ISS-A-063

主成分に基づく文書クラスタの意味抽出

吉原　堅斗　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

単語の出現頻度の類似によって得られたクラスタの意味を抽出することは難しい。本研究では主成分分析を用いて文書の意味を要素とする文書ベクトルを作成し、クラスタリングによって大量の文書を内容,意味でまとめたクラスタを生成する。さらに、生成したクラスタからの意味抽出を試みる。

2. 主成分分析と因子得点

主成分分析では、全体のばらつきを最も表す最大特徴を主成分として抽出することができる。この主成分と文書を軸にした因子得点を求めることで、単語の頻出頻度のみを表すベクトルを文書の意味を表すベクトルに変えて表現できる。さらに、因子得点を用いてクラスタに影響する主成分が求め、その意味が得られればクラスタの意味を抽出することができる。

3. 実験

(1)実験準備

本研究で用いるコーパスは､毎日新聞 2 週間分で、出現回数 40 回未満と不要語を削除した二文字名詞のみ。

(2)実験手順

因子得点で得られた文書ベクトルを k-means++ 法でクラスタリングする。なお類似比較は余弦類似度で行う。クラスタ数 80 に設定し、因子得点の寄与率が高い 200 の主成分をクラスタリングするデータとして扱う。

評価尺度には、クラスタが意味的なまとまりをもつか調べるため、各クラスタで頻出の高い固有名詞 10 個から内容を類推する内容評価と記事の 1 文目に含まれる名詞がクラスタの主成分への重みが大きい名詞(10～30 個)を含む割合をみる数値評価を行う。また、前 2 つの評価で最も内容に統一性がないとみるクラスタの意味抽出を行う。この意味抽出では主成分に影響する名詞を 50 語用いる。

(3)実験結果

表 1　内容評価と数値評価

主観的評価	数値評価	条件を満たすクラスタの割合
○	○	59%
○	×	12%
×	○	6%
×	×	23%

内容評価○:類推した内容が実際の記事と一致
数値評価○:40%以上が主成分に影響する名詞あり

表 2　内容に統一なしクラスタの意味抽出(一部)

主成分に影響する名詞はジャンルでまとまる。

ジャンル	主成分に影響する名詞	類推内容
国際	シリア,北朝鮮,EU,離脱	反米国家,EU離脱問題
社会	県警,容疑,不正,メール	国内の刑事,民事事件
五輪	東京オリンピック,委員会	2020年東京五輪準備

4. 評価

内容、数値評価から 59～77%のクラスタが内容に統一性あるか主成分の意味にまとまる。一方、内容に統一のないとみる残りの 23～41%のクラスタは、主成分に影響する単語をジャンル分けすることで複数の内容からなる意味を抽出することができる。

5. 結論

主成分に基づく文書クラスタリングを行ったことで内容に統一性のあるなしに関わらず意味抽出ができたが、複数の内容を持つクラスタが本当にその意味を持つか確かめることは難しい。