高度なテキスト分析による知識抽出の応用

Applications of Knowledge Extraction using Advanced Text Analysis

清水　航　高度情報マネジメント統括本部 IoT プロジェクト部
中安有希　高度情報マネジメント統括本部 IoT プロジェクト部
鈴木由宇　高度情報マネジメント統括本部 IoT プロジェクト部　主査
河野幸弘　高度情報マネジメント統括本部 IoT プロジェクト部　部長

不具合対応履歴，メンテナンス履歴，設計図書，打合せ議事録などのテキストデータには，不具合対応や設計改善業務などにつながるさまざまなノウハウが含まれている．そのため，テキストデータからこれらのノウハウを抽出するテキスト分析技術は，業務の効率化や知識伝承のために非常に重要である．本稿では，IHI グループにおけるテキスト分析技術を用いた業務効率化の取組み事例や，精度向上・展開に向けた取組みについて紹介する．

Text analysis is one of the essential data analysis technologies for manufacturing industries because it helps them with extracting experts' knowledge for work efficiency and training etc. from text data including business-related documents. The authors of this paper introduce applications of text analysis for work efficiency in the IHI Group. In addition, the authors explain techniques of accuracy improvement and actions to familiarize text analysis.

1. 緒　　言

IHI グループには，不具合対応履歴やメンテナンス履歴などの多数のテキストデータが蓄積されており，これらには，不具合時の対処方法といったベテランのノウハウが多く含まれている．一方，これらは情報量が膨大であり，蓄積された情報をいまだ十分には活用できておらず，ベテランのノウハウが失われつつあることが危惧されている．

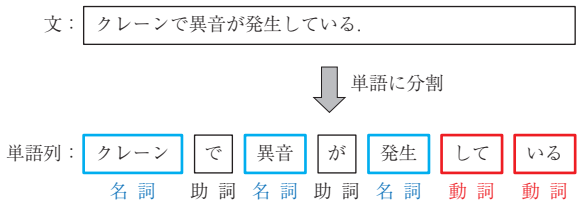
そこで，テキスト分析技術を応用することによって，蓄積したテキストデータに含まれるベテランのノウハウを有効活用する取組みを現在進めている．

本稿では，まず，テキスト分析に関する技術を説明し，次にテキスト分析を適用した業務効率化の取組み事例を紹介する．最後に，テキスト分析の精度向上・展開に向けた現在の取組みを紹介する．

2. テキスト分析とは

2.1 基本的な考え方

一般的に，テキスト分析 [1] を行う場合には，形態素解析を行う．**第 1 図**に形態素解析の例を示す．形態素解析とは，文章に含まれる単語やその品詞・活用形などを文章から求める処理のことである．形態素解析を用いて文章を単語単位に分割することによって，単語や単語の組合せの出現回数を計算できる．これによって，頻度分析やパレート分析といった統計的な解析が可能になる．また，形態素解析の結果に統計的手法や機械学習手法などを適用することによって，テキスト検索や文書分類などが可能になる．



第 1 図　形態素解析の例
Fig. 1　Illustration of a morphological analysis

統計的手法や機械学習手法などを適用する際には，形態素解析の結果を単語文書行列という行列で表現することが多い．**第 2 図**に単語文書行列の例を示す．本稿では，単語の有無を 0–1 で表現する単語文書行列を扱う．

2.2 テキスト検索

テキスト検索の前に，あらかじめ形態素解析を行うことによって，単語同士の照合による検索が可能となるため，以下の三つの観点で，単純な文字列検索より精度の高い情報抽出が可能となる．

(1)　単語の適切な検出

想定外の単語の検出，例えば，「部品」を検索したいときに，「品質保証部品質管理課」が検出され

(a) 文　書

No.	記　載　内　容
1	クレーンで異音が発生している.
2	クレーンの動きが悪くなり異音がする.
3	横行停止用 LS が劣化により破損.
4	停止用 LS の信号が入らず.
5	クレーンが途中停止をする.
6	操作盤に電源が入らず，クレーンが停止する状態.

