1R-02

# 凝縮性に基づく注釈単語検出法とその評価

小林　えり†　　斉藤　和巳†　　大久保　誠也†　　池田　哲夫†
†静岡県立大学大学院 経営情報イノベーション研究科

## 1　はじめに

近年，ビックデータの分析が盛んに行われ，テキストデータに対するトピック検出と追跡 [1] や，ホットトピック抽出 [2] などの多様な研究が展開されている．これに対し，文書クラスタリングを行い，得られたクラスターの解釈として TFIDF [3] より解釈語を抽出する方法などがあるが，TFIDF では必ずしもその文書を特徴付ける単語が抽出されるとは限らない．そこで本研究では，クラスタ係数の概念 [4] を発展させた“ 凝縮性（cohesiveness）”と呼ぶ指標を提案する．クラスタ係数は，あるノードに隣接する任意のノードペア間にリンクが存在する期待値で定義されるのに対し，我々の提案する単語の凝縮性は，文書全体での平均類似度と比較し，その単語を含む文書ペア間の平均類似度が有意に大きいかの z スコアで定義する．一般に，凝縮性の高い単語は，類似した内容で共通する話題を持つ文書集合と隣接する傾向となり，非隣接文書には出現しない識別的性質を持つ．そのため，その単語自体が文書集合の適切な解釈語になることが大いに期待できる．

## 2　提案手法

文書集合を $D = \{\mathbf{d}_1, \cdots, \mathbf{d}_N\}$，単語集合を $W = \{w_1, \cdots, w_K\}$ とする．ここで，文書 $\mathbf{d}_n$ は文書を形態素解析して得られた単語頻度ベクトルであり，$k$ 番目の要素 $d_{n,k}$ は文書 $\mathbf{d}_n$ での単語 $w_k$ の出現回数を示す．また，文書ペア $\mathbf{d}_i$ と $\mathbf{d}_j$ 間の類似度 $\rho(\mathbf{d}_i, \mathbf{d}_j)$ は以下の式で定義される．

$$\rho(\mathbf{d}_i, \mathbf{d}_j) = \sum_{k=1}^{K} \frac{d_{i,k} \times d_{j,k}}{\sqrt{\sum_{k=1}^{K} d_{i,k}^2} \sqrt{\sum_{k=1}^{K} d_{j,k}^2}} \tag{1}$$

(2)

ただし，$0 \leq \rho(\mathbf{d}_i, \mathbf{d}_j) \leq 1$，$\rho(\mathbf{d}_i, \mathbf{d}_i) = 1$ とする．ここで文書集合 D 全体での文書ペア間の平均類似度 $\mu(D)$ は次式でとなる．

$$\mu(D) = \sum_{\mathbf{d}_i \in D} \sum_{\substack{\mathbf{d}_j \in D \\ i \neq j}} \frac{\rho(\mathbf{d}_i, \mathbf{d}_j)}{|D|(|D| - 1)} \tag{3}$$

同様にして，単語 $w_k$ が出現する文書集合を $D(w_k)$ とすれば，$D(w_k)$ における文書ペア間の平均類似度 $\mu(D(w_k))$ は次式で求まる．

$$\mu(D(w_k)) = \sum_{\mathbf{d}_i \in D(w_k)} \sum_{\substack{\mathbf{d}_j \in D(w_k) \\ i \neq j}} \frac{\rho(\mathbf{d}_i, \mathbf{d}_j)}{|D(w_k)|(|D(w_k)| - 1)} \tag{4}$$

よって，文書数も考慮して $\mu(D(w_k))$ が $\mu(D)$ より有意に大きいかを示す z-スコア (単語 $w_k$ の凝縮性 $z(w_k)$) は次式で計算できる．

$$z(w_k) = \frac{\mu(D(w_k)) - \mu(D)}{s(D)/\sqrt{|D(w_k)|(|D(w_k)| - 1)}} \tag{5}$$

ここで，$z(w_k)$ には中心極限定理が成り立つ．文書集合全体での類似度の標準偏差を $s(D)$ とすると，文書集合 $D(w_k)$ での総文書ペア数 $|D(w_k)|(|D(w_k)| - 1)$ に対する標準偏差は $s(D)/\sqrt{|D(w_k)|(|D(w_k)| - 1)}$ となる．よって，極めて少数の文書にしか出現しない単語は，それら文書ペア間の類似度が高くても凝縮性 $z(w_k)$ が相対的に過度に大きくならなくなる．

## 3　評価実験

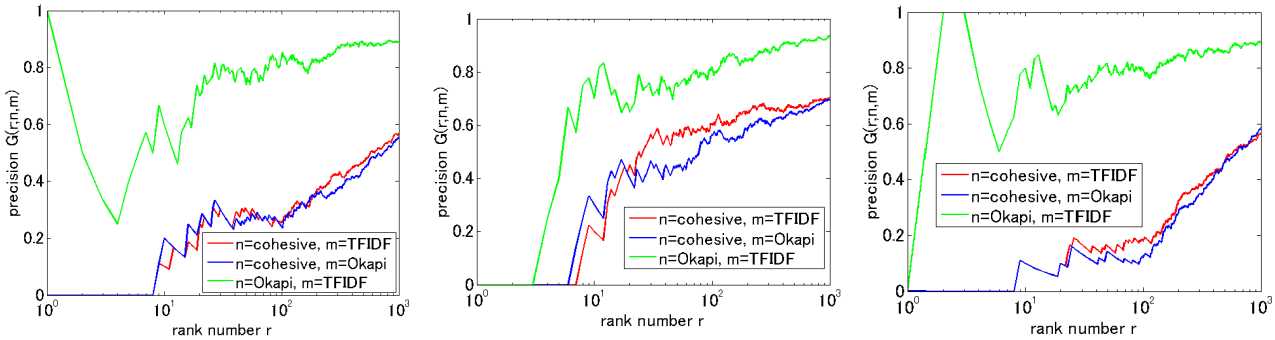
提案指標の特徴と有用性を明らかにするために，実データに対して提案指標と既存指標を適用し，それらの結果の比較を行った．具体的には，提案指標ならびに，広く用いられている“ TFIDF ”，TFIDF を応用して開発された“ OkapiBM25 ”の 3 つの手法による単語注抽出を行い，それぞれの単語抽出の相違を定量的に評価した．加えて，抽出した単語集合が文章集合の適切な解釈語になり得るかを定性的に評価した．なお，Okapi BM25 の入力パラメータは“ $b = 0.75$ ”，“ $k_1 = 2.0$ ”と設定した結果を用いる．

