テキスト分類支援ツールＦＬＵＴＥの開発（１）－機能と構成－

4Ｊ－5

間瀬久雄　森本由起子　辻　洋　絹川博之
（株）日立製作所　システム開発研究所

１．はじめに

情報の電子化、ネットワーク基盤の整備により、大量の電子化文書が身近に氾濫するようになった。その結果、これらを目的別／内容別に分類整理することが時間的・体力的に困難になっており、文書自動分類システム実現への期待が高まっている。

しかし、文書の分野や構造，分類体系によって分類精度は大きく異なるため、実際には実験システムを構築して自動分類の実現可能性を事前に評価する必要がある。特に、自然言語処理技術を用いて文書内容を解析するアプローチを採る場合、用語辞書を構築し、単語分割などの処理プログラムを個別に開発しなければならず、実現可能性の検証までには多大な時間と労力を要することになる。

我々は、テキスト文書の自動分類の実現可能性の早期検証と、自動分類システムの構築支援を目的としたテキスト分類支援ツールＦＬＵＴＥ（Filtering Lens software system for Unclassified TExts）を開発した。本ツールは、特許や新聞記事，電子メール・ニュースなどの分類に有効である[1]。

２．ＦＬＵＴＥの概要[2]

２．１　特長

本ツールは、次の特長を備えている。

（１）表層情報に基づく分類方式

文書を特徴付けるキーワードの出現傾向に基づいて分類先を決定するので、処理が高速である。

（２）知識ベースの自動初期化

過去の文書事例から、分類する際に必要な知識ベースを自動作成するため、早期評価ができる。

（３）知識ベースの構造が単純

重み付きキーワードをカテゴリ別に格納しているので、人手による補正が容易である。

（４）分類結果に対する確信度を出力

計算機が出力した分類結果をチェックするという、マンマシン分担型のシステム設計が可能である。

（５）用語／概念シソーラスが利用可能

上位・下位概念や類義語を包括した分類が可能である。

２．２　前提条件

本ツールは次の前提条件を満足する必要がある。

（１）分類体系が予め定義されていること。

（２）分類カテゴリが互いに包含関係にないこと。

（３）知識ベース作成用の分類済みの文書事例が十分な数だけ存在すること。

２．３　機能

本ツールの処理の流れを図１に示す。本ツールは、主に次の機能を提供するものである。

（１）対象文書の統計的分析

キーワード抽出エンジンにより、文書別／カテゴリ別のキーワード一覧などのテキスト内容統計データ，多くの文書／カテゴリに共通して出現する、キーワードとなりえない「共通語」の一覧，単語辞書に未登録の単語（複合語）の一覧など、文書の統計的データを出力する。

（２）知識ベースの自動初期化

知識ベース作成用の文書から抽出した重み付きキーワード群から知識ベースを自動生成する（知識ベース変換）。キーワードの重み付け方法は、文書の種類によって異なるが、基本的には、出現頻度や出現位置に基づいて行う。なお、文書の解析範囲を限定するためのテキストスキミングオプション機能[3]，および各キーワードに対する上位語／下位語／類義語を追加／置換するシソーラス展開オプション機能も備えている。

Text Categorization System Building Support Tool.

Hisao Mase, Yukiko Morimoto, Hiroshi Tsuji, Hiroshi Kinukawa.
Systems Development Laboratory, Hitachi Ltd.
{mase,yukiko,tsuji,kinu}@sdl.hitachi.co.jp

