2005 年度卒業研究概要

複数の解析手法に基づく文書分類システムの提案

指導教員
承認印

大谷　紀子研究室
0232222　村田　康宏

１．研究の目的

コンピュータで文書を作成し保存するとき、適切なファイル名をつけなかったり、全てのファイルを 1 つのディレクトリに保存するということを繰り返したりすると、一見では内容のわからないファイルが蓄積されていく。蓄積された文書ファイルの量が増えるのに従って再利用することが困難になるが、文書ファイルを適切に整理・分類することができれば再利用しやすくなる。

本研究では、文書整理の一助となることを目的として、コンピュータ内に未分類のまま蓄積された文書ファイルを簡単に自動分類するためのシステムを提案する。

２．文書分類の手法

２．１　文書分類のアルゴリズム

単語抽出には TF-IDF 法[2]と「共起の統計情報に基づく文書からのキーワード抽出アルゴリズム」[3]を組み合わせて用いる。前者は単語抽出の計算に全文書数を用いるため、文書全体からみて特徴的な単語を抽出できる。一方後者は登場回数の少ない単語もキーワードとして抽出できるため、両アルゴリズムを組み合わせることでより精度の高い分類が可能となる。

文書の分類には単語を利用するため、文書を単語に分割する必要がある。単語へ分割するために形態素解析システム ChaSen[1]を用いた。ChaSen とは文書を単語に分割し、品詞情報を得ることができるソフトウェアであり、解析によって名詞と判定された単語を単語抽出の対象とした。

分類処理の手順を図 1 に示す。

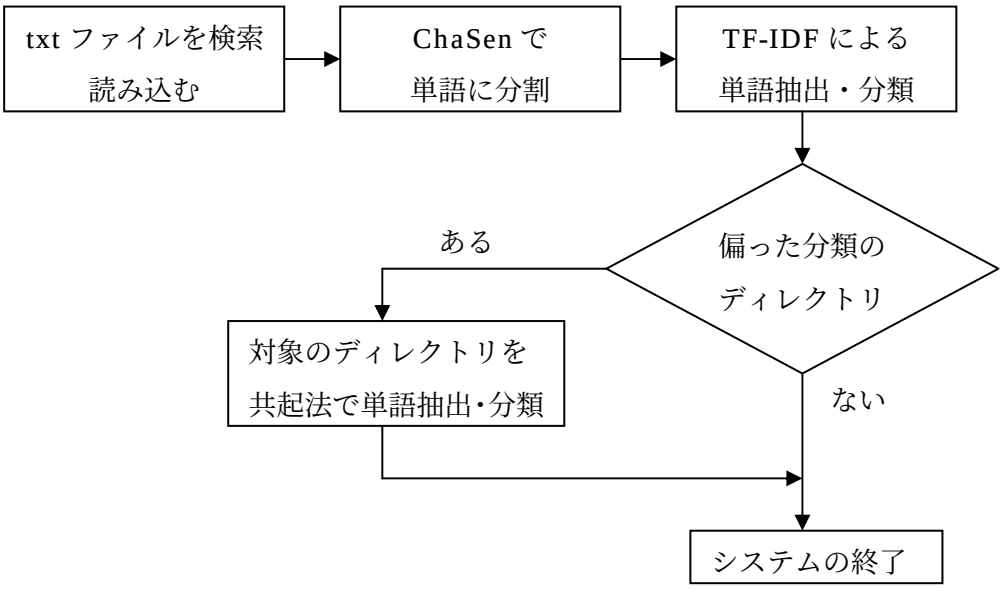

図 1：処理手順

文書の分類にはベクトル空間モデル[4]を用いる。ベクトル空間モデルとは、単語に基づいて文書を 1 つのベクトルによって表現し、ベクトルの向きから内容を判断する手法である。ベクトル空間モデルを利用することで、比較する文書同士の類似度を余弦の値から判定することができる。

２．３　システム構成

ユーザには分類するファイルが存在するディレクトリを指定させる。本システムでは指定ディレクトリ下にあるサブディレクトリも検索対象とし、全ての.txt 形式の文書ファイルを分類対象にする。分類が完了したファイルは、システムが新しく準備したディレクトリに分類された上で保存される。

３．システムの実行結果

３．１　実験の準備

メールクライアント Becky!より 1 アカウント分のメールデータをテキストファイルとして出力し、そのテキストファイルに対してシステムによる自動分類を試行する。メールクライアントのデータをテキストファイルとして出力するためにフリーソフトウェア「CircleBecky」を利用した。

３．２　結果

2004 年 6 月から 10 月の期間に受信した 60 件のメールデータに対して自動分類を試行した。実験結果の例として、大学から送信された就職関連のメールと、SA 代理募集に関するメールのシステムによる分類結果を Subject によって振り分けを行ったものと比較したデータを示す。Subject 振り分け結果の適合率を 100%とし、システムの閾値を 0.4 に設定した場合の分類結果は、就職関連のメールの適合率は 84.2%、再現率は 61.5%となった。一方、SA 代理募集に関するメールの適合率は 52.3%、再現率は 25.6%という値になった。

４．まとめ

実験によって、本システムによる分類はある程度の適合率を確保できるが、適合率に比べて再現率が低くなるという結果が導かれた。分類結果の特徴として、上半期の就職ガイダンスと下半期の就職プログラムをある程度分類できたことが挙げられる。一方、SA 代理募集の分類結果が適合率、再現率共に低かった原因としては、文書中の単語数の差が考えられる。SA 代理募集の抽出単語数は平均 20 語弱で、就職関連メールの平均 50 語前後の半分程度であった。

今後の課題は、システム単体としては再現率の向上、単語抽出アルゴリズムの橋渡し部を改良する、ユーザに合わせた分類結果を導くことができるように学習機能を持たせるなどが考えられる。また、本システムの機能をエディタなどに組み込み、ファイル保存時に自動的に保存先を選択させるという発展利用が期待される。

参考文献

[1]ChaSen's Wiki，http://chasen.naist.jp/hiki/ChaSen/

[2]岸田和明，“ 情報検索の統計モデル ”，人文学と情報処理，No.28，pp.6-15，2000.

[3]松尾豊，石塚満，“ 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム ”，人工知能学会論文誌，Vol.17，No.3，pp217-223，2002.

[4]大谷紀子，“ 情報検索におけるベクトル空間モデルの応用 ”，武蔵工業大学環境情報学部紀要，第五号，pp.99-109，2004.