(b) 単語文書行列

No.	クレーン	で	異音	が	発生	する	いる	の	…
1	1	1	1	1	1	1	1	0	…
2	1	0	1	1	0	1	0	1	…
3	0	0	0	1	0	0	0	0	…
4	0	0	0	1	0	0	0	1	…
5	1	0	0	1	0	1	0	0	…
6	1	0	0	1	0	1	0	0	…

（注）　文書ごとに単語の有無を確認し，存在する：1，存在しない：0 とする.

第 2 図　単語文書行列の例
Fig. 2　Illustration of a term-document matrix

てしまうという事象を回避できる.

(2)　表記ゆれの回避

形態素解析では，各単語の原形を評価するため，「 壊れる 」「 壊れた 」といった活用形による影響を受けずに検索することが可能になる.

(3)　類似した文書の優先表示

テキスト検索で複数の文書が検索された場合，入力した文章とより類似した文書が優先して検索結果に表示されることが望ましい．TF-IDF (Term Frequency-Inverse Document Frequency)，コサイン類似度という二つの指標を用いると，入力した文章と検索された文書との類似度を定量的に評価して，類似した文書を優先的に表示できる.

TF-IDF とは，一つの文書における特定の単語の出現回数と，全文書における特定の単語が含まれる文書の件数から算出される指標である．多くの文書に出現する一般的な単語は TF-IDF が小さくなり，特定の文書にしか出現しない単語は TF-IDF が大きくなる．そのため，TF-IDF の大きい単語ほど文書を特徴づける単語であるといえる.

コサイン類似度とは，ベクトル同士がなす角度を一般化した指標である．ベクトル同士の向きが似ているほど最大値である 1 に近づく.

文書間の類似度は以下のようにして定量的に求めることができる．まず，各文書，各単語の TF-IDF を計算する．次に，類似度を計算したい二つの文書を選択し，TF-IDF を並べたベクトル同士のコサイン類似度を計算する．これが二つの文書間の類似度である.

文書間の類似度を計算して，類似した文書を検出する例を**第 3 図**に示す．この例では，文書 No. 1 における「 クレーン 」の TF-IDF は 0.025（**第 3 図 - (b)**）であり，文書 No. 1 と文書 No. 2 における TF-IDF を並べたベクトル同士のコサイン類似度は 0.13（**第 3 図 - (c)**）である．この値が文書 No. 1 と文書 No. 2 の類似度であり，文書 No. 1 に最も類似している文書は文書 No. 2 であるといえる.

このように文書間の類似度を計算することによって，入力した文章とより類似した文書を検索結果の上位に優先的に表示できる.

2. 3　機械学習による文書分類

テキスト分析と機械学習を組み合わせることによって，文書分類を行うことが可能になる.

機械学習には，利用するデータに正解の情報を付与して学習する教師あり学習 [2] と，事前に正解の情報を付与せずに学習する教師なし学習がある．本稿では，特に教師あ

(a)　単語文書行列

No.	クレーン	で	異音	が	発生	する	いる	の	…
1	1	1	1	1	1	1	1	0	…
2	1	0	1	1	0	1	0	1	…
3	0	0	0	1	0	0	0	0	…
4	0	0	0	1	0	0	0	1	…
5	1	0	0	1	0	1	0	0	…
6	1	0	0	1	0	1	0	0	…

(b)　TF-IDF の値を並べた行列

No.	クレーン	で	異音	が	発生	する	いる	の	…
1	0.025	0.111	0.068	0.000	0.111	0.025	0.111	0.000	…
2	0.022	0.000	0.060	0.000	0.000	0.022	0.000	0.060	…
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	…
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.060	…
5	0.029	0.000	0.000	0.000	0.000	0.029	0.000	0.000	…
6	0.018	0.000	0.000	0.000	0.000	0.018	0.000	0.000	…

(c)　文書間の類似度

No.	1	2	3	4	5	6
1	1.00	0.13	0.00	0.00	0.04	0.03
2	0.13	1.00	0.00	0.11	0.04	0.03
3	0.00	0.00	1.00	0.21	0.02	0.09
4	0.00	0.11	0.21	1.00	0.02	0.23
5	0.04	0.04	0.02	0.02	1.00	0.05
6	0.03	0.03	0.09	0.23	0.05	1.00

