機械学習を用いた文章データの解析・可視化技術

専門知識がなくても膨大な文書から有用な情報・知識を自動で抽出

IoT（Internet of Things）やビッグデータ技術の進展で情報量が飛躍的に増加しており，企画・設計・製造からサービスに至るまで生産活動のあらゆるフェーズで，いかに価値のある情報を素早く探し出し，有効に利活用するかが課題になっています。

東芝は，設計書類や技術文書といった定型化されていない電子テキストから，専門知識なしで有用な情報・知識を自動で抽出して可視化する技術を開発しました。設計や調達業務へ適用した場合，過去の設計事例やノウハウの取得や，調達戦略への活用などにより，業務の効率化や高度化への効果が期待できます。

情報・知識獲得の効率化に向けて

設計関連部門では，設計開始時点において，過去の設計事例や，使用部品，不具合事例などの検索に業務の約40％を費やしていると言われています。また製品機能の高度化に伴い，高い専門性と幅広い技術領域をカバーすることが求められますが，必要な情報が広範囲にわたって保管されており，欲しい情報がどこに在るか分からなかったり，また分かっても欲しい情報がどの文書に載っているかを探すために膨大な時間が掛かったりしています。

検索時間の短縮や，情報が見付からないために発生する後戻りの抑制が喫緊の課題になっていますが，過去分を含めた大量の文書や，失敗・優良事例集，関連する論文などを読み解いて活用するには限界があります。

東芝は，このような課題を解決するため，機械学習を利用して短時間で膨大な文書から単語間や文章との関連性を自動で取得し，その関係性を可視化する技術を開発しました（**図1**）。

テキスト解析技術

検索する電子テキストの形式や文書の言語に制約を設けないため，テキスト解析のコアエンジンとして，言語を問わず自然言語文章の解析が可能で，かつオープンソースであることからツールへの組み込みが容易なword2vec[1]を選定しました。その上に単語間や文章の関連性を自動で取得し，可視化する環境を構築しました。

word2vecは，多数の文書をリファレンスとして，文章中の単語同士の関連性を，注目単語からの距離やその単語の近くに出現する頻度などに着目し，機械学習により単語を多次元の特徴ベクトルとして表す手法です。似た意味を持つ単語は特徴量が似ており，特徴空間の中で近い位置に配置されることから，距離を指標に意味の近い単語を自動で抽出します（**図2**）。

この手法では，前処理として学習に利用する文章データを品詞分割（形態素解析）する必要があります。その際に，抽出したい情報・知識と関連性が薄い品詞（助詞や冠詞など）を除去する機能や，人の情報を有効利用するために名字と名前を連結する機能などを組み込んでいます。

情報・知識の連鎖的検索と可視化技術

word2vecの基本的機能は，単語を数値化した辞書を生成する機能と，与えた単語に近い意味の単語（複数）を表示する機能の二つですが，学習に必要な文書の前処理を含めて計算機のコマンドとして実行する必要があり，言語解析や計算機を利用するための専門知識が求められます。また，表示された解析結果を保存できませんでした。

そこで今回，専門的な知識がなくても文書ファイル（Excel®・PowerPoint®ファイル，テキストファイル）を用意してアップロードさえすれば，文書から情報・知識を自動で抽出できる環境を構築しました（**図3**）。特長として，関連性が高いとして抽出された単語群をグラフィカルに表示し，表示された単語を次々にクリックしていくことで，連鎖的に関連性の高い単語を検索して表示することができます。

適用事例

この環境を使用して情報・知識を抽出し可視化した二つの事例について述べます。

数百ページ相当のオーブンな記事を学習した事例1では，調達分野だけに関連するユーザーやサプライヤーを抽出でき，またカテゴリーも自動で分類できていることを確認しました（**図4**）。

また製品種別ごとの報告書を使用した事例2では，PowerPoint®形式の数十ファイルを学習に使用して連鎖的検索を実施したところ，現象名から関連性の高い部品や要因が抽出できることを確認しました（**図5**）。それぞれの情報・知識の抽出作業は，10分以内で完了できました。

