～ステップ3)を用いて，階層的に分類を行った。その結果，類似事象が発生した際に，例えば“x号機のyがzの信号で停止した。”といったテキストを入力することで過去の事象との類似箇所を2章のステップ4の手法にて即座に特定し，適切な対応・処置を検索できるようになった。従来は，FTA を実施すること自体に高度な知識と経験が必要であったが，テキストマイニング手法を使った階層分類を行うことで，膨大なテキスト形式の記録を有効に活用できるようになった。

(3) 知財部門による特許調査業務支援についての実証例

技術動向等に係る特許調査業務においては，従来は技術分野ごとに付与された特許分類コード及び当該分野の製品や機器に係るキーワード等により，他社特許調査範囲を絞り込んでいた。しかし，“AI&IoT”のように抽象的なキーワードに対して適切な特許分類コードが選定しにくい新技術観点の調査では，好適な絞り込みができず，特許調査範囲が広範囲に及んでしまう。

そこで，調査範囲の絞り込みを効率的に行うためにテキストマイニング手法を活用し，抽象的なキーワードに対して対象範囲の中から関連性の高い技術分野に自動分類できるようになり，分野ごとの専門者にて詳細調査を実施するといった後段の作業に対する効率化や，より精度の高い調査を行う取組みを推進できるようになった。

4. まとめ

本報では，テキストマイニング技術の活用により，膨大なテキスト情報の数値化・構造化を実現し，クラスタリング(分類)や検索を実現する手法の紹介と，この手法を不適合情報，戦訓集，特許などのテキスト情報に適用することで得られた成果について，事例を基に紹介した。

テキスト情報に対する汎用的な分析手法を確立したことで，数値情報以外に活用可能なデータ種類が拡張され，効率的な設備運営や業務プロセスのリードタイム改善を実現するために，予測以外に分類・検索を活用したこれまでとは異なるソリューション提案の可能性が示唆された。

今後は，テキスト情報の発生時刻などの付随情報や現場に蓄積されている膨大な数値情報を結合して新たな知識を発掘し，ソリューション提案の幅を拡大していく。