（注）　TF-IDF の値を並べた行列から 2 行を選択して，そのコサイン類似度を計算する．

第 3 図　文書間の類似度の例
Fig. 3　Illustration of degrees of similarity between text data

り学習による文書分類を扱う．

教師あり学習による文書分類では，各文書に分類したいグループの情報を付与して機械学習を行い，分類モデルを構築する．例えば，以下の手順に従って文書分類を行うことによって，製品の状況（故障事象）を表す文章から適切な処置内容を提示することができる．

(1)　学習データの準備

分類対象となる文書に対して分類したいグループを表すラベルを付与する．そして，分類対象となる文書から作成した単語文書行列と結合して，学習データを準備する．**第 4 図**に学習データの例を示す．

(2)　機械学習

単語の有無のベクトルから対応するラベルを分類できるように，単語文書行列内の各行のデータから機械学習を用いて分類モデルを構築する．分類モデルを構築する機械学習の手法には，例えば，ロジスティック回帰，決定木，ランダムフォレストなどがある．

機械学習では，学習データから自動で重要な単語を選定する．例えば，決定木を利用した文書分類では，特定の単語を含むかどうかに着目して，複数のルールを自動で選定して分類モデルを構築する．これらのルールには適用する順番も決まっており，より早く適用するルールに含まれる単語ほど，より重要な単語であると解釈できる．

ラベル分類の精度を向上させるためには，単語の選定が重要である．前述のとおり，機械学習では自動で重要な単語を選定しているが，パラメータを調整し，より良い単語を選定させることによって，精度を向上できる場合が多い．さらに，機械学習を適

(a)　文書とラベル

No.	記　載　内　容	ラベル
1	クレーンで異音が発生している.	チェーン給油
2	クレーンの動きが悪くなり異音がする.	チェーン給油
3	横行停止用 LS が劣化により破損.	LS 交換
4	停止用 LS の信号が入らず.	LS 交換
5	クレーンが途中停止をする.	基板交換
6	操作盤に電源が入らず，クレーンが停止する状態.	基板交換

(b)　単語文書行列とラベル

No.	異音	クレーン	停止	LS	発生	動き	…	ラベル
1	1	1	0	0	1	0	…	チェーン給油
2	1	1	0	0	0	1	…	チェーン給油
3	0	0	1	1	0	0	…	LS 交換
4	0	0	1	1	0	0	…	LS 交換
5	0	1	1	0	0	0	…	基板交換
6	0	1	1	0	0	0	…	基板交換

（注）　文書ごとに単語の有無を確認し，存在する：1，存在しない：0 とする.

第 4 図　学習データの例
Fig. 4　Illustration of a training data

用する前に人手で単語を選定することによって，精度の向上が期待できる．例えば，**第 4 図**に示す学習データでは助詞や「する」，「いる」といった，頻出するが重要でない単語を除外するように単語を選定すると，後に適用する機械学習の精度向上も見込める.

(3)　診　　断

新しく得られた文書，つまり，ラベル付けされていない文書に対して，形態素解析を用いて単語単位に分割して，単語の有無のベクトルを作成し，機械学習によって得られた分類モデルを適用することによって，分類されたラベルを得る.

なお，診断時において，学習データに含まれない単語は分類に利用できない．そのため，新しく得られた文書にそのような単語が多く含まれる場合は，分類精度が悪くなる可能性が高い．学習データに含まれない単語も利用して診断する方法として，例えば **4 章**に示す分散表現の利用が挙げられる.

3.　テキスト分析の適用事例

IHI グループにおけるテキスト検索の適用事例として，点検報告書作成支援ツール [3] を紹介する.

