ユーザーの意図を反映した対話型文書分類技術

システムとの協調作業で，誰でも，簡単に，大量の文書を分類できる

情報を効率良く整理し，新しい知見を得るという目的で，大量の情報を有効活用するための文書分類技術が求められています。しかし，従来の全自動の分類システムでは，ユーザーの意図を解釈してそれを十分に反映した分類結果になっていません。

東芝ソリューション(株)は，ユーザーとの協調作業を通じて，システムが解釈したユーザーの意図に基づいて文書を分類する，対話型文書分類技術を開発しています。長期にわたって蓄積されるクレーム情報を継続的に分析し，既存の構造に分類できない新しいクレーム情報をも発見できます。

文書分類のニーズと課題

企業内に蓄積された様々な情報を有効に活用するには，内容や用途に応じて情報を適切に整理する必要があります。また，製品のクレーム情報の分析では，問題の発生頻度を調べたり，新しい問題を発見したいというニーズもあります。

企業の情報は大部分が文書であり，その内容に基づき自動的に文書を分類する必要がありますが，従来のシステムには以下の問題があります。

(1) 全自動で分類すると，意図した結果が得られないことがある。
(2) 大量の文書の分類には，専門的なスキルと労力を要する。
(3) 過去に作成した分類構造では，新しい情報を分類できないことがある。

大量の文書を分類する方法は目的によっていく通りも考えられますので，システムはユーザーの意図を解釈したうえで，自動分類を行う必要があります。一方，ユーザーは最初から明確な意図を持っているとは限らず，多くの場合，分類作業を進めながら自身の意図を明確化していきます。したがって，システムには，分類作業を通じてユーザーの意図を分類構造に反映させていく仕組みが必要です。

対話型文書分類システム

東芝ソリューション(株)は，ユーザーとシステムの協調作業を通じて分類の意図を明確化し，これに基づいて文書を分類する，対話型文書分類技術を開発しています。このシステムは，次の3段階の処理で構成されています(**図1**)。

(1) システムによる自動分類
(2) ユーザーによる分類構造の編集
(3) システムによる再分類

これらの処理を繰り返すことで，ユーザーの意図どおりの分類構造を効率良く作成し，継続的に改良，保守していくことができます。各処理の概要を以下に述べます。

システムによる自動分類

大量の文書の分類には多大な労力がかかります。数千から数万の文書を手作業で分類することは現実的でありません。また，どのような観点で分類すべきかを決めるために，ユーザーは文書群の内容をおおまかに知る必要があります。

このシステムでは，主に初期の分類作業を支援するため，クラスタリングの手法を用います。クラスタリングとは，大量の文書の中から内容が類似した文書どうしのまとまりを自動的に発見する手法です。ユーザーはその結果を用い，例えば文書数の多いカテゴリーを調べることで，文書群の全体像を把握できます。

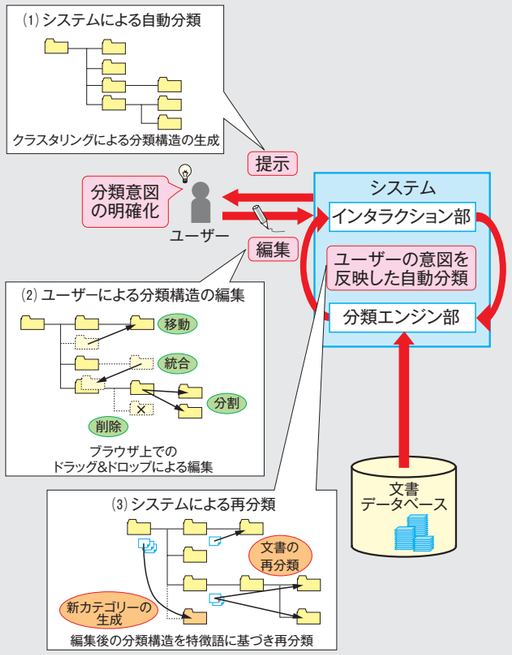

図1. 対話型文書分類システムの構成 — ユーザーとシステムの協調作業を通じて明確化したユーザーの意図を用いることで，分類構造を洗練します。

ユーザーによる分類構造の編集

クラスタリングによる自動分類の結果は，分類作業の手がかりとしては有用ですが，ユーザーの意図に合っているとは限りません。例えば，クレーム情報を分類する場合，製品別にすべきか，故障の原因別にすべきかをシステムは判断できません。そこで，以下の二つの枠組みによって，ユーザーが意図する分類構造を容易に作成できるようにします。