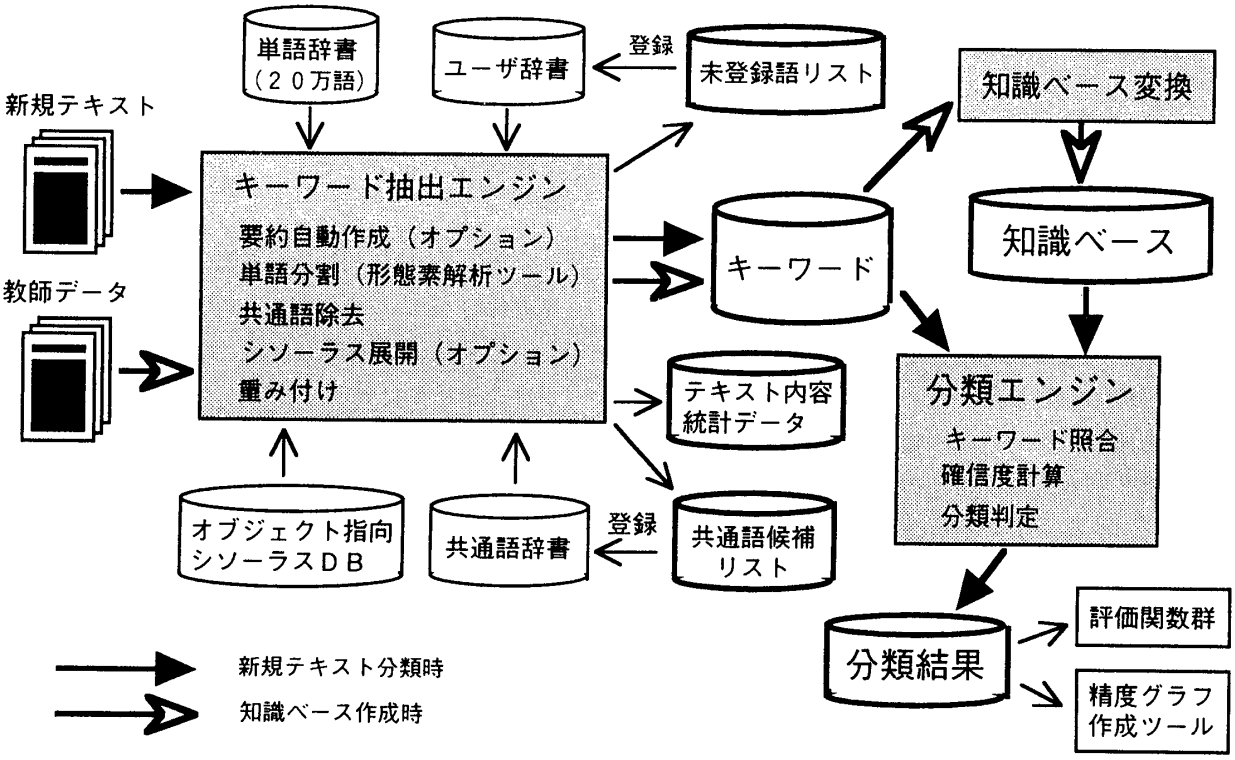

図１　ＦＬＵＴＥの処理の流れ

（３）分類エンジンによる分類実験

新規文書から抽出した重み付きキーワードと、知識ベースに格納されたキーワードとの照合により類似度を計算する。カテゴリ毎に算出した類似度の得点分布から確信度を算出し、あるしきい値以上の確信度をもつカテゴリに分類する。

（４）分類精度評価

分類結果について、全体の再現率／適合率および、カテゴリ別の再現率／適合率を算出する。また、評価結果をグラフ化することもできる。

（５）ＧＵＩによる分類結果の確認，分析

個々の文書の分類結果および解析ログデータを参照しながら分類結果を分析したり、知識ベースをチューニングしたりすることができる。

２．４　性能

ワークステーション３０５０ＲＸ（メモリ１９２Ｍバイト）での１件当たりの処理速度は、平均的な新聞記事（５００文字）で、約４秒である。大部分の処理は単語分割処理で費やされるので、処理速度は、文書の長さにほぼ比例する。

３．終わりに

自動分類の実現可能性の早期検証を実現し、自動分類システムの構築を支援するツールＦＬＵＴＥについて述べた。

参考文献

１）森本他３名：テキスト自動分類支援ツールＦＬＵＴＥの開発（２）－障害事例分類への適用－，情報処理学会第52回全国大会講演論文集，1996.

２）辻他３名：テキスト自動分類エキスパートシステムの一構成法，情報処理学会第49回全国大会講演論文集（３）3-93，1994.

３）間瀬他２名：パラメータ設定による文章要約支援システム，情報処理学会第48回全国大会講演論文集（３）3-103，1994.