ISS-P-045

MCMC 法で導き出された定常的単語

垣上　南帆　　三浦　孝夫

法政大学理工学部創生科学科

1.前書き

本研究では文書集合に属する単語の延べ語数や文書数の最大値や平均を計算する以外の方法で、平均的で代表的な単語の導出をするために MCMC 法を用いた極限分布の導出を利用している。

2.提案手法

文書集合の平均的・代表的な単語・単語集合を導き出すことを目的とする。代表的な単語とは複数の文書に一定の割合で使われており、単語の意味が一定である語とする。語の導出及び代表的な理由を解析する。

導出はマルコフ連鎖モンテカルロ法(MCMC) 特に Gibbs Sampler 法を用いる。MCMC は多次元の確率から近似的に極限分布を求められる。

単語を当該語の文書内延べ語数(TF)ベクトルで表現する。テスト語ベクトルとして{1,1,1…1}のような全 TF=1 の適当なサイズのベクトルを生成し、この TF を 1 つ削除・1 つ挿入を全ての TF で繰り返す。例にテスト語ベクトル$TW = \{スポーツ記事1, コラム2,, 経済記事5\}$からスポーツ記事 1 を削除、この文書の確率分布 $P(x|0,$コラム 2,,経済記事 5)から直接期待文書生成はできないため{コラム 2,,経済記事 5}と最類似ベクトルを持つ単語の文書生成確率を利用して文書生成し、この文書をテスト語ベクトルに追加する。一連を 1 ラウンドとして収束するまで繰り返す。

しかし現実的には完全に収束するまで試行を繰り返すのは難しく、語ベクトル同士の類似度が上昇したまま安定し最類似単語がほぼ変化しない状態を疑似収束:Burn-in したとする。この時のテスト語ベクトルを GS の収束ベクトルとして実験を終了する。収束ベクトルと最類似するベクトルを持つ単語・単語集合を定常的単語として結果の評価をする。

3.実験

本研究では 2017 年 1 月 1 日~14 日の毎日新聞記事 2443 件、2274 見出し語を扱う。

見出し語に含まれる不要語 Stop Word、動詞・形容詞・数詞などは取り除き、文書サイズの小さい記事:コマーシャルや天気予報、ほぼ数値で書かれた記事は取り去り、実際に扱ったデータは記事数 2051 件、総見出し語が 1692 個である。

テスト語ベクトルはサイズ 20 で生成し、初期値の文書を変えて二回実験を行う。

類似度が上昇し安定した状態で特定の単語や単語の持つ意味が近い単語集合から 100 ラウンド以上最類似単語ベクトルが遷移しなくなると Burn-in したとみなす。ベクトル同士の類似度計算は余弦類似度を用いる。

定常的単語は客観的にも定常的であると考え、同じ時期の報道やその年の重大ニュースと比較。定常的単語の登場する文書や持つ意味など全体的に共通点が大きいほど良い。また新聞は重要度の高いものほど紙面を割いているものとして、毎日新聞記事の中で定常的単語・単語集合が登場する文書や、そのジャンルの記事数から実験結果を評価する。

4.結果

ラウンド数	類似単語	類似度
1	事前	0.77302
63	帝京	0.88107
68	帝京大	0.92621

図 1:実験 1 の類似単語と類似度の遷移

「帝京」と「帝京大」は共通した記事に登場し同じ意味の単語として使われ、実験 2 でも同じ値に収束したためこれを定常的単語とする。

2 単語が共通して登場する文書は 6 つ。いずれもラグビー全国大学選手権第 53 回に触れた記事である。2 単語は“帝京大学”を指す単語だがこれだけでは大学自体を表しているのか属する何かを表すのかわからない。

参考に実験 1 の収束ベクトルとの類似度が高い上位単語から「FW」(類似度:0.475)を挙げて「帝京」「FW」を含む 2017 年 1 月 1 日～14 日のニュースを調べると、サンスポや Yahoo! Japan で定常単語が属する記事と定常単語の意味と内容が近い記事が見つかる。

また毎日新聞記事 2443 件を 13 のジャンルに分けた時、スポーツ記事は 487 件で二番目に大きいジャンルであり(平均 169 記事)、毎日新聞記事内で多く紙面を割かれているといえる。よってスポーツ関連の単語が定常的単語・単語集合になるのは悪くないと評価する。

5.結論

GS に従って収束した単語・単語集合を定常的としたが、収束ベクトルとその類似ベクトルを持つ単語集合を定常的単語集合とする方が文書集合の平均や内容を正確に表すことができるのではないかと考える。

文書集合を変え違う実験をした際定常的単語集合がすべて曖昧で特徴を欠いた場合も同じ評価ができるよう確立した評価方法が必要である。

6.参考文献

1. 手塚太郎「マルコフ連鎖モンテカルロ法」『しくみがわかるベイズ統計と機械学習』朝倉書店,2019,p164~175