記事集合の話題抽出

横沢 薫　白井匡人　三浦 孝夫
法政大学理工学部創生科学科

1. 前書き

インターネットが普及し、スマートフォンが広く用いられる現代、膨大な量の情報で、必要な情報の発見が極めて困難となった。大量の情報を素早く要約しわかりやすく表現する技術は必要不可欠である。それらを目的とし、Reuter Corpus('98.8.20-)の新聞記事から、話題の自動抽出、テーマの推測を行う。

1. 前処理

まずは記事本文の不要語を取り除く。次にステミングを行う。最後に Zipf の法則を使い、中頻度以下の単語の削除を行う。

2. ベクトル化

ベクトル化は記事本文について行う。これらを文書ベクトルと呼ぶことにする。1 つの記事を文書ベクトルで表現する。

3. 近傍と交叉

記事同士の余弦類似度を求める。閾値を設け、ある記事に類似した記事の集合を作る。今回は閾値を 0.45 とした。また閾値を設け、集合同士を比べる。記事の共通部分が幾つあるのかを調べる。

4. クラスタリング

生き残った記事集合同士、文書ベクトル同士を合計する。これらの余弦類似度を求める。ベクトルとベクトルの距離を「余弦類似度を 1 から引いたもの」と定義し、クラスタリングで 20 個のクラスタに分割する。

5. 分布化

記事トピックごとに出現する単語の頻度を分布化する。本研究では、トピック自体の出現回数が多いものトップ 20 のみの語分布を求めた。クラスタごとの文書ベクトルも語分布化する。

6. 話題抽出と検証

両者の分布の KL 情報量の逆数を求め、比較する。KL 情報量の逆数が大きかったトピックの語分布のトップ 5 をそのクラスタのトピックと仮定し本文を実際に読み検証する。

7. 実験結果

表 1 検証結果

		記事番号							
クラスタ番号	1	2606	2583	6825	6086	4452	1930	10089	3171
適合率	0.5	TRUE	FALSE	TRUE		TRUE		TRUE	FALSE
クラスタ番号	2	11053	2583	6620	6825	925	4452	10089	3171
適合率	0.666667		FALSE	TRUE	TRUE			TRUE	
クラスタ番号	3	11053	2606	2583	6620	4452	925	6086	10398
適合率	0.5		TRUE	FALSE	TRUE				
クラスタ番号	4	11053	2606	2583	6825	4452	6086	10398	10089
適合率	0.642857		TRUE	FALSE	TRUE				TRUE
クラスタ番号	5	11053	6620	6086	925	10398	10089	1930	3171
適合率	0.454545		TRUE				TRUE		
クラスタ番号	6	6645	5175	10316	916	4452	6086	1081	10089
適合率	0.333333							FALSE	TRUE
クラスタ番号	7	11053	2606	2583	6620	6825	925	6086	4452
適合率	0.6875		TRUE	FALSE	TRUE	TRUE			

記事によっては最初の 2 行に記事内容がなく、検証できないところもいくつかみられた。適合率の最大値は 0.6875、最小値は 0.25、平均は 0.52 だった。

8. 考察

適合率が高かった要因は、記事数が少なかったことが考えられる。同じ話題について言及している可能性が高い。また、検証が実際に読んでの検証で主観が介入している。ここの改善も、より精度の高い結果への道なのかもしれない。

9. 結論

記事に前処理をし、ベクトルで表した。ベクトル化した記事の近傍、交叉をとり、淘汰した。それらを分割、語分布化し、KL 情報量を用いて比較を行った。適合率は 0.5 を超え、高い値が得られた。