テキスト自動分類エキスパートシステムの一構成法

6J－8

辻　洋，間瀬久雄，木山忠博，絹川博之
（株）日立製作所

１．まえがき

多くの人が自ら計算機を用いて文書を作成し、それを送受信するようになってきている。その量が多くなると共に文書を中心とした外部から着信するフロー情報の自動分類技術が必要となってきている。このような背景で、一つのアプローチは、文書に予め付加された構造を持ったフィールド情報を設け分類するものである[1]。別のアプローチは、文書内容を解析して、そこに現れる語句を用いて分類するものである[2]。

前者のアプローチは、分類のための情報をあらかじめフローの送信側で設定しておく必要がある。分類のルールを書くことは比較的容易でルールを書けば意図通りの分類がなされると考えられるが、適用範囲にはおのずと制限が課せられる。一方、後者では、分類のための情報を送信する側で意識する必要が無く、適用範囲が広くなると考えられるが、着信する側で分類のためのルールを書くことは必ずしも容易ではなく、また表層的な処理では意図通りの分類は困難であった。

２．問題の記述

本稿の目的は、後者のアプローチとして、分類知識の半自動生成（自動初期化）を特徴とするテキスト自動分類エキスパートシステムの構成法を明らかにすることであり、次のことを前提とする：

（１）分類済みの教師テキストデータが各分類当たり複数ある、

（２）同じ分類に入るテキストには、分類の特徴を表す単語が出現する可能性が高い、

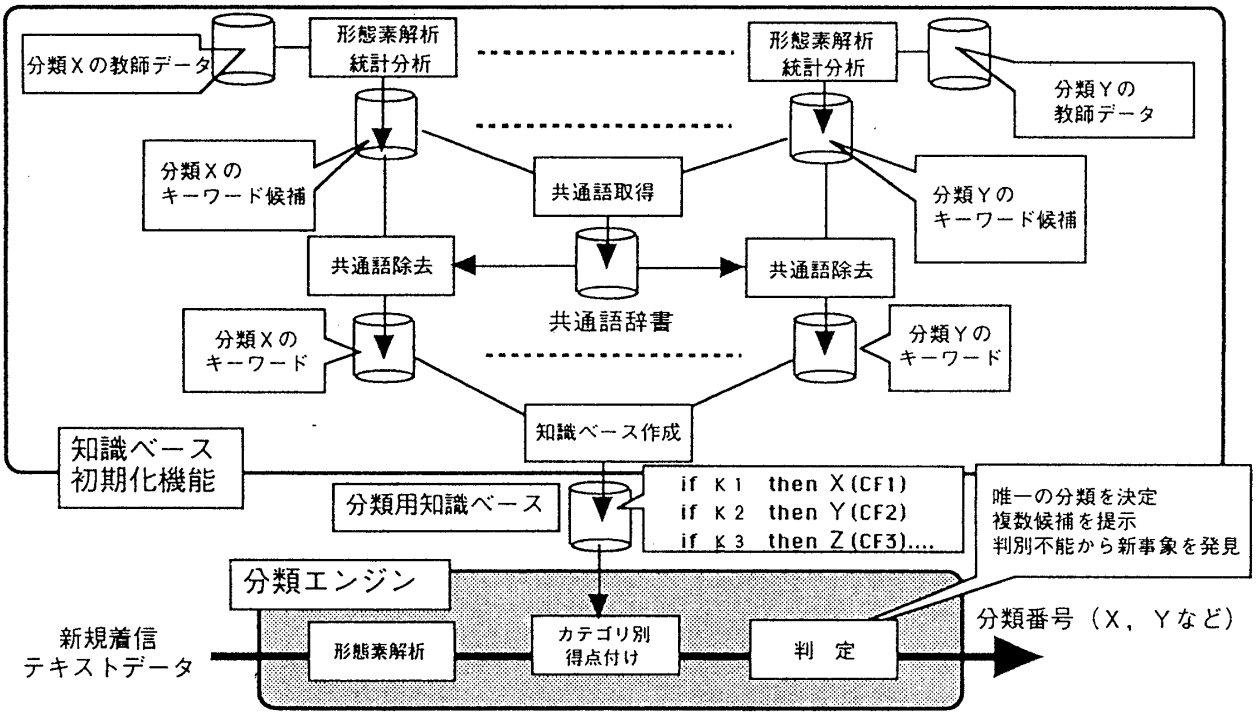
（３）ある分類に入る分類の特徴を表す単語は、他の分類ではあまり現れない。

つまり、「複数ある」「可能性が高い」「あまり現れない」といったファジーな状況で、問題は

（１）教師テキストデータから分類の知識を取得する機構を設計すること、

（２）取得した知識をもとにした分類エンジンを設計すること、

である。



第１図　テキスト自動分類エキスパートシステムの構成

An Architecture for Text Classification Expert System
Hiroshi Tsuji, Hisao Mase, Tadahiro Kiyama, Hiroshi Kinukawa
Systems Development Laboratory, Hitachi, Ltd. 3-6-1 Bakuroh-Machi, Chuo, Osaka 541 Japan

3．システムの構成

提案するシステム構成を図１に示す。

知識ベース初期化機能の概略アルゴリズムは以下のとおりである。

（１）分類済みのテキストデータを形態素解析し、キーワード候補（名詞、サ変動詞、複合語など）を出現頻度と共に取得する、

（２）どの分類にも一様に現れる語を共通語とする、

（３）各分類毎にキーワード候補から共通語を取り除いて各分類のキーワードとする（つまり、出現頻度が特定分類に偏った語がキーワードとして残り、そのキーワードに対し決まる分類がルールとみなされる）、

（４）分類毎のキーワードを知識ベースとしてまとめる。ここで、出現頻度をもとに分類の確信度を［0.0, 1.0］の間で設定する。