実験データとして，以下の 3 種類の記事データを用いた．1 つ目は，YahooNews の 2013 年 12 月 21 日から 30 日までの 10 日間内で記載された約 10,000 記事，2 つ目は，毎日新聞の 1992 年から 2002 年の国際ジャンルの記事のみを対象とした約 70,000 記事，3 つ目は，明治時代の静岡県の新聞記事データベースである近代新聞の約 120,000 記事である．それぞれの記事データセットの総出現単語（ターム）数は順に 38,425，51,030，30,889 である．

Extracting annotation words based on cohesiveness and its evaluation
†Eri KOBAYASHI †Kazumi SAITO †Seiya OKUBO †Tetsuo IKEDA
†Graduate School of Management and Information of Innovation, University of Shizuoka



Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.




(a) Yahoo News　(b) 毎日新聞　(c) 近代新聞

図 1: 一致度による評価

### 3.1 定量評価

本研究では抽出単語集合の違いに対する定量評価として，各指標より抽出された解釈語の一致度を採用する．$r$ をランク数とし，$V(r)^{\text{cohesive}}$ を凝縮性の高い上位 r 位以内の単語集合，$V(r)^{\text{TFIDF}}$ を TFIDF 値の高い上位 r 位以内の単語集合，$V(r)^{\text{Okapi}}$ を Okapi BM25 指標値の高い上位 r 位以内の単語集合とすると一致度は以下の式で定義される．

$$G(r;n,m) = \frac{|V(r)^n \cap V(r)^m|}{r} \tag{6}$$

ここで $n,m$ は {cohesive,TFIDF,Okapi} のいずれかをとる．図 1 に $r = 1$〜1000 までの各データセットでの結果を示す．横軸にランク数 r を，縦軸に各手法ペアでの一致度をとる．図 1 より，どのデータセットでも提案指標と既存指標との一致度は低く，ランク数を増やせば一致度は上がるものの，最大ランク数 ($r = 1000$) でも半分近くが異なる単語を解釈語として抽出している．一方，TFIDF とその応用指標である OkapiBM25 による一致度はどのデータセットでも比較的高く，ランク数を増やすと抽出語の差異はほとんど見られない．このことから，提案指標は従来指標とは異なる観点から特徴語を抽出していると定量的に評価できる．

### 3.2 定性評価

抽出される単語集合が解釈語として適切か否かを定性的に評価する．YahooNews の場合，提案指標の高い上位 5 位の単語は { 靖国神社，参拝，安倍，首相，晋 } であり，この時期に騒がれた“ 安倍首相靖国参拝問題 ” に関する単語が抽出された．さらにランク数を増やしていくと，“ スケート ”,“ ソチ ”などソチ五輪（予選）に関する単語が見られ，抽出単語が文章集合内のイベント，話題等の把握に貢献できると期待できる．一方，TFIDF では { 円，月，人，年，日本 } が，OkapiBM25 では { 円，さん，日本，市，氏 } が抽出され，ランク数を増やすと，“ 韓国 ”,“ 中国 ”などの靖国問題関連の単語が抽出され始めるものの，解釈語とは言い難い単語が上位に抽出した．この他の毎日新聞，近代新聞の場合でも提案指標は文章内のイベントやジャンルといった一種のまとまりを示すような単語が上位に抽出されているのに対し，既存指標は解釈語には不適な単語が抽出されていた．このことから，提案指標は既存指標よりも文書集合を特徴付ける単語の抽出が期待できることが確認できた．

## 4 おわりに

本研究ではテキストデータに対するトピック検出，キーワード抽出に関する研究への貢献を目的に，新たな凝縮性と呼ぶ指標を提案した．そして評価実験により，定量的に既存指標との差異を，定性的に抽出単語が解釈語として適切だと期待できることを確認した．今後は多様なデータセットでの検証や時系列を考慮したキーワード抽出への応用を目指す．

**謝辞** 本研究は科学研究費補助金基盤研究 (C)(No. 23500312) の補助を受けた．

## 参考文献

[1] J. Allan, “ Topic detection and tracking: “ event-based information organization ”,Kluwer Academic Publishers, (2002)

[2] J.Kleinberg, “ Bursty and hierarchical structure in streams ”, Data Mining and Knowledge Discovery, 7: 373-397, (2003)

[3] C.Manning, P. Raghavan, H. Schutze, “ Introduction to information retrieval ”, Cambridge University Press, (2008)

[4] D. Watts, S. Strogatz, “ Collective dynamics of ’small-world’ networks, ” nature 393: 440-442, (1998)

[5] Stephen Robertson and Hugo Zaragoza, “ The probabilistic relevance frame-work: Bm25 and beyond ”, Now Publishers Inc, (2009)



Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.