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

Interactive Document Classification System to Accelerate Information and Knowledge Utilization

後藤　和之　　平　　博司　　宮部　泰成
■ GOTO Kazuyuki　■ TAIRA Hiroshi　■ MIYABE Yasunari

企業内の大量の情報を整理して活用するために，文書を内容に応じて自動分類する技術の必要性が高まっている。しかし，文書を分類する観点は様々であり，一様に自動分類するだけで有用な分類結果を得ることは難しい。

東芝ソリューション（株）は，数万件規模の文書をユーザーとシステムとの協調作業によって対話的に分類するシステムを開発した。ユーザーは，様々な観点や手法によってシステムに自動生成させた分類結果を自由に組み合わせることで，目的に合った分類構造を効率よく作成できる。日々増加する文書にも柔軟に対応でき，顧客のクレームを内容に応じて整理する作業や，自社と他社の特許の傾向を分析する作業などに活用できる。

Document classification technologies are expected to provide a solution to the need for effective use of large amounts of corporate information and knowledge. However, one-way automatic classification methods have been insufficient to generate classification structures suitable for the various purposes and viewpoints of users.

Toshiba Solutions Corporation has developed an interactive document classification system. This system supports the construction of classification structures by automatically generating categories based on various viewpoints and methods, and arranging them in an interactive manner. The classification results can be continuously refined to handle increasing volumes of documents, and flexibly applied to activities such as the sorting of customer claims and analysis of patent information.

1　まえがき

企業内に大量に蓄積された情報を，業務に役だつ知識として積極的に活用するためには，文書情報を分類し，その内容を分析する活動が不可欠である。例えば，顧客から日々寄せられるクレーム情報を分類した結果の各々に対して回答を用意しておくことで，迅速かつ的確に対応できる。また，苦情の原因となった製品の不具合の発生傾向を調べることで，いち早くリスクを回避できる。更に，様々な顧客の要望を分析することで，新製品の開発に顧客のニーズを反映させることができる。

しかし，数千，数万という規模の文書を人手で分類するには多大な労力を要するため，これを自動化する文書分類技術が必要となる。東芝ソリューション（株）は，企業の知識経営を支援する“情報知識利活用技術”[(1)]の一つとして，この文書分類技術の開発に取り組んでいる。応用として，特許情報を自動分類することで調査作業を支援するシステムを実現した[(2)]。

今後，更に多くの業務支援ソリューションに文書分類技術を適用するには，次の課題を解決する必要がある。

（1）　専門的なスキルを持たないユーザーでも，少ない労力で簡単に分類作業ができること。

（2）　情報の整理，調査，分析など多様な目的に合った分類構造，つまり文書の入れ物であるカテゴリーとその階層構造を作成できること。

（3）　日々増加する文書に対しても，継続的に分類構造を改良し保守していくことができること。

ここでは，このような課題を解決するために，当社が開発した“対話型文書分類システム”の機能と効果について述べる。

2　対話型文書分類システム

大量の文書を分類する観点は目的に応じて様々であるため，分類の方針を定めずに分類システムに全自動で分類させても，ユーザーの目的に合った結果が得られるとは限らない。一方ユーザーは，文書の全体像を把握できて初めて分類の方針を決定できるが，大量の文書の内容を一つずつ調べるのは労力を要する。分類作業の効率と分類結果の品質を両立するには，ユーザーとシステムの対話的（インタラクティブ）な協調作業によって，分類構造を洗練させていく枠組みが必要である。当社はこの考えに基づき，対話型文書分類システムを開発した。その構成を**図1**に，画面例を**図2**に示す。このシステムでは，次の手順で分類作業が行われる。

（1）　分類構造の自動生成　　分類作業の初期段階では，すべての文書を対象に，分類構造をシステムに自動生成させる。ユーザーは，この結果を手がかりに文書の内容の全体像を把握し，分類の方針を決めていくことができる。