初期化した知識ベースは、

ｉｆ　キーワード　ｔｈｅｎ　分類（確信度）

という形式をとっているので、人手で洗練化することが容易である。

一方、分類エンジンの概略アルゴリズムは以下のとおりである。

（１）分類対象のテキストから、キーワード抽出範囲を特定する、

（２）その範囲のテキストデータを形態素解析し、含まれている用語を抽出する。

（３）知識ベースを参照して、分類のための確信度計算を行う、

（４）判定を下す。

判定に当たっては

①確信度の高い分類が一つのみの場合

⟶　唯一の分類を決定、

②確信度の高い分類が複数ある場合

⟶　複数の候補を提示、

③確信度の高い分類がない場合

⟶　判別不能

の３種類を設ける。

システムの構築に当たっては、知識ベースを初期化した後、教師データを上記アルゴリズムで分類する。その結果、全テキストデータに現れた語句の統計情報を参照しながら、知識ベースを洗練化していく。そして、教師データにたいして行う分類結果を検討して、対象とする分野が先に上げた前提を満たすか否かを判断する必要がある。

4．本手法の特徴

提案手法は、

（１）本手法は形態素解析に基づく表層処理であり、１件当たりの処理速度が速い、

（２）分類のための知識ベースを事例から初期化するため早期フィージビリティテストが可能、

（３）知識ベースの構造が単純なため、人手による保守が容易、

（４）分類結果に対して、一律全自動分類にこだわならないクレジット付けを行っている（１００％正解を目指すことは現実的でない）ので、マンマシン分担型の設計が可能、

である。

一方、

$$\text{再現率}=\frac{\text{分類Ａでありかつ分類Ａと判定されたデータ}}{\text{分類Ａのデータ総数}}$$

$$\text{適合率}=\frac{\text{分類Ａでありかつ分類Ａと判定されたデータ}}{\text{分類Ａと判定されたデータ総数}}$$

と定義すると、

（５）知識ベースをチューニングしても再現率と適合率にはトレードオフの関係が残るが、

さらに、

（６）表記の揺れ（送りがな、大文字小文字、ほか）、同義語の吸収

（７）固有名詞（人命、社名、組織名）などの辞書の充実

（８）語の共起関係のうまい利用による補正ルールの人手追加

により、分類精度を上げることが可能である。

5．あとがき

テキスト自動分類エキスパートシステムの一構成法について述べた。システムの外部から着信する膨大な量のテキスト情報を解析・分類し、それらを必要とする人に送る技術の必要性は今後益々高まっていくものと考えている。

参考文献

［１］K. Lai, et al.: Object Lens - A "Spread Sheet" for Cooperative Work, Proc. of CSCW, 1988.

［２］P. Hayes, et al : CONSTRUE/TIS - A System for Content-Based Indexing of a Database of News Stories, Proc. of Second Conf. IAAI, 1990.