大量文書の内容が一目でわかる代表文生成技術

文書群特有の内容を文で表現

製品不具合の調査において，対策が必要な不具合を見つけ，対策を立案するためには，大量に蓄積された不具合情報を人が読んで内容を理解しなければならず，手間や時間がかかっています。

このような問題を解決するために，東芝ソリューション(株)は，大量文書の内容が一目でわかる代表文生成技術を開発しています。この技術により，“亀裂部から燃料が漏れる”，“燃料タンクが脱落する”，“ホースの材質が不適切”など，文書群特有の内容を表す文，すなわち代表文を生成することができます。これらから対策立案の必要性が高い不具合がわかり，的確な対策をすばやく立案できます。

大量文書の活用におけるニーズ

企業内では，業務を通じて収集及び作成された情報が大量に蓄積されています。従来は，こうした大量の文書を，製品名や年月などの属性値やキーワードを基に，いくつかのグループ(文書群)に自動分類していました。しかし，属性値やキーワードだけでは，文書群にどのような内容が記述されているか十分に把握できないため，人が全部の文書を読んで内容を理解しなければならず，時間がかかります。そこで，文書群の内容をすばやく把握したいというニーズが出てきています。

文書群特有の内容を把握できる代表文生成技術

東芝ソリューション(株)は，文書群の内容を効果的に把握することを支援するために，文書群特有の内容を表す文を生成する代表文生成技術を開発しています。これは，複数の文書群から，文書群の内容を表す文を生成し，文書群特有の内容を表す文を選択する技術です(**図1**)。

●文書群の内容を表す文の生成

例えば，自動車の不具合情報の分析では，**図2**(1)に示すように，複数の文書群の中の“亀裂が発生する”という内容は，従来，“亀裂”というキーワードで表現されていました。しかし，キーワードだけでは，“亀裂部から燃料が漏れる”，“亀裂により異音が発生する”などの内容を把握することができません。内容を適切に把握するために，“亀裂が発生する”といった文で内容を表現する必要があります。

文書群の内容を表す文を生成するためには，図2において，まず，複数の文書群(1)に含まれるすべての文を構文解析し，その結果(2)に抽出パターン(3)を適用し，木構造形式で表現された内容(4)を複数抽出します。そして，木構造形式の内容を用いて，内容を表す文，すなわち代表文候補(5)を複数生成します。抽出パターンは，主語，述語という，文の骨格として重要な単語と，単語間の構文的関係を表したものです。

●文書群特有の文の選択

図2(5)の代表文候補に示す例では，内容が“亀裂”と表現された文書群から①“不具合が発生する”，④“亀裂部から燃料が漏れる”などの代表文候補が，更に上位の“燃料タンク”文書群からも⑪“不具合が発生する”などの代表文候補が生成されます。このように，“不具合が発生する”という文は，両方の文書群に含まれるため，“亀裂”という文書群に特有な内容を表す文としては，“亀裂部から燃料が漏れる”のほうが適切です。このような文書群特有の内容を表す文を，上位などの文書群から生成される文との比較に基づき，代表文候補から選択します。選択基準として，複数の文書群に共通する代表文候補の出現度合いを定義した量(相互情報量)を採用しました。相互情報量は，代表文生成対象となる文書群だけでなく，ほかの文書群内の文書の統計情報も用いて算出するものです。相互情報量を用いると，“亀裂”の文書群特有の内容を表す文として“亀裂部から燃料が漏れる”を選択することができます(図2(6))。

このように，抽出ルールと相互情報量を用いて，文書群特有の内容を表す文を生成できます。

代表文から得られる情報

蓄積された大量の文書を活用する例として，製品サポート部門において，大量の不具合情報の中から対策の必要性が高い不具合を見つけて，その対策を立案する場合があります。

この作業では，まず，不具合情報の分類で得られた個々の文書群について，文書群の名前を見て全体傾向を把握します。例えば，不具合件数の多い“亀裂”文書群に注目します(**図3**(1))。この文書群の代表文を見て，文書群内の不具合の内容を把握することで，対策立案の必要性が高い不具合を探します(図3(2))。着目した代表文に対応する文書を読んで，必要性が高い不具合に関する詳細な情報を得ることができます(図3(3))。