（2）　分類構造の編集　　ユーザーは，システムによる自動

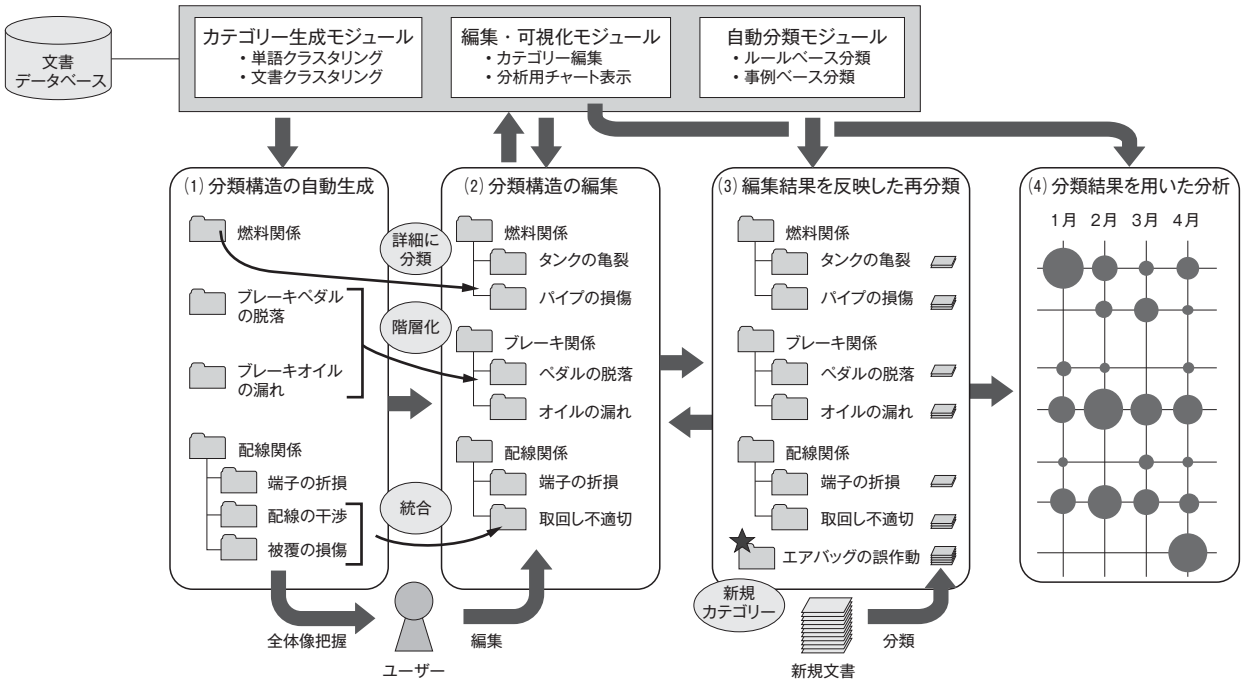

図1. 対話型文書分類システムの構成 — ユーザーとシステムの協調作業によって，分類構造を作成し，対話的に洗練していく枠組みを提供する。
Interactive document classification system

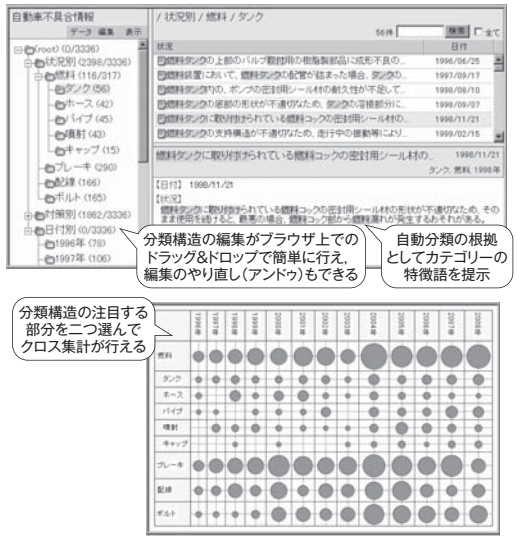

図2. 対話型文書分類システムの画面例 — ユーザーはWebブラウザ上の簡単な操作によって，大量の文書を分類し，その結果を分析できる。
Example of user interface display

分類結果のうち，利用できる部分はそのまま利用し，修正すべき部分についてはカテゴリーの移動，統合，削除などを行って，目的に合った分類構造を作り上げていく。着目するカテゴリーを掘り下げて分類することもできる。

