5W-4: 分類階層の自動生成機能を備えた文書分類システムの構築

筒井秀樹、福井美佳、真鍋俊彦
(株) 東芝 研究開発センター

1 はじめに

近年、電子化された文書が大量に流通するようになったことを背景に、文書データベースを簡単に整理して閲覧したり、どのような情報が登録されているかを提示する手段として文書分類システムが求められている。文書データベースの全体像を分類階層で提示することにより、ユーザはそれを辿って必要な知識情報のまとまりを参照できる。また、アンケートなどの文書情報の内容分析にも利用できる。しかし、これまでの文書分類システムでは、初期分類階層や分類のためのプロファイルをサイト管理者が作成しなければならず、システム立ち上げ時や更新される文書データベースのメンテナンスが大きな負担となっていた。

本システムでは分類階層 (カテゴリ構造) と各カテゴリのプロファイルを自動生成する機能を連携させることで、管理者の負担を大幅に軽減させる。本報告ではシステム構成を中心に、本システムがどのようにカテゴリ構造構築から新着文書の分類までを支援するかについて報告する。

2 アプローチ

文書分類システムにおいて管理者の負担となる作業を以下にあげる。

(1) 立ち上げ時、初期分類階層を作成する
(2) 運用開始後、新着文書の分類先を決める
(3) 新着文書にあわせてカテゴリ階層を修正する

これらの負担を軽減するために分類階層の自動生成技術と文書の自動分類技術を組み合わせたシステムを開発した。

システム立ち上げ時には文書データベースを対象として、初期分類階層を自動生成する。運用時には新着文書を自動分類する。また、多くの文書が溜ったカテゴリではサブカテゴリを自動生成する。未分類文書を対象として分類階層を自動生成する事で話題の変化に対応する。

以降では本システムのうち、分類階層の自動生成と文書の自動分類を連携させる事で、上記 (1)(2) の作業を支援する方法について述べる。

Development of Text Categorization System using Hierarchical Index Generator,
Hideki TSUTSUI, Mika FUKUI, Toshihiko MANABE, Research and Development Center, Toshiba Corp.
1Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 210, Japan.

3 システム構成

処理の流れを図 1 に示す。

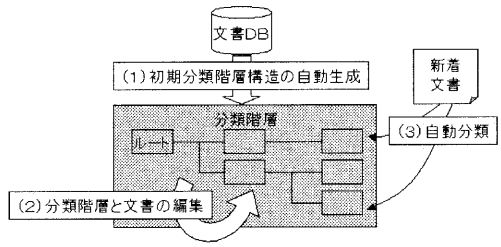

図 1: 処理の流れ

(1) 分類階層自動生成機能

データマイニングでの相関ルールの抽出方式 [3] を語の出現相関の検出に適用し、語を階層配置する事で分類階層を構築する。例えば、語 A、Bに対して、Bが出現する時あらかじめ定められた割合以上で同じ文書に A が出現する場合、両者の間に相関関係 $B \rightarrow A$ が成立しているとする。このとき、語 A に基づくカテゴリを親カテゴリ、語 Bに基づくカテゴリを子カテゴリとするカテゴリ階層を構築する。文書の表題など指定された部分についてだけ処理を行うことで、組合せの爆発を防ぐ。

(2) 分類階層と文書の編集機能

ユーザインタフェースはブラウザ上に構築した。システムはサーバークライアントモデルをとり、分類を編集・管理する管理者用インタフェースと、分類を参照する一般ユーザ用インタフェースを備える。管理者用インタフェースではカテゴリの作成・移動・削除と、文書の追加・コピー・移動・削除をおこなうことができる。管理者用インタフェースで確定した文書の分類結果を一般ユーザ用インタフェースで参照、利用する。図 2に管理者用インタフェースの画面例を示す。

(3) 新着文書の自動分類機能

新着文書を自動分類するために、各カテゴリ内文書から、カテゴリを特徴付けるプロファイルを求

める。ここでのプロファイルとは特徴語集合のことで、語の統計量を計算する手法 [2] を応用して求める。新着文書とプロファイルとの類似度はベクトル空間モデルに基づき計算し、分類先のカテゴリを決定する。