例えば，“亀裂部から燃料が漏れる”に対しては部品の品質改良や，“燃料タンクが脱落する”に対しては組立工程の改良といった対策の立案が可能です。

このように，代表文によって，短時間で容易に，対策が必要な不具合を見つけ，対策立案に関する情報も収集し，迅速に対応することができます。

今後の展望

今後，代表文生成技術を，企業の情報と知識の利活用を促進する対話型文書分類システム(この号のp.60－63参照)へ搭載する予定です。

また，特許マップの作成にもこの技術を適用していきます。従来，キーワードで表現されていた特許文書群の名前を代表文で表現することにより，特許文書群の内容を効率的に把握しやすい特許マップの作成を支援します。

倉田　早織

東芝ソリューション(株)
IT技術研究所　研究開発部

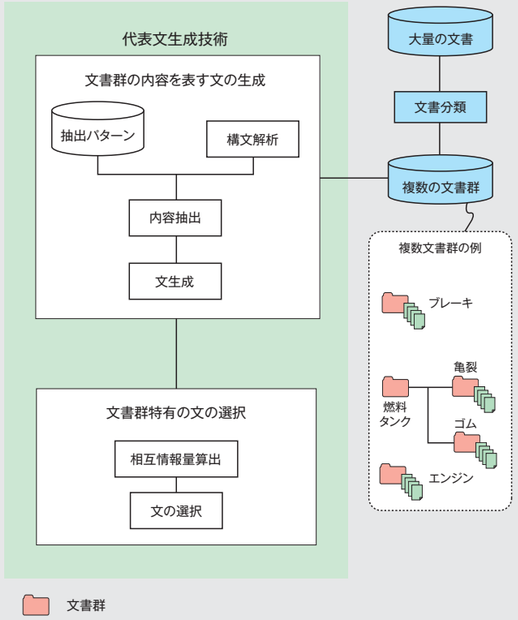

図1．代表文生成技術の構成 ― 代表文生成技術は，文書群の内容を表す文の生成と，文書群特有の文の選択から構成されます。

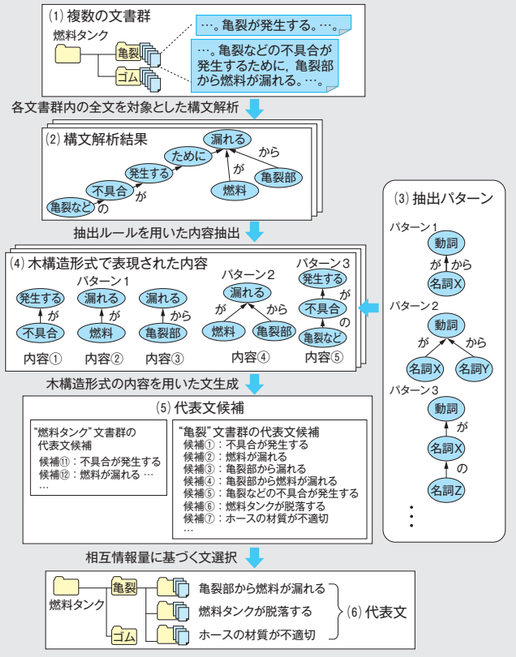

図2．代表文生成の処理の流れ ― 文書群の内容を表す文を生成し，生成した文から文書群特有の文を選択して代表文とします。

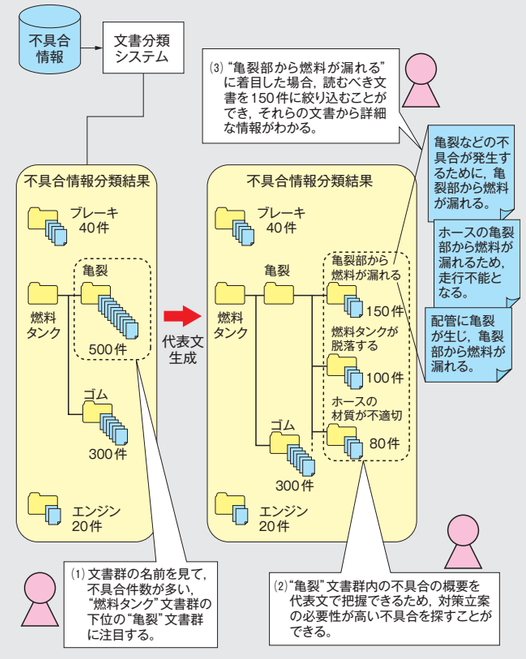

図3．文書分類システムへの適用 ― 代表文があると，短時間で容易に，文書群特有の内容を把握でき，読むべき文書を絞り込むことができます。