(3) 編集結果を反映した再分類　　ユーザーによって編集された分類構造に従って，システムは文書を再分類する。例えば，ユーザーが二つのカテゴリーを統合すると，両カテゴリーに共通する特徴を持つ文書が統合後のカテゴリーに再分類されるようになる。再分類の結果，修正が必要であれば(2)と(3)の作業を繰り返す。

(4) 分類結果を用いた分析　　ユーザーは，分類構造の中で注目する部分を選んで，クロス集計などの分析を行うことができる。各カテゴリーの文書数や，カテゴリー間の相関が，バブルチャートなどの形で可視化できる。

このように，対話型文書分類システムは，分類作業のすべてのプロセスを支援する。新しい文書が日々増加する場合にも，前述の(2)と(3)の処理で対応できる。すなわち，既知の内容の文書は既存のカテゴリーに分類し，新しい内容の文書に対しては新しくカテゴリーを作成することで，分類構造を長期にわたって改良しながら利用していくことができる。

3　クラスタリングと自動分類

対話型文書分類システムは，分類作業を効率化する機能として，新規のカテゴリーを自動生成する機能と既存のカテゴリーに文書を自動分類する機能を持つ（**図3**）。ユーザーはこれらの機能を，目的に応じて自由に使い分けることができる。

3.1　クラスタリングによるカテゴリーの自動生成

クラスタリングとは，データの集合をデータ間の類似性に基づきクラスタと呼ぶ部分集合に分割する，発見的な自動分類

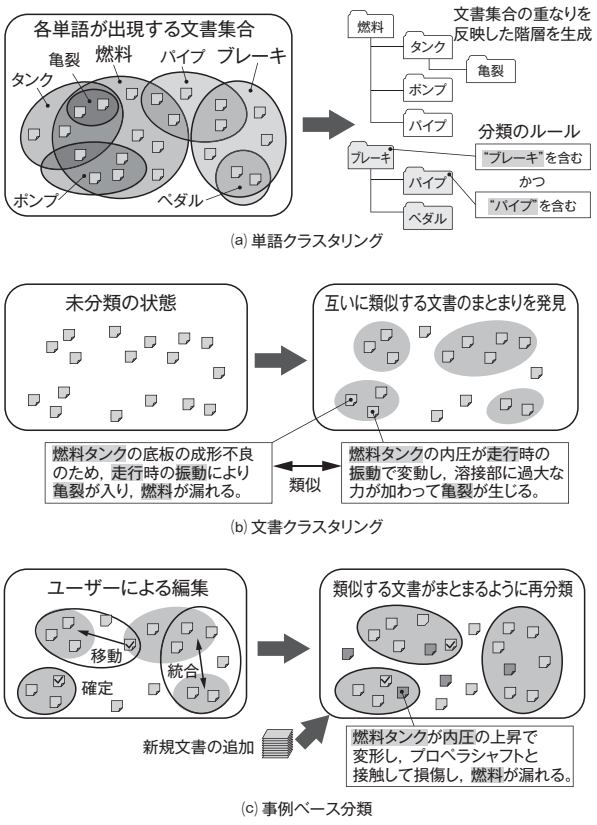

図3. クラスタリングと自動分類の処理 — 文書の内容に応じて，新たなカテゴリーを自動生成する機能と，文書をカテゴリーに自動分類する機能を持つ。

Document clustering and classifying processes

手法である。対話型文書分類システムは，文書の内容に応じたカテゴリーを自動生成するために，次の二つのクラスタリング機能を持つ。

(1) 単語クラスタリング　　文書集合を分類するのに適した単語を自動抽出する機能である。その処理を図3(a)に示す。複数の単語が同じ文書に出現する共起頻度が多い場合，例えば，“燃料”，“タンク”，“亀裂”という単語をすべて含む文書が数多く存在する場合には，これらの単語はある一つの話題，すなわち，“燃料タンクの亀裂”という不具合を表すまとまりであることが多い。このような単語のまとまりを自動的に抽出し，単語間の上位下位関係を推定することで，各単語に対応するカテゴリーとその階層構造を生成する。

(2) 文書クラスタリング　　内容が類似した文書のまとまりを自動生成する機能である。その処理を図3(b)に示す。各文書の内容を，文書中に出現する単語の頻度に基づいて，単語ベクトルと呼ぶ形式で表現し，文書間の内容の類似度を，ベクトルの余弦などで定義する。クラスタを生成するアルゴリズムにはleader-follower法[3]を用いる。この手法は比較的高速である点と，文書をまとめる際の類似度のしきい値を調節できる点で，対話的な分類作業に適している。

