大量の文書の内容を様々な観点で把握できる特徴語抽出技術

情報に埋もれた特徴を浮き彫りにする

企業内に大量に蓄積される製品の不具合情報や他社も含めた特許情報などの中から，注目すべき情報を見つけ業務に役だつ知見を得るには，大量の文書を様々な観点で捉えて，内容の特徴を明らかにすることが必要です。

東芝ソリューション(株)は，大量の文書情報の中からユーザーが着目する部分について，その内容の特徴を表す単語や，別の部分との内容の違いを表す単語を自動的に抽出する，特徴語抽出技術を開発しました。この技術によって，製品の不具合情報を基に部品や製造時期の観点で原因を把握したり，特許情報を基に各社の注力技術の違いを明確にすることができます。

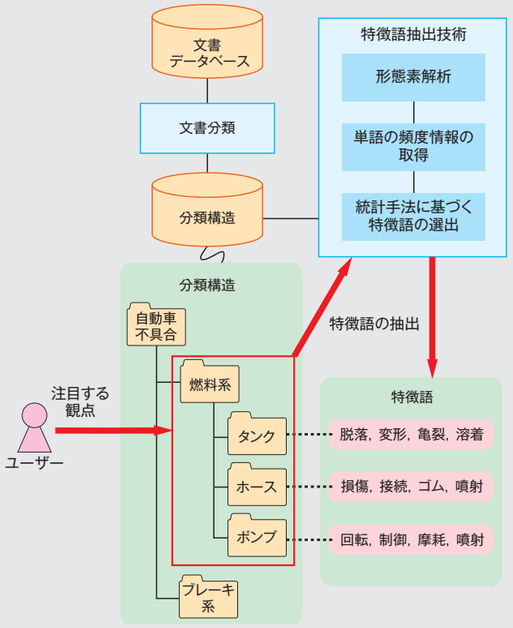

図1．特徴語抽出技術の構成 — 階層的に分類された構造を利用して，様々な観点から特徴語を抽出することができます。

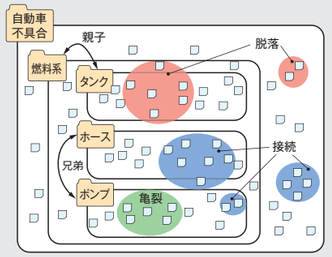

図2．特徴語として適切な単語の抽出 — 兄弟カテゴリーの中で特定のカテゴリーに偏って出現する“脱落”，“接続”，“亀裂”という単語を特徴語として抽出します。

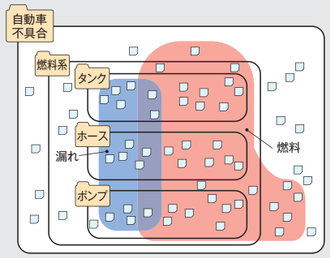

図3．特徴語として不適切な単語の抽出 — 親カテゴリーで一般的な単語の“燃料”や，兄弟カテゴリーで一様に出現する単語の“漏れ”は特徴語とはしません。

図4．電子レンジ関連特許の分析例 — 大量の特許文書に対して，各社で共通する技術や，特有な技術といった特徴を分析し，一目で把握することができます。

大量の文書を活用する手がかりとなる特徴語

企業には大量の文書情報が日々蓄積されており，これらの情報を業務に積極的に活用したいというニーズが高まっています。これらの情報について，個々の内容を確認するだけでなく，蓄積された情報を様々な観点から分析することで，業務の品質や効率を高めるための知見が得られます。

例えば，製品の不具合情報では，部品ごとに不具合を分析することで，各部品に特有な不具合の特徴を把握でき，原因の特定と迅速な対策が可能になります。また，同じ部品に関する不具合でも，異なる製造時期で比較することで，各時期の製造上の問題点が明確になり，注意すべき時期や，改善すべき工程などを発見できます。

