5W—6　多様な要求に対応するテキストの自動分類システム

片山佳則
(株)富士通研究所 ドキュメント処理研究部
e-mail: katayama.yoshin@jp.fujitsu.com

1．はじめに

WEB ページの急増に代表されるように、膨大な量の情報があふれ続けている現在、情報の収集技術や検索技術などの情報処理のための各種の要素技術の整備が進んでいる。その中でも、特に情報提供や共有を目的とした情報の整理や絞込みのために、テキストの自動分類技術は欠かせない重要な要素技術である。以下では、テキストの自動分類のことを単に自動分類と呼ぶ。これまで人手作業中心で行っていた様々な仕分け作業に対して、自動分類技術を活用することで、処理時間の短縮、分類の一貫性保持、大量の情報処理などが可能になる。

本稿では、WebClassify[1)]のサービスで活用されている自動分類システムの概要とその特徴の一つである組み込まれている様々な分類方法を述べ、その情報サービスへの活用可能性をまとめる。

2．自動分類システムの概要

既に作成されている自動分類システムには、複数の分類方法が組み込まれており、適用場面に応じてその分類方法が選択できる点と、分類ルールの蓄積に大きな特徴がある。

2．1 自動分類システムの基本構成

自動分類システムは、図1に示すように、分類ルール作成のためのサンプルテキストの入力処理(自動学習)と作成された分類ルールを用いて、入力された対象テキストを分類（自動分類)の2つの処理からなる。

本稿の自動分類システムは、一般的な自動分類がもつ特徴（処理時間短縮、分類の一貫性保持、大量の情報処理など）に加えて、以下のような特徴を備えている。

- ○ 複数の分類方法を比較して最適なものを選択可能。
- ○ 簡単なパラメータ操作により分類結果の適合率や再現率を操作可能。

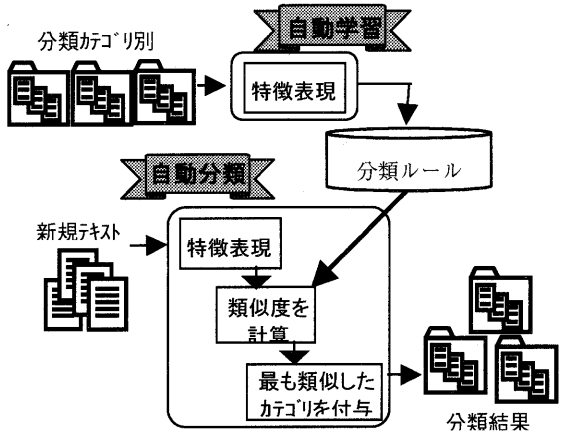

図1　自動分類システムの基本構成図

- ○ 単一カテゴリへの分類や複数カテゴリへの分類を選択可能。
- ○ 分類ルールを直接変更して分類精度を調整可能。

図2は自動分類システムの動作画面である。図中左側の小ウインドウでは、分類カテゴリを一覧できる。

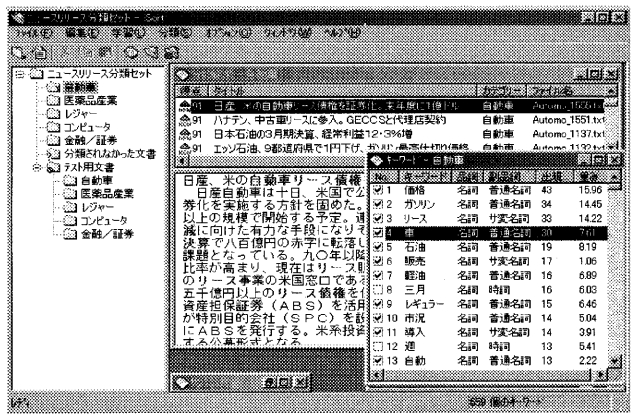

図2　自動分類システムの起動トップ画面

右側では、複数のウインドウを開いて以下の情報の確認、操作が行える。(1)分類結果（新たに分類されたテキストのアイコン表示、分類カテゴリとの関連度を表す得点やテキスト本文表示)、(2)キーワード(カテゴリ毎の特徴語)、(3)サンプルテキスト

Text Categorization System using Several Methods
Yoshinori KATAYAMA
FUJITSU Laboratories LTD.

