# 文書分類方法及び文書再利用システム

6R−2　○永井 愛之† 森田 靖† 矢野 理† 高木 勝則† 工藤 裕‡ 平井 千秋‡
†日立エンジニアリング(株) 機電システム本部　‡(株)日立製作所 システム開発研究所

## 1. はじめに

オフィス業務における電子文書には、ワープロ、スレッドシートやスライド等がある。一般的には個人のマシンで作成・管理され、その殆どはごく一部の関係者やグループ内にのみ公開されるのが常である。しかし、その中には文書の雛型として、また参考資料として再利用することで、次回の文書作成時間を短縮するようなものもある[1]。

一方、どのようにしてこの再利用可能な文書を全社や全部門で共有できる文書として扱い、いかに手間をかけず整理された形で分類し利用者に提供する方法の問題もある。

そこで、文書の構成要素や記述内容と分類先をルール化したデータベースを用いてカテゴリ別に分類する方法と、利用者がその分類結果を辿ることで目的の文書を検索できる文書再利用システムの概要、特徴について述べる。

## 2. 文書再利用

### 2.1 文書再利用の現状

社内の設計部門で運用中の再利用支援システムには、次のような問題が発生していた。

- 検索のためのキーワードの登録に手間がかかる
- 分類が面倒なために、個人のマシンに未整理の状態で眠っていることが多い
- 開発部門や管理部門でも、文書再利用のニーズが発生した

これらに対処するため、文書登録作業を軽減する文書の自動分類機能を付加することにした。また誰にでも検索しやすいインタフェースにすることにより、文書の再利用促進、各部門における知識情報の共有、文書作成効率の向上を図る必要がある。

### 2.2 設計方針

これまでに社内に構築してきた再利用支援システムの運用過程も含め、以下の設計方針を定めた。基本的には、運用してきたプロジェクト管理資料及び成果物再利用システムの機能拡張という形で設計方針をまとめた。

(1) 文書の自動分類により、文書登録時の煩雑な作業の時間短縮
(2) 自動的に分類された文書へのリンク表示によって、目的の文書を閲覧、保存可能にする
(3) 自動分類方法は固定されたものではなく、導入部門の要求に合わせたカスタマイズを可能にする
(4) 部門間をまたがった文書の共有を可能とする

## 3. 文書分類方法

### 3.1 文書情報

文書を自動で分類する際の手がかりとして、表1の文書ファイルを対象に、表2の分類アイテムを利用することにした。

表1 対象文書

| |
|---|
| Microsoft Office (Word97/98/2000, Excel97/2000, PowerPoint97/2000) |
| HTML |
| XML |
| TEXT |

表2 分類アイテム

| 分類アイテム | 内容 |
|---|---|
| パス名 | パス名に含まれる文字(ファイル名も含む) |
| 見出し | 文書の見出し |
| 指定単語出現回数 | 指定した単語の出現回数 |
| 表紙 | 文書表紙の記述内容 |
| 範囲指定 | 本文指定範囲の記述内容 |
| Microsoft Office ドキュメントプロパティ | 組込みプロパティに設定可能なアイテム（作者、コメント、会社名等、11アイテム） |

### 3.2 条件定義

#### 3.2.1 条件

表2で示した分類アイテムを条件としてテキストマイニングを行う際、同じ条件を設定する場合でも、文書のタイプによってその要素の指すものが異なる場合がある。例えば、「見出し」条件の場合、Microsoft Word(以下、Word)文書であれば、文書スタイルに「見出し」スタイルが存在する。+VBA レベルでそれを検索して「見出し」と判定することは可能である。しかし、テキスト文書のような、文書スタイルを持たない文書の場合はWord文書と同じ検索基準での判定ができない。また、Word 文書であっても、アプリケーション固有の設定に従った文書で無い場合は、「見出し」というスタイルでVBAを用いて検索した場合は、それを見つけることができない。

つまり、文書固有のスタイルに従った文書、そうでない文書、あるいは文書タイプに関連付けされているアプリケーションが文書の要素を取得するための API を用意していない場合など様々なケースに対応可能とする為に、「~を"見出し"とする」といった条件の定義を行う為の機能が必要になる。

#### 3.2.2 アーキテクチャ

文書タイプ別の条件の定義は、文書タイプごとに用意する。また、今後サポート対象となる未知の文書タイプへの対応を容易とし、かつ既存の処理モジュールに対するエンハンス時の他モジュールへの影響を排除する為に、文書タイプごとに処理モジュールを独立させる。これにより、対象文書タイプ定義を拡張できる。

---

The Documents Classification Method and Development of a Web-based Artifact Reuse Database
Yoshiyuki Nagai, Yasushi Morita, Katsunori Takagi, Osamu Yano, Yutaka Kudo, Chiaki Hirai
Electrical Machinery & Digital Control Systems Group, Hitachi Engineering co., Ltd.
Systems Development Laboratory, Hitachi Ltd.