カテゴリ内文書を基にしてプロファイルを作成する。しかし初期分類階層を自動生成した場合、各カテゴリ内には文書が入っていない。そこで本システムでは、分類階層自動生成機能により生成された語を用い、文書データベースから文書を検索することで初期のカテゴリ内文書を収集する。例えば図 3 の「特許→出願」のカテゴリでは、「特許 出願」をキーワードとして、文書データベースから文書を検索し、検索スコアの高いものを初期のカテゴリ内文書とする。

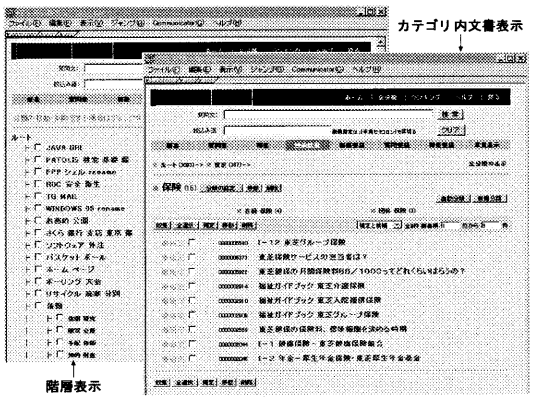

図 2: 管理者用インタフェース

4 実験

4.1 分類階層の自動構築

社内で収集した事務手続などのノウハウ情報 [1] 約 10,000 件から分類階層の構築実験を行った。処理時間は約 2 分 (UltraSparc 300MHz) で、全カテゴリ数 74、平均 1.27 段の分類階層が構築された。約 32 %の文書が初期カテゴリ内文書となった。大きな文書データベースに対して、実用に耐える速度で分類階層を構築できることを確認した。本方式で生成した分類階層の一部を図 3に示す。

4.2 新着文書の分類精度

初期分類階層を自動生成した際、検索により収集したカテゴリ内文書を基にプロファイルを作成しているため、人手により適合する文書を集めてプロファイルを作成した場合に比べて新着文書分類の精度が低下することが予想される。そこで、BMIR-J2[4] のグループ A(キーワードの存在確認で検索できる質問文) から 10 件選択し、正解文書 20 件からプロファイルを作成した場合と、検索結果の上位 20 件からプロファイルを作成した場合で分類精度の比較実験を行った。その結果、平均適合率は 0.83 から 0.72 への低下に留まり、自動的に収集した文書でプロファイルを求めた場合でも大きな性能低下にならないことを確認した。

図 3: 分類階層の例

5 おわりに

分類階層自動生成機能、分類階層と文書の編集機能、新着文書の自動分類機能を使い、初期分類階層の作成から運用までをサポートする文書分類システムを開発した。分類階層の自動構築は実用に耐える速度であった。新着文書の自動分類では 10%程度の性能低下が見られるが、管理者による編集機能を備えており、分類において最もコストのかかると考えられる初期分類階層の作成と、運用でコストのかかる新着文書の分類について、管理者の負担を軽減できる見込みを得た。

本システムで分類階層の自動生成に用いた手法は、単語の出現相関を基にしているため、文書の中に単語が直接現れなければ、カテゴリを作る事ができない。例えば新聞記事からは「野球」「テニス」「サッカー」などの具体的なカテゴリは生成できても、「スポーツ」などの概念的なカテゴリの生成は難しいと考えられる。概念的なカテゴリをどのように扱うかが今後の課題である。

参考文献

[1] 中山他:知識情報共有システム (KIDS) の開発と実践－組織におけるノウハウ共有の促進－, 人工知能学会 AI シンポジウム'99, SIG-J-9901, pp.137-142, 1999

[2] Robertson, S.E. et al.: Simple, Proven Approaches to Text Retrieval, Computer Laboratory, University of Cambridge, 1994

[3] R. Agrawal, et al.: Fast Discovery of Association Rules, Advances in Knowledge Discovery and Data Mining(AAAI Press), pp.307-328, 1996

[4] 木谷他:日本語情報検索システム評価用テストコレクション BMIR-J2, 情報処理学会データベースシステム研究会, DBS-114-3, pp.15-22, 1998