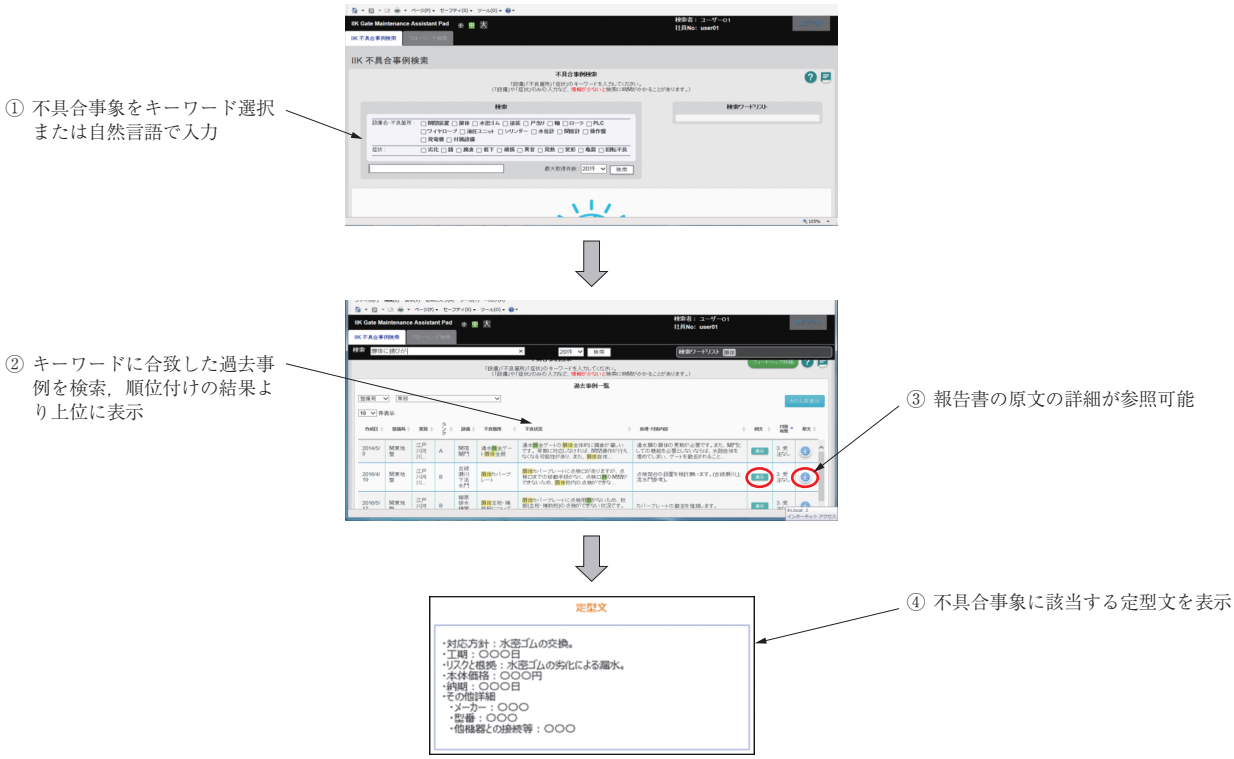
水門の点検業務では，点検時に発見した不具合事象とその事象に対する対処方法を点検報告書にまとめている.

特に経験が浅いメンテナンス員は報告書を作成するうえで，過去の経緯や類似不具合の対処方法を参考にして，基準値と比較参照しながら報告書をまとめている．そのため，書類作成に多大な時間を費やしているという問題や，経験が浅いメンテナンス員とベテランのメンテナンス員との間で報告書の質にばらつきがあるという問題があった.

そこで，報告書作成における業務負荷の低減と質の向上を目的として，メンテナンス員が文章または単語を入力すると，類似した不具合事象が記載されている点検報告書を検索する点検報告書作成支援ツールを開発した.

第 5 図に点検報告書作成の支援データの提供フローを示す．点検報告書作成支援ツールでは，事前に過去の点検報告書と辞書データを登録し，キーワードを抽出する．システム画面上でメンテナンス員が文章を入力すると，文章からキーワードを抽出し，過去の点検報告書から抽出されたキーワードと照合して，照合した際の類似度が高い検索結果を表示する．また，同時に箇条書きによる定型文を自動的に作成し，対応や方針，リスクと根拠を提示し，分かりやすい報告書ができるようメンテナンス員を支援する.