今後の展望

この技術は，開発設計・調達・生産から物流・保守・営業に至る各フェーズでの情報・知識の獲得に広く活用でき，業務効率・生産性向上に寄与することが期待されます。今後，各フェーズ間の情報をつなげて活用する技術の一つとして発展させ，実務への展開を図っていきます。

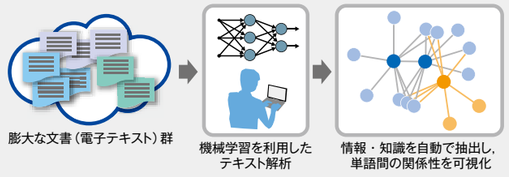
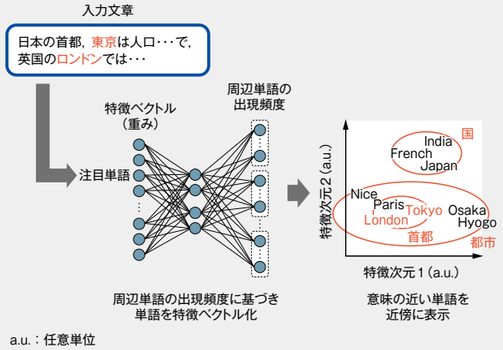


図1．機械学習による文章データの解析・可視化技術の概要 ― 膨大な文書群から情報・知識を抽出し，関係性を可視化します。



a.u.：任意単位

図2．word2vecの概要 ― 文書の形式や言語の制約を受けることなく，文章内の単語を多次元の特徴ベクトルで表し，意味の近い単語を自動で抽出する手法です。

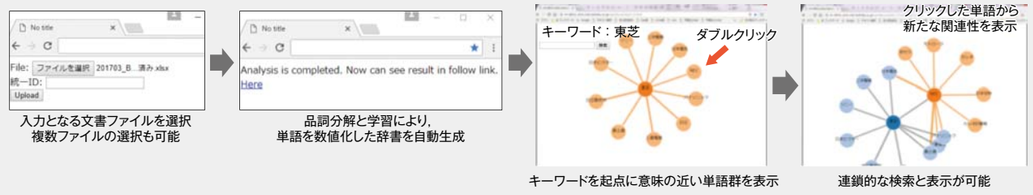


図3．情報・知識を自動抽出できる環境の構築 ― 言語解析や計算機の専門知識がなくても，文書ファイルをアップロードするだけで解析とその結果を可視化できます。

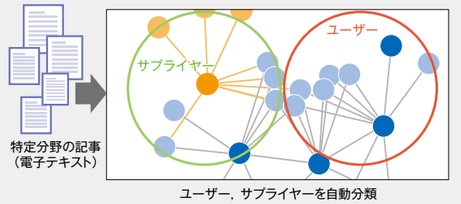


図4．事例1 ― 調達分野だけに関連した業者の抽出とその分類を自動で実施することができました。

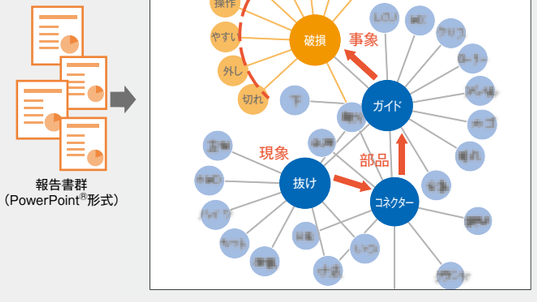


図5．事例2 ― 報告書を用いて学習し，連鎖的な検索を行った結果，現象の起点から要因まで抽出することができました。

文　献

(1) Mikolov, T. et al. "Distributed Representations of Words and Phrases and their Compositionality". Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013)-Volume 2. Lake Tahoe, Nevada, USA, 2013-12, NIPS. Curran Associates, Inc., 2013, p.3111－3119.

池田　弘行
研究開発本部 生産技術センター
設計生産システム変革推進部