従来の技術では，このような分析のために，文書の中で多用されている単語を抽出することで，大量の文書の中で代表的な内容を示します。しかし，頻度の高い単語に埋もれがちな特徴までは示すことができません。前述の例のように，部品という観点で詳細を調べたり，製造時期の観点で比較したりするには，それぞれの目的に適した単語(特徴語)の抽出が必要になります。更に，着目する観点は分析の目的や文書の種類によって様々に異なるので，この観点をユーザーが自在に設定できる枠組みが必要です。

統計的手法に基づく特徴語抽出

東芝ソリューション(株)は，ユーザーが着目する観点に従って，特徴語を自動的に抽出する特徴語抽出技術を開発しました(**図1**)。この技術では，様々な観点で階層的に分類された構造を利用します。ユーザーは，この分類構造の部分を指定する形で着目する観点を設定します。 図1では，自動車の燃料系の不具合を分類したカテゴリーに，ユーザーが着目したと仮定します。この観点に対して，タンクやホースなどの部品に対応する子カテゴリーから，特徴語をそれぞれ抽出します。この場合，以下に示すような特徴語がユーザーにとって有用と考えられます。

(1) 燃料系に対するタンクのような，親カテゴリーに対する子カテゴリーの特徴がわかる単語

(2) タンクやホースのような兄弟カテゴリー間での相違点がわかる単語

このような特徴語を自動的に抽出するために，カテゴリーの階層構造に基づいた統計的手法を用います(**図2**，**図3**)。

例えば“問題”などの単語は，不具合の記述で一般的に用いられており，不具合の内容を表してはいません。また，“燃料”などの単語は，燃料系におけるほとんどの不具合の記述に用いられ，タンクなどの部品に特有の不具合を表してはいません。(1)の点で，このような不適切な単語を除き，適切な特徴語を得るためには，統計的な頻度の偏りを表すT統計量を用い，親カテゴリーに対して子カテゴリーに偏って多く出現する単語を選別します。例えばタンクという部品に対して，“脱落”などの特徴的な単語だけが抽出できます。

また，“漏れ”という単語は，タンクやホースなど，多くの部品で生じる不具合を表し，部品ごとの相違点を表すものではありません。ホースの“接続”や，ポンプの“亀裂”のように，漏れの具体的な原因や箇所を表す単語のほうが，(2)の点では適切です。このような単語を優先して抽出するために，分散分析の手法を用いました。これにより，兄弟カテゴリーの中で特定のカテゴリーに偏って出現する単語を自動的に選別します。

特徴語を用いた分析のメリット

この技術は，不具合情報や，新聞記事，特許文書など，様々な種類の文書に適用できます。これらの文書での抽出結果を評価したところ，多くの場合，自動抽出した上位10個の単語のうち，8個程度が特徴語として有用でした。TF-IDF (Term Frequency - Inversed Document Frequency) 法などのように，カテゴリーの階層構造を考慮せず単純に文書中での出現頻度に基づいて特徴語を抽出する従来手法に比べて，この技術の手法では10～20％ほど精度良く抽出できます。

また，クロス集計などの機能と組み合わせることで，カテゴリーの特徴語を可視化して分析し，価値のある知見を得ることができます。特許文書の分析にこの技術を適用した例を**図4**に示します。各セルの数字は，横軸の企業の出願特許の中で，縦軸の特徴語を含む特許の件数です。例えば，いちばん左上のセルの“14”は，A社が出願した特許の中で，“表示”という単語を含む特許が14件あることを示しています。また，セルの値が大きいほど背景色を濃く，小さいほど淡くしています。

これにより，特許を出願した企業ごとに技術の内容を表す特徴語を抽出することで，複数の企業が取り組んでいる共通の技術や，少数の企業が注力している技術の傾向が，一目で把握できます。更に，発明者や出願年度など，任意のカテゴリーについて特徴語を抽出することができます。

今後の展望

今後は，この特徴語抽出技術を洗練し，精度を向上させるとともに，様々な分析技術や可視化技術を組み合わせることで，大量の文書を効率よく活用するための製品やソリューションを提供していきます。

岩崎　秀樹
東芝ソリューション(株)
IT技術研究所
研究開発部