このツールを利用することによって，報告書作成に掛かる時間を短縮することが可能になった．また，経験が浅いメンテナンス員でもベテランのメンテナンス員と同一品質



第 5 図　報告書作成の支援データ提供フロー
Fig. 5　Procedure of providing support data for creating reports

の報告書を作成することが可能になった．

4.　精度向上・展開に向けた取組み

4.1　現状の問題点

2 章に示した単語単位で照合を行う基本的なテキスト分析技術には，主に二つの問題点がある．

(1)　表記ゆれ

例えば，「 フィードバック 」と「 F/B 」といった表記ゆれがある場合は，内容が同じでも記載が異なるため，別の単語として扱われてしまう．そのため，「 フィードバック 」と「 F/B 」は同じ意味であることを示すデータ（ 類義語辞書データ ）を別途準備することによって表記ゆれに対応する．

また，「 損傷する 」と「 破損する 」は，意味が似ている単語であり，同一の概念をもつ単語としてグルーピングする方が，テキスト検索や文書分類の精度向上が期待できる．

しかし，人手による類義語辞書データの準備や，単語をグルーピングする作業は，単語量が多くなったときに作業負荷が非常に大きくなるという問題がある．

(2)　テキストデータ内の情報不足

人が自由に記述した文章の多くは，文を構成するための情報が省略されている．

例えば，「 装置 A がうまく動作していなかった．修理が必要である．」という文章があったとき，2 文目では修理の対象は不明である．また，修理を担当している企業や人物をこの文章からは判断できない．そのため，お客さまからこのような修理依頼があった際，テキスト分析技術を用いて自動で処理しようと思っても，修理の依頼先を適切には判断できない．

人は文脈を理解したり，別の情報を基にテキストデータに不足している情報を補完したりすることによって，文章の意味を解釈している．しかし，**2 章**に示した単純な単語照合だけでは，このような高度な情報処理は実現できないという問題がある．

そこで，以上の問題点を解決するために，分散表現の利用，および照応解析と知識グラフを用いた情報付加の二つの取組みを紹介する．

4.2　分散表現

表記ゆれに対応するためには，例えば「 損傷 」と「 破損 」という単語について，単語の意味が近いということが表現できればよい．そこで，単語を数値ベクトルに置き換えることができれば，単語同士の近さを計算できる．

単語を数値ベクトルとして扱うために広く扱われているものは，ある単語の意味は，周囲の単語（ 文脈 ）によっ

て形成されるというアイデアである．このアイデアに基づき単語をベクトル表現する手法を，単語の分散表現[4]という．

分散表現を実現するためのツールである word2vec では，ある単語を出力とし，学習データに含まれる文におけるこれらの単語の前後にある単語を入力として，ニューラルネットワークモデルによって学習を行い，数値ベクトルの要素を算出する．

大規模な日本語文章データセットを用いて単語の数値ベクトル表現を学習[5]させ，例として「損傷」という単語とコサイン類似度が近い単語群を抽出した結果を**第 6 図**に示す．**第 6 図**から，「破損」の類似度が最も高く，「損傷」と「破損」は意味が近いことを判断できており，分散表現を用いることが有効であることを確認できた．

分散表現を利用した高度なテキスト分類を実現するために，株式会社エヌ・ティ・ティ・データ（NTT データ）から提供された AI 技術である corevo® を活用している．corevo を活用することによって，上述の例に示した「損傷する」と「破損する」といった同一の概念をもつ単語同士の照合が可能になっており，単純な単語照合だけを用いた場合より分類精度が向上している．

4. 3　照応解析と知識グラフを用いた情報付加

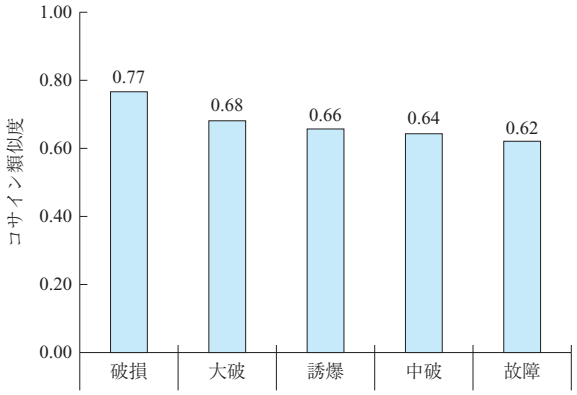
テキストデータ内の情報不足に対応するためには，照応解析や知識グラフを用いた情報付加の手法などが有用である．

