ISS-A-068

主成分分析とギブスサンプリングによる代表文書の解釈

長井 慶成　　三浦　孝夫
法政大学理工学部創生科学科

1. 目的

代表文書の持つ意味解釈の方法の提案とその妥当性の検証を目的とする。

2. 主たる主張

はじめにこの論文では代表文書の定義として文書集合全体の意味を表す平均的な文書であるとする。

文書は単語の集合であり、出現する単語から文書の意味解釈を行うことは容易に思われる。しかし、単語の中には類義性、多義性、共起性を持つものも存在し、これらは文書の解釈の幅を広げるため、単語の分布から解釈を行うと様々な解釈が可能であり詳細な内容を把握するのは困難である。

3. 提案

主成分分析を行い類義語、共起する語をまとめ、単語の分布ではなく意味ベースで考える。各主成分は意味を表すことから、これらをもとに代表文書の詳細な意味解釈を行う。

文書を意味単位としてみると、意味と意味にもつながりがあると思われ、単語集合と見た時と同じように極限があることが期待されることからギブスサンプリングを用いる。

4. 実験データ

毎日新聞の 2017 年 1 月 3 日から 14 日までのスポーツ記事

5. 結果

ギブスサンプリングを 1000 ラウンド行い、終わり 30 ラウンドに最大類似文書として複数回出現した文書に収束したと判断する。

ラウンド	最大類似文書	ラウンド	最大類似文書	ラウンド	最大類似文書
971	文書812	981	文書988	991	文書907
972	文書296	982	文書2184	992	文書1807
973	文書686	983	文書1809	993	文書2156
974	文書1573	984	文書529	994	文書908
975	文書1465	985	文書700	995	文書1807
976	文書2056	986	文書988	996	文書907
977	文書2057	987	文書1143	997	文書907
978	文書335	988	文書994	998	文書1806
979	文書165	989	文書753	999	文書907
980	文書1416	990	文書783	1000	文書1699

表 1：最大類似文書の推移

結果として文書 907、文書 988、文書 1807 が収束文書と判断できる。今回はセンバツ高校野球の運営委員会に関する記事である文書 1807 との類似度が最も高いベクトルを代表文書ベクトルとした。このベクトルが大きい成分を持つ主成分の意味から代表文書の解釈を行う。

第x主成分	解釈
4	スキージャンプ
5	高校センバツ野球２1世紀枠
6	毎日新聞
16	大会優勝候補
23	事前の対策
26	安全対策と都道府県高野連の承認
29	ある付近
35	和歌山
40	会長の任期満了

第x主成分	解釈
44	選手の評価
45	選手の獲得メダル
54	出場資格停止処分
59	女性との協力
67	若手の指導
71	障害を持つ選手
78	大統領との会談
81	結婚のお祝いムード

表 2：主成分の解釈

各主成分の解釈から代表文書の解釈は次のようなものが考えられる。

トピック	詳細
スキージャンプ	大会優勝候補の紹介
	試合の結果
高校センバツ野球	安全対策と高野連の承認
	授与されるメダル
	女子部員の参加

表 3：代表文書の解釈

6. 考察

代表文書の主な内容はスキージャンプと高校センバツ野球の 2 つであると考えられる。

スキージャンプは w 杯や国内大会が開催されるなど当時の関心事であると推測でき、また高校センバツ野球に関しても毎日新聞社が主催していることや前年問題となった女子部員の練習参加の可否の動向が注目されていた点を踏まえると、代表文書の意味として合理的である。

7. 結論

ギブサンプリングによって得られた代表文書ベクトルは、各主成分から意味を詳細に解釈することができた。またその内容も当時の世間の関心事であることから、この解釈方法の妥当性は保証されるものと考えた。

Copyright © 2021 IEICE