ISS-P-040

主成分分析とギブスサンプリングを利用した代表文書の抽出

小林　千真　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

現代では文字の情報溢れている。その中から代表的な意味を持つ文書を取り出すことは有益である。

2. 提案方法

文書を単語頻度として扱うと単語自体が持つ曖昧性や多義性の影響から文書の意味を捉えきれないことが考えられる。そこで、あいまい性や多義性を除去した文書ベクトルを扱うために主成分分析で次元削減をして、共起する単語をまとめることで、似た文脈で出現する単語をより大きいが概念でとして捉える。さらに、マルコフ連鎖モンテカルロ法を使用することで、文書の関連性を考慮する。これはある文書からある文書へ遷移する際に意味のつながりを持ちながら遷移し、最終的に定常分布へと収束する。これこそが代表的な分布といえると考える。そして、定常分布との cos 類似度が最大の文書が代表文書であると考える。

3. 実験

3.1. 実験手順

文書集合に主成分分析を行い、その文書集合にギブスサンプリングを実行し、定常分布を得る。主成分分析の寄与率の閾値を 0.5 とし、50000 ラウンド行った。

また、評価方法として、定常分布における絶対値に各主成分の寄与率をかけた値の上位 5 主成分に解釈を与え、その内容が代表足りうるかを判断する。代表文書は定常分布とどの程度一致しているのか実際の文書を読み、定常分布において大きな特徴を表す主成分の内容を代表文書が含むか否かという視点で筆者が判断した。ただし、〇:含む、×:ふくまない、△:判断が難しいという評価基準とした。

3.2. データセット

本研究ではコーパスとして、毎日新聞 2017 年 1 月 3 日から 2 週間分の経済ジャンルを対象とした。

3.3. 実験結果

30000 ラウンド以降代表文書は文書 161 と文書 221 の 2 つのみであり、最終的な代表文書はこの 2 つの文書と判断した。

3.4. 評価と考察

3.1 で述べた評価方法の結果を以下に示す。まず、定常分布の上位 5 主成分の解釈結果を表 1 に示す。

表 1 で示した各主成分は自動運転や旅行などさまざまな内容であったが、すべてに共通して「日本と世界の比較・関係性」というと内容であった。文書集合におけるこれら以外に多く出現していた内容として、「トランプ大統領就任による米中間の関係」や「日本国内の電気とガスの小売り自由化」といった「日本と世界」という構図になっていないものであった。対象コーパスでもっとも多かった内容は「日本と世界」を比較した内容であり、新聞記事の経済ジャンルの内容として「日本と世界」の比較をする内容は代表いえると考えた。また、これは網羅的な意味をカバーしているという意味の代表ではなく、意味的に文書集合で最も述べられている内容を示しており、文書集合の特徴を表すような代表である。次に代表文書と主成分の関係を表 2 に示す。

表 1. 主成分の解釈

第二主成分	日本企業と世界
第三主成分	自動運転の開発
第四主成分	中国との貿易
第五主成分	旅行や観光
第六主成分	企業と投資

表 2. 代表文書と主成分の関係

	文書161	文書221
第二主成分	△	〇
第三主成分	〇	×
第四主成分	×	×
第五主成分	×	〇
第六主成分	〇	△

表 2 から、文書 161 と文書 221 では定常分布に対して含んでいる内容が異なることが分かる。このことから、代表文書は複数抽出する必要があると考える。また、代表文書を 1 つに絞ると、定常分布の 2/5～3/5 の内容を含むことが分かった。

4. 結論

本研究では、主成分分析による語義曖昧性の除去とマルコフ連鎖モンテカルロ法による意味の追跡により定常分布として文書集合における代表性を示した。また、定常分布に対応する代表文書はその特徴の 2/5～3/5 の内容を 1 文書で表現するが、定常分布の表す内容を表現するためには代表文書を複数抽出する必要があることを確認した。

参考文献

[1] 木村　淳,吉富　康成, 田伏　正佳,”単語頻度を用いた文書分類と代表文書の抽出”,情報処理学会報告自然減資後処理,Vol12,p117-120,2013
[2] 手塚太郎.仕組みが分かるベイズ統計と機械学習.朝倉書店,2019