(図２では最小化表示されている)。

対象テキストの分類作業やサンプルテキストの追加作業はエクスプローラ等からのドラッグ＆ドロップで簡単に行える。

２．２ 選択可能な分類方法

現在自動分類の分類方法[2]としては、ベクトル空間法、確率に基づく法、ルールベース法などの各種の方法が提案されている。どの方法も、それぞれに精度向上に向けた改良や拡張が進められており、一意に絞り込める状況にない。

本稿の自動分類システムは、複数の分類方法を組み込める形態を取っている。現在は次に示す分類方法が組み込まれており、必要に応じて選択できる。

ベクトル空間法：分類カテゴリの特徴語と分類対象テキストの特徴語をベクトル表現し、両ベクトルの類似度をベクトルの余弦や距離計算などにより計算、類似度の高い分類カテゴリを分類対象テキストに付与(分類)する。

ルールベース法：テキストをカテゴリに分類する条件(カテゴリとの関連度は数値化されている)を記述した分類規則を用意し、分類対象テキストに適用される分類規則の数値をスコアとして、スコアの高いカテゴリを分類対象テキストに付与(分類)する。

ブースティング法：分類方法をサンプルテキストの分布を操作しながら繰り返して実行し、それぞれで作成される分類方法の多数決で分類対象テキストに分類カテゴリを付与(分類)する。

２．３ 分類方法の選択に関する考察

様々な分類対象テキストの自動分類作業から、各分類方法の傾向として次のようなことが明らかになっている。

ベクトル空間法は、サンプルテキストから分類ルールを学習するのに時間がかかる(2000 ﾃｷｽﾄからの学習３分弱)が、対象テキストの分類は速い(2000 ﾃｷｽﾄの分類２分)。

ルールベース法は、サンプルテキストから分類規則を学習するのは速い(2000 ﾃｷｽﾄからの学習１５秒)が、対象テキストの分類に時間がかかる(2000ﾃｷｽﾄの分類８分弱)。尚、ルールベース法に関しては、ビジネス関連の分類規則の蓄積･流用の実績がある。

ブースティング法は、サンプルテキストに矛盾が無ければ、繰り返し回数に応じて、サンプルテキストの分類精度が非常に高い(98%以上)。

組み込まれている分類方法の比較として、対象テキストの形式や内容に応じて、分類方法の得手・不得手があるという仮定のもとで比較してみた。

分類対象テキストとして、新聞記事、特許要約、WEBページなどを試みた。新聞記事では、サンプルテキストのある４～１２の分類カテゴリに対する新規新聞記事約1500テキストの分類。特許要約では、公開広報からIPCコードによる分類として５～２５の分類カテゴリを取り上げ、関連する特許要約約7000テキストの分類。WEBページでは、サンプルテキストのある１８の分類カテゴリに対して、ロボットで集めたWEBページ約3500ページの分類。

これらのテキストでの分類比較を試みた結果、分類対象テキストごとの、精度としてそれぞれの分類方法による有意な差は得られなかった。このことから現時点では、分類対象テキストに応じた分類方法の選択は、与えられたサンプルテキストの分類精度で判断すること、および、各分類方法の特徴と自動分類システムの運用方法の対比から行うことになる。後者の具体的な選択観点は、分類ルールの学習時間、対象テキスト分類時間である。

３．自動分類の活用場面

対象テキスト側からの扱いの区別では、すべてにカテゴリ付与されるべきものと精度を保ってカテゴリ付与されればよいものに分かれる。前者は、整理・格納などの推薦ツールとしての活用となり、ほとんどの場合、後に人の判断や作業がある。後者は、主に検索の絞込みなどの情報収集系との組合せとなる。それぞれ分類結果後の作業内容に応じてシングルラベルとマルチラベルの使い分けが行われる。分類方法の選択は、前記の各分類方法の特徴およびシステムの運用方法に基づいて行う必要がある。

４．おわりに

今後は、自動分類システムに組み込まれている分類方法の改良や精度改善を進めるとともに、このシステムに基づく各種のサービスの提案と実践を進める。

謝辞 本稿で述べた自動分類システムは、富士通(株)ﾈｯﾄﾜｰｸｻｰﾋﾞｽ本部DBｻｰﾋﾞｽ部の内野氏、及び同部の皆さんとの協力により開発したものである。皆さんの協力に感謝いたします。また、各種の分類方法については、富士通研究所ﾄﾞｷｭﾒﾝﾄ処理研究部の星合氏、塚本氏、及び同部の関係各位に感謝いたします。

参考文献

1)WebClassify:FENICS EIP ｻｰﾋﾞｽ,富士通 NS 本部 http://www.fujitsu.co.jp/hypertext/solution/fenics/fenics.html

2)徳永:言語と計算５ 情報検索と言語処理、東京大学出版会