D-019

語の関係性を抽出した特徴ベクトルによる重要語抽出法の評価

Evaluation of Keyword Extraction Method with Feature Vector of Term Correlation

今井 智宏 † Tomohiro Imai　　**望月 久稔** † Hisatoshi Mochizuki

1. はじめに

ウェブ上では多種多様な人によって，非常に高い頻度で様々な文書が更新される．これらの文書は自動で解析することで市場調査や動向調査などへの利用が期待できる．解析法の多くは頻度情報を用いるが，文書の特徴を表す重要語のうち，頻度が小さい語を抽出することが難しい．

そこで，本論は語の関係性を解析することによって，重要語を抽出する方法を提案する．そして，重要語抽出の精度において，頻度情報を用いた解析との比較により評価する．提案手法は，語の共起関係からグラフを構築して，$PageRank$ [1] を用いて解析することで特徴ベクトルを抽出する．続いて，抽出した特徴ベクトルから閾値を用いて，重要語を抽出する．

2. 語の関係性を用いた重要語抽出

はじめに，文書から特徴ベクトルを抽出する方法を説明し，次に重要語の抽出方法について述べる．

2.1. 特徴ベクトルの抽出と idf の付加

はじめに解析対象とする文書を形態素解析して，文書の特徴を表す上で大きな役割を果たすと考えられる名詞を抽出する．続いて，語の共起から無向グラフを構築する [4]．ここで，文書グラフ G を式 (1) に示す．無向グラフを表す行列 H に対して，原始性を付加して，さらに α によって H への依存率を決定する．n は対象文書に出現した形態素の種類数である．

$$G = \alpha H + \frac{1-\alpha}{n} \tag{1}$$

続いて構築したグラフを PageRank [1] で解析する．文書グラフ G からべき乗法によって定常ベクトルを抽出する．計算式を式 (2)，式 (3) に示す．i は収束までの繰り返し回数，π は G に対する左固有ベクトル，式 (3) の左辺 $\mathbf{1}$ はベクトル，右辺 1 はスカラーを表す．π が収束したとき，これが G の定常ベクトルとなる．

$$\pi^{(i+1)T} = \pi^{(i)T} G \tag{2}$$

$$\pi^{(i+1)T}\mathbf{1} = 1 \tag{3}$$

抽出した定常ベクトルは文書の特徴を表す特徴ベクトルであり，ベクトルの値は文書中に出現した形態素の，その文書における重要度を表す．

頻度情報を用いる方法として，tf-idf がある．tf は対象文書から，idf は複数の文書からの情報を利用する．また，提案手法は tf と同様に文書単体から重要度を求めるため，idf を付加することで他の文書の情報を利用できると考えられる．文書数を D，形態素 w が 1 回以上出現した文書数を $DF(w)$ としたとき，idf を式 (4) に表す．idf を提案手法で求めたベクトル値に掛け合わせて，正規化することで重要度を更新する．

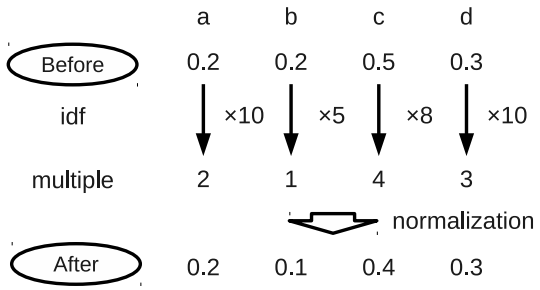

図 1: idf 付加の例

$$idf = \log \frac{D}{DF(w)} \tag{4}$$

図 1 に例を示す．語 $\{a, b, c, d\}$ からそれぞれ重要度を求めると，大きさは $c > d > a = b$ である．それぞれ，予め求めた idf の値を掛け合わせて正規化すると，大きさは $c > d > a > b$ となる．これにより，対象文書以外の頻度情報を付加する．

2.2. 重要語の抽出

抽出した固有ベクトルから，対象文書の重要語を抽出する．固有ベクトルの各値は対象文書に出現した形態素の重要度の合計値を 1 としたときに，各形態素に割り当てた重要度を示す．したがって，値が大きい形態素ほど対象文書において重要である．そこで，重要度の平均値よりも重要度が大きい語が重要語であると考える．対象文書において，出現した形態素の数を m としたとき，重要度の平均値 $1/m$ を閾値として，閾値よりも重要度が大きい形態素を重要語とする．図 1 では，$m = 4$ から閾値は 0.25 であり，c, d が重要語となる．