照応解析とは，代名詞や指示詞などの指示対象を推定する処理のことであり，特に省略された名詞句を求める処理をゼロ照応解析という．例えば，「装置 A がうまく動作していなかった．修理が必要である．」という文章において，2 文目には「装置 A の」という言葉が省略されている．照応解析を用いることによって，このような省略された単語を推定することが可能になる．

また，「修理が必要である．」という文章だけでは，修理をだれが行うべきかを判断することはできないが，あらかじめ概念や概念同士の関係性を表現した知識グラフに情報を蓄積しておくことによって判断が可能になる．**第 7 図**に知識グラフの例を示す．

この例では，業務の担当者などを知識グラフとしてあらかじめ保有しておくことによって，「修理が必要である．」という文章が入力された場合に，修理を担当しているのは D 会社であるということが推定できる．

これらの技術を利用することによって，「装置 A がうまく動作していなかった．修理が必要である．」といったお客さまからの修理依頼に対して，修理方法を適切に提案することが，自動的に行えるようになると期待できる．

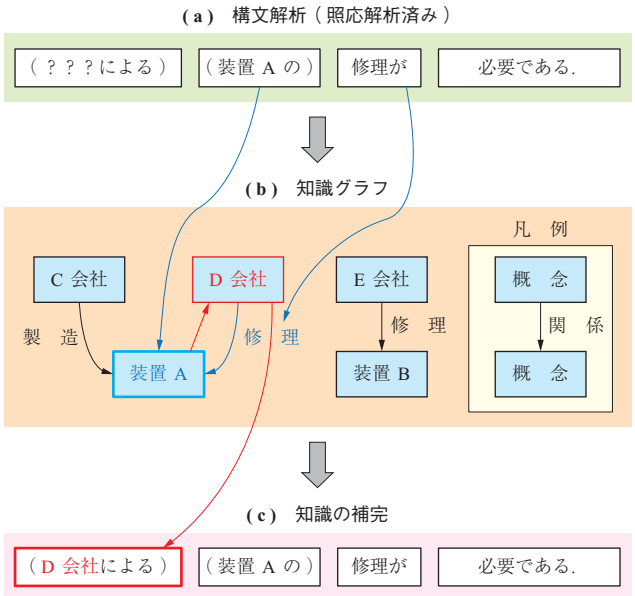


第 6 図　word2vec で計算した単語「損傷」との類似度
Fig. 6　Degrees of similarity between “damage” and other words calculated by word2vec

5.　結　　言

本稿では，テキスト分析技術について，過去に蓄積された文書を効率的に検索するためのシステムを，事例を示しながら紹介した．テキスト分析技術を利用することによって，過去に蓄積されたテキストデータから有益な情報を抽出することが可能になり，業務の効率化や知識伝承の実現が期待できる．

IHI グループでは，お客さまに納めた製品の状態や稼働データを蓄積し，データ解析を行い，保守サービス支援などのお客さま価値の創出に活用できる IoT プラットフォー



第 7 図　知識グラフの例
Fig. 7　Illustration of a knowledge graph

ム ILIPS (IHI group Lifecycle Partner System) を保有している．テキスト分析技術も ILIPS 上に搭載しており，例えば，製品の稼働データと設計図書やメンテナンス記録などのテキストデータの分析結果などを合わせることによって，より付加価値の高いサービスの提案が可能になると期待できる．今後もテキスト分析技術の高度化を進め，お客さま価値の向上に貢献していく．

参 考 文 献

(1) 黒橋禎夫：改訂版 自然言語処理，NHK 出版，2019 年

(2) C. M. ビショップ著，元田　浩，栗田多喜夫，樋口知之，松本裕治，村田　昇訳：パターン認識と機械学習 上，丸善出版，2012 年

(3) 熊谷公雄：水門点検支援システム『 GBRAIN 』ジーブレイン ― 点検業務の質の向上・効率化に向けて ―，JACIC 情報，第 119 号，2019 年，pp. 51 － 56

(4) 坪井祐太，海野裕也，鈴木　潤：機械学習プロフェッショナルシリーズ 深層学習による自然言語処理，講談社，2017 年

(5) 柳井孝介，庄司美沙：Python で動かして学ぶ自然言語処理入門，翔泳社，2019 年