単語クラスタリングは，速度を重視した方法で実現されており，大量の文書の全体像を短時間で把握する目的に適している。一方，文書クラスタリングは，多数の単語の出現傾向に基づき，総合的に求めた類似度によって，文書のまとまりを生成する機能であるため，文書を詳細に分類する目的に適している。また，文書クラスタリングの結果に対して，後述する事例ベース分類を実行することで，より精度よく分類が行えるように分類構造を洗練できる。

3.2 カテゴリーへの文書の自動分類

対話型文書分類システムは，カテゴリーに文書を自動的に分類する機能として，次の二つの機能を持つ。

(1) ルールベース分類　　カテゴリーに設定したルール，例えば，“作成日が2010年以降”，“タイトルに東芝という単語を含む”といったルールによって，文書を分類する機能である。自動分類の処理は，ルールを満たす文書をデータベースから検索することで行う。このルールは前述の単語クラスタリングなどによって自動生成できる。

(2) 事例ベース分類（教師あり分類）　　ユーザーがカテゴリーに分類した文書（教師文書）にならって，未分類の文書を自動分類する機能である。その処理を図3(c)に示す。ユーザーが分類構造を編集すると，各カテゴリーの文書集合が増減するため，各文書の単語ベクトルの総和であるカテゴリーの単語ベクトルも変化する。この変化に従い，各文書をより類似度の大きいカテゴリーへ分類し直すことで，ユーザーの編集結果を反映した分類が行える。新しい文書が追加された場合も同様に，各文書を，もっとも類似したカテゴリーに自動的に分類する。

ユーザーは，これらの機能を自由に組み合わせて，階層の上位から下位に向かって文書を絞り込む形に分類構造を作成できる。分類構造の一部分をクラスタリング機能で自動生成し，別の部分を手作業で作成することもできる。このような枠組みにより，ユーザーは，目的に合った方法や観点を見つけて分類を試行し，その結果を検証して修正するサイクルを，システムとの対話を通じて効率よく進めていくことができる。

4 実作業による有効性の評価

対話型文書分類システムを用いて，特許調査という実際の作業を行うことで，分類作業の効率化と分類結果の品質向上に対するシステムの有効性を評価した。ここでは，その評価活動の内容と結果について述べる。

4.1 分類作業と分類対象文書

評価活動では，作業者すなわちシステムのユーザーを，当社の2名の研究者とした。この研究者たちが従事する研究

テーマに関連のある，自社と他社の特許文書約1,200件を対象とし，これを分類して各々の作業者がパテントマップを作り上げることを，この評価活動での分類作業とした。

作成するパテントマップは，当社の特許調査の方針に従い，特許が解決しようとする“課題”と，課題を解決するための“技術”の，二つの観点の相関によって出願傾向が把握でき，実際の研究業務で活用できることを，品質上の要件とした[2]。

4.2　作業手順と評価結果

評価の結果を**図4**に示す。分類作業は，手作業による特許調査とほぼ同じく，(1)調査不要な特許の除去，(2)“技術”観点での分類，(3)“課題”観点での分類，という手順で行われた。結果として，品質上の要件を満たす(4)パテントマップの作成までのすべての作業を，このシステムを用いて行うことができた。

各作業で有用な機能も明らかになった。例えば(2)の作業では，特許をおおまかに分類するには“発明の名称”を対象に文書クラスタリングを行うとよく，“技術”の内容で詳細に分類するには，“請求項”を対象とした文書クラスタリングと事例ベース分類が有効であった。一方，(3)の作業では，まず，発明の“課題”の表現として“速度”，“精度”などを洗い出しておき，これらの単語でルールベース分類を行った後，それまで想定していなかった表現を，単語クラスタリングによって発見するという方法がとられた。このように，多様な観点に適した分類方法が，ユーザーの試行錯誤と創意工夫によって得られる点も，このシステムの特長である。

分類作業の効率化については，このシステムを用いた所要時間が約4人日であるのに対し，従来の手作業による特許調査では，同程度の件数の特許で約10人日，また，当社で試行運用を行っている従来の特許自動分類システム[1], [2]を用いた場合は約7人日を要していた。調査の対象や目的が異なるため単純な比較はできないが，約2倍の効率化が見込まれる。

