2E-2

文書間の差異に着目したクラスタリング手法の提案

渡辺　匡[†]　　太田　学[‡]　　片山　薫[‡]　　石川　博[‡]

[†]東京都立大学工学部電子情報工学科 [‡]東京都立大学大学院工学研究科

1. はじめに

現在、インターネット等によって大量の文書データを容易に得ることが可能である。しかし、全ての文書に目を通すことは難しく、膨大な情報の中から我々の必要な情報だけを取得する手法は有用だと考えられる。

現在、大量の文書データに対し、クラスタリングによって類似文書をカテゴリーごとに分類する手法は広く研究されている。しかし、従来のクラスタリング手法ではカテゴリー分類が行われた後の処理については考慮されていなかった。そのため、いくつかの知りたい情報を含む文書が存在するとき、従来の手法ではカテゴリーによる分類情報のみが与えられ、文書間の違いについて知ろうとする際には各文書をひとつひとつ確認しなければならないという手間が生じていた。

そこで我々は文書間の差異に着目した新たなクラスタリング手法を提案する。この手法によって他の文書との差異部分についての情報を取得することで、ある文書を参照する際に他の文書から情報を補完することや、文書の取捨選択を効率的に行うことが可能となる。

今までにも同じカテゴリーに属する文書について差異を比較する方式は研究されているが、本研究ではカテゴリーの違いに関わらず差異の比較検討が可能であるという点で従来の研究と異なっている。

2. 関連研究

差異増幅機能を有する適合フィードバック検索[1]では、類似画像の検索のために差異増幅を用いる。我々はテキストに注目した。

差異に注目した複数文書融合手法[2]では、同じトピックの複数文書からそれぞれの文書に他文書の差異情報を付加する手法をとる。カテゴリ内での技術として我々と相補的である。

A document clustering method using differences between documents
Masashi Watanabe[†] ,Manabu Ohta[‡] ,Kaoru Katayama[‡] ,Hiroshi Ishikawa[‡]
[†]Electronics and Information Engineering, Faculty of Engineering, Tokyo Metropolitan University
[‡]Graduate School of Engineering, Tokyo Metropolitan University

3. 提案手法

3.1 概要

本手法は文書に対し形態素解析を行い、名詞を特徴語として抽出し、階層的凝縮型クラスタリング手法 (hierarchical agglomerative clustering algorithms)[3]によるクラスタリングを行う。結果を MDS(multidimensional scaling)[4]表示によって視覚的に表現し、任意のクラスを選択し、クラスの持つ要素をもとに再クラスタリングを行う。

3.2 手順

(1) はじめに形態素解析システム「茶筌」[5]を利用して各文書に対し形態素解析を行い、得られた品詞情報から名詞を抽出し、特徴語として扱う。

(2) この情報を元に各文書を特徴ベクトルとして表現する。このとき重み付けとして tf-idf 法を用いる。

(3) これにより階層的凝縮型クラスタリングを行う。n個の文書があるとき、はじめに 1 個の対象だけを含むn個のクラスタがある初期状態を作る。特徴ベクトルより対象X_iとX_jのユークリッド距離$D(X_i, X_j)$からクラスタ間の距離$D(C_i, C_j)$を計算する。そしてもっともこの距離の近い二つのクラスタを逐次的に併合する。この併合をクラスタがひとつになるまで繰り返すことで階層構造を獲得する。

(4) こうして得られた階層構造を元にいくつかのクラスタを取得し、MDS による二次元表示によって表現する。簡略化のため、各クラスタの最短距離及び最長距離を元に二次元配置し、表示するものとする。

(5) 任意の文書より文書間の差異となる特徴を各文書の持つ特徴量の絶対値差より計算し、上位数個を表示する。選択された任意の文書の持つ特徴のみを用いて再クラスタリングを行い上記の手順(3)～(5)を繰り返す。