### 3.3 ルール定義

#### 3.3.1 ルール

一つ以上の条件と分類先が設定されているものをルールとし、以下のように定義する。

**分類ルール ::= 条件部**
**アクション**
**条件部 ::= [NOT]条件[LOP 条件部]**
**条件 ::= 分類先を決定するための条件**
**LOP ::= AND | OR**
**アクション ::= 分類先の決定**

例えば［ルール１］として「見出しに"○×システム概要"が含まれる場合、"/○×システム/説明資料"に分類する」と設定した場合、登録されている全ての文書のうち、「見出し」に、"○×システム概要"が含まれている文書は、 " /○×システム/説明資料"カテゴリに分類される。

また、複数のルールが成立して、それぞれの条件で定義している分類が異なる場合に対応するため、ルールには優先順位を設ける。

#### 3.3.2 ルール生成

ルールの生成には、専用のルールカスタマイズツールを用いて行う。ここでは、表２に示した分類アイテムの詳細設定を行い、具体的な分類条件を作成、AND/OR/NOT の論理演算子の設定と分類先となるカテゴリツリーの作成と指定を行い、ルールを生成していく。

### 3.4 自動分類処理の流れ

あるファイルサーバに登録された文書を対象にした、自動分類処理の流れを示す。

共有フォルダに登録されている全ての文書に対し、１文書単位に全てのルールを実行する。実行するルールが複数存在し、結果が競合する場合は優先順位に従い分類を行う。また、ルールの記述内容によって、同一文書に到達するルートは複数存在させることも可能とする。つまり、"○○製作所の××仕様書"は、「○○製作所」というカテゴリからと「仕様書」というカテゴリから引くことができる。

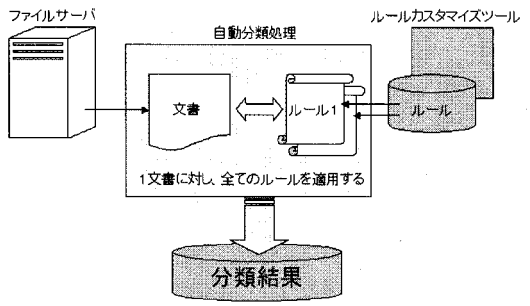


図１ 自動分類処理

## 4. 文書再利用システム「AWARD」

### 4.1 システム概要

これまでに述べた設計方針、分類方法を実現するために、以下を特徴とする Web ベースのシステムを構築した。

(1) 文書登録には、共有フォルダを用いる
(2) 分類は定期的に行う
(3) 分類結果表示は、カテゴリ別階層構造とする

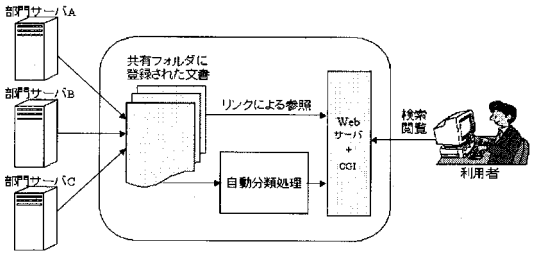


図２. システム概要

従来、目的の文書を取り出すときは、その検索対象となる文書がサーバマシン内に分散して登録しているため、検索に手間がかかっていた。このシステムでは事前に検索範囲を極小化しているため、目的の文書を短時間で検索できる。分類処理は、ルールの追加・変更があった時はそのルールのみを、また文書の修正が発生した場合はその文書のみを対象に処理を行うようにし、マシンの負荷を軽減した。

また Web 上では、従来のフォルダ構成を生かしたフォルダツリー表示、文書のプレビュー表示、全文検索機能、文書のダウンロード、アップロードが可能である。

### 4.2 分類画面

自動分類結果の画面は、図３のようなカテゴリ別の階層構造にした。このリンクを辿ることで登録された文書を段階的に絞り込んでいくことが可能となり、検索の際キーワード入力の必要が無くなる。また、キーワードで検索する時に比べ、これまで検索漏れで見落としてきた文書も一覧の中に表示されるので、何度も検索をしなおす必要が無い。

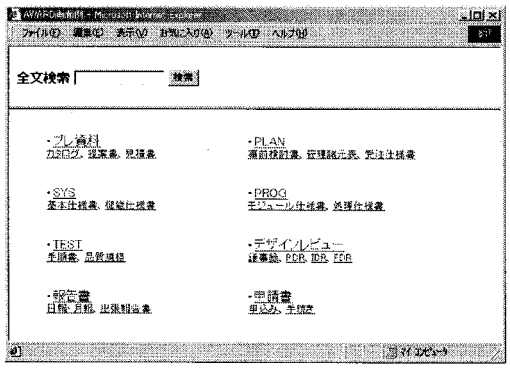


図３. AWARD 画面例

## 5. 終わりに

本稿では、文書分類方法とそれを生かした文書再利用システムについて述べた。この自動分類機能により、文書の登録作業の軽減と、今まで再利用されることのなかった文書が利用者の目的に応じ検索可能となった。また、検索したい文書を複数の観点から探り当てることが可能になり、検索目的とする文書に関連する知識情報の取得も期待できる。

今後は、さらにユーザビリティの向上を目指す。

## 参考文献

[1]:工藤 他、“再利用支援データベース「AWARD」の開発”、2K-02、第 61 回情処全大、Sep. 2000