7T-06

単語の出現頻度に基づくテキストの話題分割とラベリング

柿本　雄輝†　　毛利　元昭‡　　打矢　隆弘†　　船瀬　新王†　　内匠　逸†
名古屋工業大学　工学研究科　情報工学専攻†　　愛知大学　経営学部　経営学科‡

1. はじめに

近年，インターネット上の情報提供や配信サービスなどにより，電子ニュースなどのネットワークを介したテキスト情報の配信が盛んに行われている．膨大なテキスト情報が日々蓄積される中で，ユーザが欲しい情報を得るために，クラスタリングや話題抽出などのテキストマイニング技術が重要視されている．ユーザがテキストから欲しい情報を得る際に，テキストの構成を把握することができれば，テキスト全体を読むことなく必要な部分のみを得ることが可能となる．

文書から話題を抽出する方法として，複数の文書を用いるものであれば，独立成分分析や潜在意味解析(LSA) といった手法が挙げられる．これらは，文書間で比較を行い，話題に特有な単語を抽出することが多く，単一の文書の中の話題分割には用いることができない．そこで，単一の文書から得られる単語の出現頻度や共起情報に着目し，話題分割及び話題の推移について調査する．

2. 従来手法

テキストの話題分類を行う研究は多く行われているが，その中でも単一のテキストから話題区分を行うTextTiling[1]について述べる．TextTilingは，テキストを出現している単語の並びとして置き換え，各文間から前後N語の単語群のコサイン類似度に応じて，話題の境界かどうかを判別する．コサイン類似度は以下の式(1)で求める．

$$\mathbf{sim}(\mathbf{b_l}, \mathbf{b_r}) = \frac{\sum_t w_{t,b_l} w_{t,b_r}}{\sqrt{\sum_t w_{t,b_l}^2 \sum_t w_{t,b_r}^2}} \qquad (1)$$

$\mathrm{b_l}, \mathrm{b_r}$は各文間から前$N$単語，後ろ$N$単語分の窓幅，$w_{t,b_l}, w_{t,b_r}$はそれぞれの窓幅における単語$t$の出現頻度を表す．このコサイン類似度の値が極小値を取った点を話題境界であると定める．

この手法では，文間の前後のみを用いているため，離れた位置の文章の関係性などは考慮できていない．そこで本稿は，テキスト中の話題の推移に着目し，時間の推移に伴った話題の変化の可視化を目的とする．

3. 提案手法

特定の単語群がある区間で複数回出現している時に，話題を形成していると定義し，テキスト中の単語の出現傾向を確認する．以下の手順で作成した表を頻度分布表と定義する(図 1)．

1. テキスト中に出現する全単語を抽出
2. 連続 5 文での各単語の出現した文数を記録
3. テキスト全体を通し連続 5 文での出現文数の最大値が 1 であったものを表から削除
4. 連続 5 文での出現文数の最大値を取る位置がテキストで早く訪れている順に単語を並び替える

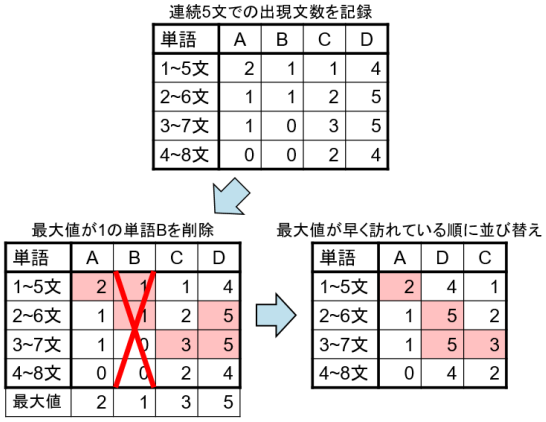

図 1　頻度分布表の作成手順

手順 1,2 では，形態素解析を利用し，名詞のみを抽出する．形態素解析は日本語形態素解析プラグイン lucene-gosen[2]を利用した．横に単語を出現順で並べ，縦に連続 5 文中の各単語の出現文数を記録した表を作成する．

手順 3 では，出現文数の最大値が 1 であった単語を表から除外するが，該当する単語は話題を形成しないと判断したためである．

手順 4 では，各単語が最も出現している箇所に注目し，同じタイミングで多く出現した単語同士を近づける目的がある．

また，これらの手順とは別に，視覚的にわかりやすくするため，出現文数の値が大きいほど濃い色で表示させる．これにより，縦に濃い部

Text topic segmentation and labeling based on word frequency
† Yuki Kakimoto, Takahiro Uchiya, Arao Funase, Ichi Takumi, Nagoya Institute of Technology
‡ Motoaki Mouri, Aichi University

