G-18

「K-means 法」によるテキストクラスタリング処理

Text clustering by K-means method

○横山航典[1]

*Kosuke Yokoyama[1]

Abstract: K-means method is one of the well-known methods for data clustering. The method has been successfully used in various fields, including natural language processing. This paper focuses on the application of K-means to corpus-division problems and reports some experimental results.

1. はじめに

「K-means 法」はクラスタリングの手法の 1 つで，「K 平均法」とも呼ばれており，多次元空間をクラスタリングするための汎用の教師なし分類法のことである[1].

本研究では「K-means 法」を実装し，「20Newsgroups」コーパスをテキストクラスタリングする．また，その結果どれくらいの精度でクラスタリングができたのか評価を行い，その結果について報告する.

2. クラスタリング

クラスタリングは，集団をある規則や共通項によって分類する手法である．中でもテキストクラスタリングとは，多数のテキストを自動でグループ分類する処理である[2]．また，「教師なし学習」と呼ばれる機械学習処理の 1 つでもある．「教師なし学習」は正解データがない問題で学習を行うことである.

3. コーパス

コーパスは，一般に言語資源で膨大なボリュームのことを指し，構造化されたテキストデータのことである．身近な例としては，「Wikipedia」が挙げられる[2]．コーパスを使用した理由は，実験データをあらかじめ収集・編集をしなくて良いことや品質が保証されているからである．本研究では公開コーパスである「20 Newsgroups」コーパスを用いて実験を行った．「20 Newsgroups」は 20 のカテゴリから成り，約 19, 000 件のニュースグループの記事で構成されている.

4. 自然言語処理

自然言語処理とは，文字や音声をコンピュータ上で処理するための技術である[2]．つまり，人間が理解できる文字・音声をディジタルデータに変換し，コンピュータに理解させることである[2]．最近では，自然言語処理の技術を応用した「Siri」や「自動翻訳」などのサービスが提供されており，更なる進化を遂げている.

5. TF-IDF

本研究では自然言語処理分野における主な手法である，TF-IDF やコサイン距離による類似度を使用して K-means 法の実装・評価を行った．TF-IDF (Term Frequency – Inverse Document Frequency)とは，文書に含まれる単語の相対的な重要性を表す数値として広く用いられている[1]．この値が大きいほど，特徴的な単語であることを示す．また，TF-IDF は TF (Term Frequency) と IDF(Inverse Document Frequency) を掛け合わせて値を求める.

TF はある文書中の単語の出現頻度として，TF =（ある文書中の単語の個数）/（単語の総数）で表される[2]． TF は分析する文書において，単語の個数が少ないほど値が減るような式である． IDF は，文書全体の単語の出現頻度として DF の対数を取ったものであり， DF =（単語が出現した文書数）/（文書の総数）で表される[2]．DF は文書全体を通して，出現頻度が高い単語ほど値が減るような式である．本研究では類似度についてコサイン距離を用いる．すなわち，類似度 =（単語 A と B の内積）/｛(単語 A の TF-IDF の絶対値）×（単語 B の

1：日大理工 ・ 学部 ・数学

TF-IDF)｝である[2].

6. K-means 法の実装-1

K-means 法は非階層型クラスタリングのアルゴリズムであり，自然言語処理分野においても広く使用されている．また，クラスタの平均によって数値やテキストなどの与えられたデータを K 個のクラスタに分類することから K-means 法と呼ばれている[2].

基本的な手順の流れについては以下の通りである:

(1) 与えられたデータを何個のクラスタに分類するのかをはじめに与える（K の値を決める）.

(2) K 個の重心点を用意する.

(3) 各データを最寄りの重心点に所属させる（Figure 1）.

(4) 再度重心点を計算しなおす.

(5) 更新された K 個の重心点を用いて，改めてそれぞれのデータを最寄りの重心点に所属させる.（Figure 2）

(6) (4)の重心点の再計算と(5)の各データを最寄りの重心点に所属させるという処理を繰り返し行う.

(7) 次第に重心点は動かなくなる．すると，クラスタリング処理が終了となる.

重心点 データ

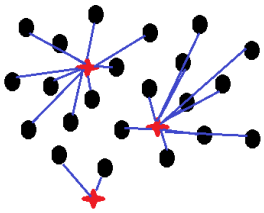

Figure 1
もとの重心点に所属

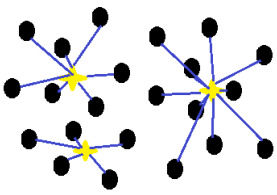

Figure 2
新しい重心点に所属

7. K-means 法の実装-2

実装効率改善のため，必要事項をクラスとして実装し，下記のような機能を加えた.

(1) 初期重心点の割り当て

初期重心点の割り当てによって処理速度が変化するため，互いに離れた重心点を割り当てることによって性能が改善される.

(2) 各データの重心点への割り当て

各データ間の距離を比べ，最寄りの重心点に所属させる．ここで各データとして，テキストから生成された TF-IDF ベクトルを使用する．また，各データ間の距離はコサイン距離を用いた類似度を使用する.

(3) 重心点の再計算

各重心点に属するデータより， 新たな重心点を計算する.

(4) クラスタリングの終了判定

新重心点と前重心点を比べ，変動が見られなくなったらクラスタリングを終了とする.

(5) TF-IDF

毎回 TF-IDF ベクトルを計算するのは通常非常に時間がかかるので，クラスとして実装する.

上記(1)～(5)を 1 つのクラスとし，これを含めて 5 つのクラスを用い以下 Table 1 に示す.

クラス名	主な内容
Corpus_20NewsGroups	コーパスを読み込む
Morpheme	形態素解析を行う
Vectorizer_TFIDF_English	TF-IDFベクトルの生成
Vector_Similarity	類似度の計算
Cluster_Kmeans	7節で述べたもの

Table 1.「K-means 法」の実装に用いたクラス

8. 最後に

上記のクラスを使用することにより，スムーズなクラスタリング処理を行うことができた．また「K-means 法」を実装し，評価した結果については学術講演会当日に報告する予定である.

9. 参考文献

［1］Matthew A. Russell 「入門 ソーシャルデータ」，株式会社オライリージャパン, July 2014

［2］斉藤常治，高橋佑幸 「PHP による機械学習入門」，株式会社リックテレコム，July 2014