(1) システムによる分類の根拠の提示　システムは，生成したカテゴリー内の文書に多く含まれる特徴語をユーザーに提示します。これによりユーザーは，カテゴリーが適切かどうかを判断できます。
(2) 分類構造の編集　ユーザーは，システムが生成した分類構造を自由に編集できます。カテゴリーの作成，削除，移動，及び統合や，文書の移動などが可能です。

ユーザーインタフェースの例を**図2**に示します。ユーザーは，ブラウザ上でのドラッグ＆ドロップなどの操作によって，分類構造の確認や編集ができます。

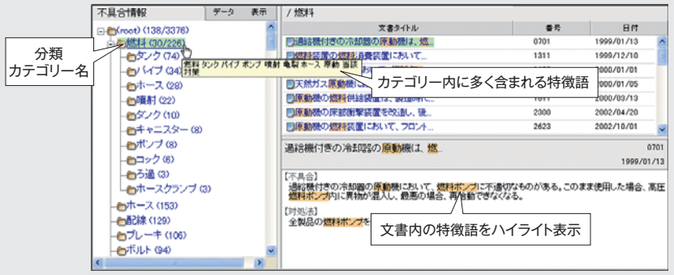

図2. ユーザーインタフェースの例 — ユーザーは，ブラウザ上でのドラッグ＆ドロップなどの操作により，分類構造の確認や編集ができます。

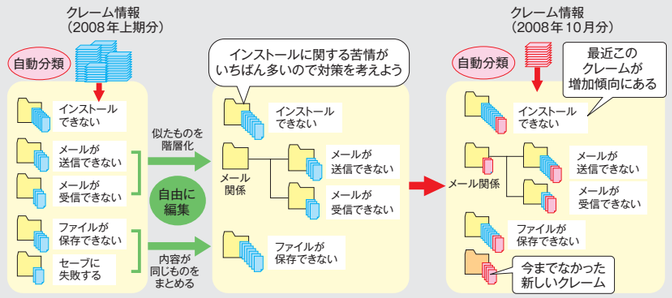

図3. 分類作業の流れ — 既知の内容のクレームは，ユーザーの意図を反映した既存の分類構造に従って分類され，新しい内容のクレームに対しては新たなカテゴリーが自動生成されます。

システムによる再分類

ユーザーが行う分類構造の編集操作には，ユーザーの意図が含まれています。例えば，二つのカテゴリーを統合した場合，両カテゴリーの共通点を見つけ，これらを統合すべきと判断して操作を行ったと解釈できます。システムは，このような解釈に基づき，ユーザーの意図を反映した文書の再分類を行うので，統合前の両カテゴリーに共通の特徴語を重要であると解釈して，このような語を持つ文書を統合後のカテゴリーに再分類します。

また，ユーザーの意図を反映した分類構造は，長期的に利用できることが望まれます。例えば，日々蓄積されるクレーム情報を分析する場合，このシステムを用いると，既知の内容のクレームは既存の分類構造に従って分類される一方で，新しい内容のクレームに対しては新たなカテゴリーが自動生成されます(**図3**)。ユーザーは，長期にわたって蓄積される情報を継続的に分析しつつ，新しい情報をも発見できます。

今後の展望

今後は，クレーム分析，特許調査，及び製品の不具合情報の分析などにこの技術を適用していきます。例えば特許調査では，発明の内容に従って分類構造を作成し，これに基づいて定期的に特許を分類することで，出願の傾向や新しい技術の内容を知ることができます。

この対話型文書分類技術を用いることで，誰でも，簡単に，大量の情報を整理，分析でき，そこから得た知識を種々の業務に生かせるようになります。様々な適用を通じてこの技術を洗練し，情報資産を積極的に利活用するための基盤技術へと発展させていきます。

宮部　泰成

東芝ソリューション(株)
IT技術研究所
研究開発部