分が広がっている場合，少ない単語数で長い間話が続いていること，逆に横に広がっている場合，多くの単語が使われているにも関わらず話が長く続かないことがわかる．

4. 検証

4.1 使用データ

提案手法の適用にあたり，解析の対象は論文とした．論文は数あるテキスト情報の中でも話の繋がりや流れを追えるように書かれていることが多く，単語の出現傾向の調査に適していると考えられる．今回は文献[3]を用いた．あらすじ部分は話の流れと別に書かれていることと，章タイトルや図表などは文の形になっていないことを考慮し，それらを削除した本文のみをテキスト情報として利用した．単語の種類が 186 語，文数が 81 文から成り立っている．図 2 にこのテキスト情報を解析して得られた頻度分布表を示す．

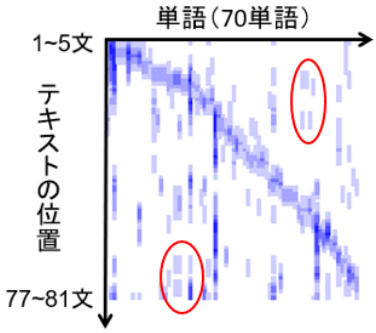

図 2　頻度分布表

4.2 考察

各単語がテキスト中でどのように出現しているかを確認することができた．また，テキスト全体で幅広く出現している単語とそうでない単語があることも確認した．幅広く出現している単語の中でも，図 2 中の赤丸で示したような同タイミングで出現（共起）している単語同士に強い関わりがあると考え，各単語同士での出現頻度を用いて相関行列を作成する．

5. 相関行列

単語の並びは頻度分布表作成の手順 4 で並び替えたものを使用し，各単語同士の出現頻度の相関を要素にした相関行列を作成する．

相関はピアソンの相関係数を用いて算出し，その値が 0 の時に白色，正の相関があるとされる 1 に近づくほど赤色に，負の相関があるとされる -1 に近づくほど青色になるように示した表を図 3 に示す．

頻度分布表作成の手順 4 により，出現文数が最大となる箇所が固まっているため，対角行列を挟む正方形の範囲で相関係数が大きくなっている．また，正方形の範囲を広げていくと，負の相関が混ざるようになるが，その単語は話題に含まれない単語であると考えることができる．そこで，対角を挟む正方形の全要素が，弱い相関があるとされる 0.2 以上で構成される場合に一つの話題であるとして分割したものを図 4 に示す．

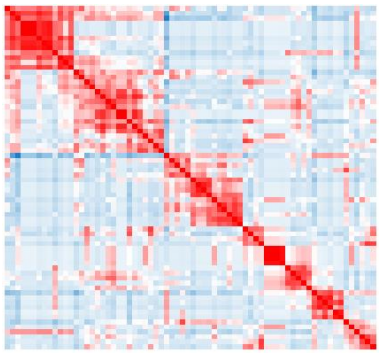

図 3　文献[3]の出現頻度の相関行列

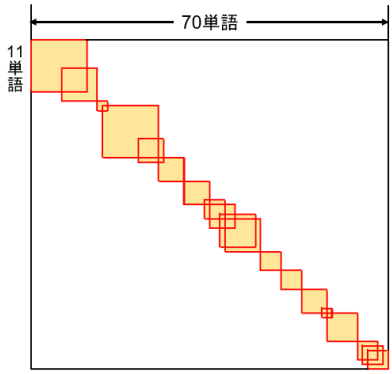

図 4　単語の話題による分割

19 個の話題として分割され，最大 11 単語，最小 2 単語で構成されていた．また，正方形が重なっている部分の単語によって話題が推移していると言える．

6. まとめ

本稿では，テキストの出現単語から頻度分布表を作成し，単語の出現傾向を調査した．その出現傾向の相関を取ることにより，テキストに出現する単語の話題分割を行った．

相関から話題分割を行う上での細かい検討や他のデータについても検証と評価を行う必要がある．

参考文献

[1] Marti A. Hearst, “TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages”, (1997).

[2] 日本語形態素解析プラグイン lucene-gosen https://code.google.com/archive/p/lucene-gosen/

[3] 丸田要他，“自然な対話継続のための推移する話題推定”，言語処理学会（2018）.