作業内容*		機能の使用方法	所要時間*
(1)調査不要な特許の除去	約1,200件の特許から調査が不要な約600件を除き，残りの約600件を分類対象とする	• 既存の分類コード(国際特許分類など)でルールベース分類 • 不要な特許のまとまりを文書クラスタリングで見つけて除去	約4.7 h
(2)“技術”観点での分類	“技術”の内容(課題を解決する手段)の観点で分類し，カテゴリー約40個(2階層)を作成	• “発明の名称”を対象にした文書クラスタリングによっておおまかに分類 • “請求項”を対象にした文書クラスタリングと事例ベース分類によって詳細に分類	約19.4 h
(3)“課題”観点での分類	“課題”の内容(発明が解決する課題)の観点で分類し，カテゴリー約10個(1階層)を作成	• 課題を表現する既知の単語を用いてルールベース分類 • 要約の“課題”を対象にした単語クラスタリングによって新たな課題の表現を発見	約5.5 h
(4)パテントマップの作成	“技術”と“課題”両方の観点のカテゴリーを，名称や並び順を調整してクロス集計	• クロス集計用バブルチャートを使用 • “技術”と“課題”の観点のほかにも様々な観点で分析できる	約1.0 h
合　計			約30.6 h

*2名の作業者がそれぞれ行った作業結果の平均的な値

図4．特許調査への適用結果 — 約1,200件の特許を調査する作業に適用して評価した結果，約4人日でパテントマップを作成できた。

Results of application to patent information analysis

5　あとがき

ここでは，企業内に日々蓄積される大量の文書情報を，ユーザーの目的に合った形に分類し，継続的に活用する枠組みを提供する，対話型文書分類技術について述べた。

今後は，日本語の文章の詳細な意味内容に基づいた，高精度な分類機能や，分類結果の要約・可視化機能を実現するために，パラフレーズ技術[4]や代表文生成技術（この号のp.68－69参照）に取り組んでいく。

一方，クレーム分析や特許分析をはじめとする実適用の面でも，実践を通じて技術を洗練していく。当社の知識継承ソフトウェアKnowledgeMeisterSucceed$_{TM}$[5]や，遠隔監視プラットフォームTMSTATION$_{TM}$[6]など，幅広い分野に適用し，企業の情報の“知識化”と“見える化”を促進する業務支援ソリューションを実現していく。

文　献

(1) 早川ルミ，ほか．日本語解析技術を活用した業務支援ソリューション開発への取組み．東芝レビュー．64，2，2009，p.30－34.
(2) 平　博司，ほか．特許調査に役だつ特許情報分類技術．東芝レビュー．62，2，2007，p.68－71.
(3) 岸田和明．文書クラスタリングの技法：文献レビュー．Library and Information Science．49，2003，p.33－75.
(4) 齋藤佳美，ほか．パラフレーズ技術を利用した情報・知識利活用ソリューション．東芝レビュー．64，8，2009，p.66－69.
(5) 小林賢治．知識継承ソフトウェアKnowledgeMeisterSucceed$_{TM}$．東芝レビュー．61，7，2006，p.41－44.
(6) 沖谷宜保，ほか．診断技術の進化を支える遠隔監視プラットフォームTMSTATION$_{TM}$．東芝レビュー．64，8，2009，p.41－44.
(7) 宮部泰成．ユーザーの意図を反映した対話型文書分類技術．東芝レビュー．64，2，2009，p.58－59.

後藤　和之　GOTO Kazuyuki
東芝ソリューション(株) IT技術研究所 研究開発部研究主務。情報検索，文書分類，ナレッジマネジメントなどの研究・開発に従事。情報処理学会会員。
Toshiba Solutions Corp.

平　　博司　TAIRA Hiroshi
東芝ソリューション(株) 技術統括部 商品・技術推進部主任。商品・技術戦略の企画業務に従事。情報処理学会，人工知能学会会員。
Toshiba Solutions Corp.

宮部　泰成　MIYABE Yasunari
東芝ソリューション(株) IT技術研究所 研究開発部。文書分類技術の研究・開発に従事。
Toshiba Solutions Corp.