3.3 プロトタイプ

上記の提案手法を用いたシミュレーション用のプログラムを作成した。実行画面を図 1、図 2

に示す。なお、このプログラムは開発段階のものとなっている。

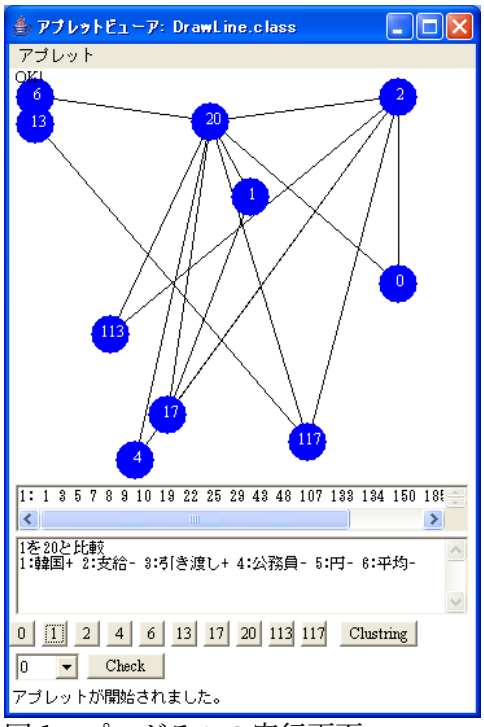

図 1　プログラムの実行画面

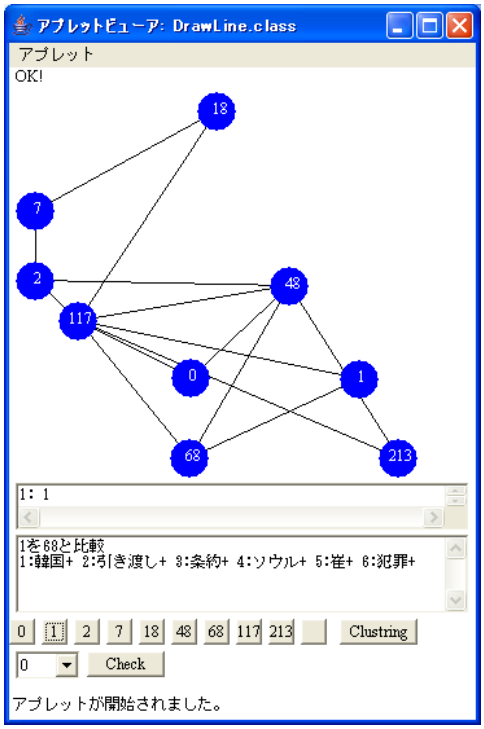

図 2　クラスタリング後の実行画面

図 1 はプログラムの実行画面を示す。画面上部にはクラスタの MDS による二次元地図が表示されている。それぞれの円が各クラスタを表し、円内に代表文書の ID が示されている。クラスタ間の直線はクラスタ間の距離を意味する。MDS 画面の下にある数値は選択されたクラスタに含まれる全文書 ID を示し、その下のテキストエリアは特徴語の差異を示す。「+」であれば比較したクラスタに対して特徴語がより多く含まれることを意味し、「-」はその逆を意味する。画面下部のボタンまたはセレクトタブから比較対象とする任意のクラスタ番号を選択し「Check」ボタンを押すことで特徴語の差異部分をいくつか表示する。図 1 では 1 と 20 を比較している。任意のクラスタ番号を選択し「Clustering」を選択することにより選択されたクラスタの代表文書内の特徴のみを用いて再クラスタリングを行う。番号 1 を選択し、再クラスタリングを行った結果を図 2 に示す。代表文書 1 の持つ特徴による再クラスタリングによって特徴語の差異部分が新たに計算されていることが分かる。

4. おわりに

本研究では、各文書の差異に着目し比較検討するクラスタリング手法について提案した。この手法によって任意の文書の持つ特徴から再クラスタリングを繰り返し、他文書との差異をより詳細に分析することが可能となったと考えられる。今後はより実践的な有用性を測るため、評価手法について検討する予定である。

謝辞

本研究の一部は，文部科学省科学研究費補助金特定領域研究(2)(課題番号:16016273)による。

参考文献

[1]木下真一,中島伸介,田中克己:差異増幅機能を有する適合フィードバック検索,情報処理データベースシステム 125-72,2001.
[2]渡邊拓也,大野成義,太田学,片山薫,石川博:差異に注目した複数文書融合手法,日本データベース学会論文誌 vol.3 no.1,2004
[3]クラスター分析: http://case.f7.ems.okayama-u.ac.jp/statedu/hbw2-book/node115.html
[4]F.W.Young: MULTIDIMENSIONAL SCALING, http://forrest.psych.unc.edu/teaching/p208a/mds/mds.html
[5]茶筌: http://chasen.aist-nara.ac.jp/hiki/ChaSen/