3. 評価

実験データとして，CD-毎日新聞データ集 95 版 [2] に収録されている新聞記事約 10 万件を使用する．このデータは人間が抽出した重要語が各記事に登録されており，これを重要語の解とする．また，中には「国際」と「国際会議」のように包含関係にある語が存在する．実験では，包含関係にある語同士では，長い語ほど優先して，その文書内で重要であると考える．実験データから無作為に抽出した 3,000 件の文書を文書集合として使用して，解析した文書ごとに適合率，再現率を

†大阪教育大学, Osaka Kyoiku University

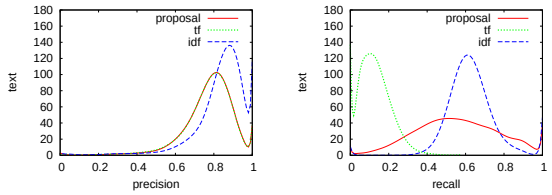

図 2: tf と idf との比較（左:適合率，右:再現率）

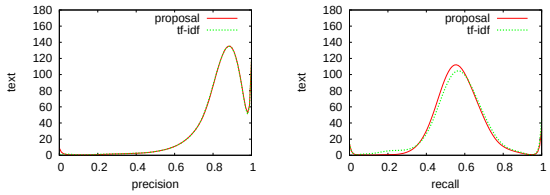

図 3: tf-idf との比較（左:適合率，右:再現率）

求めて，その分布から評価する．実験は Intel Celeron G550 2.60GHz，Memory 4GB，CentOS6.4 上で行う．形態素解析器は MeCab [3] を使用する．

評価指標として，重要語の抽出における再現率と適合率を順に定義する．各記事に登録された重要語を解集合として，解析により抽出した重要語群と解集合を比較する．抽出した重要語群のうち解集合に含まれた語の数を t，解集合の語の数を a としたとき，以下に再現率の式を示す．

$$\text{再現率} = \frac{t}{a} \tag{5}$$

続いて，適合率を定義する．重要語を抽出する上で，1 文書あたりの重要度の合計値は 1 である．本論では，重要度が解集合に，より多く割り当てることが適切であると考えて，解集合に割り当てた重要度の合計値と，1 文書あたりの重要度の合計値との割合を適合率とする．重要語の解 i の重要度を S_i としたとき，適合率の式は以下のとおりである．

$$\text{適合率} = \frac{\sum S_i}{1} = \sum Si \tag{6}$$

提案手法において，idf を付加しないものを提案手法 1，付加するものを提案手法 2 とする．頻度情報を用いた tf による解析を対象手法 A，idf による解析を対象手法 B として，提案手法 1 と比較する．また，tf-idf による解析を対象手法 C として，提案手法 2 と比較する．

まず，提案手法 1 と対象手法 A と対象手法 B との比較を図 2 に示す．グラフは横軸に適合率，再現率を，縦軸は文書数を示した分布である．分布が右に偏るほど各精度は高い．適合率について，対象手法 B が最も右寄りで，提案手法 1 と対象手法 A はほぼ等しいことがわかる．平均値は，提案手法 1 が 77.6 %，対象手法 A が 77.4 %，対象手法 B が 83.6 %となり，対象手法 B が分布と平均値で最も高い値を示した．一方で，再現率について，対象手法 A の左に大きく傾いた分布に比べて，提案手法 1 と対象手法 B は中央より右寄りな分布である．平均値は，提案手法 1 が 55.3 %，対象手法 A が 12.7 %，対象手法 B が 62.6 %となり，対象手法 B が最も高い値を示した．

対象手法 B が高い精度を示した理由は，実験データに適応したからであると考えられる．まず，idf は複数の文書から得られるその語の稀少度であり，実験ではより長い熟語に集約するため，語の重複が抑えられて，稀少度が高まる．そして，熟語は人間が重要語として抽出しやすいと考えられ，idf による抽出が有効だと考えられる．

一方で，提案手法 1 と同様に文書単体から解析する対象手法 A と比べて，提案手法 1 は適合率において同等となった．しかし，提案手法 1 は，語の関係性を解析したことにより，対象手法 A に比べて頻度が小さい重要語をより多く抽出できたため，再現率を大きく上回った．また，対象手法 B と比べて再現率 0.8 以上で多く分布しており，再現率の高い文書数が最も大きい．

続いて，提案手法 2 と対象手法 C との比較を図 3 に示す．グラフの見方は図 2 同様である．適合率，再現率はほぼ等しく，平均値もほとんど差が見られなかった．平均値は適合率が，提案手法 2 が 83.5 %，対象手法 C が 83.4 %であり，再現率がともに 56.9 %となった．

提案手法は，idf を付加することで，精度が改善された．しかし，idf のみの対象手法 B に比べると精度は低い．これは，提案手法が低い重要度を割り当てた重要語に対して，idf を付加しても大きな値の改善が難しいことが原因である．このような傾向は，グラフ上で文間の接続がない文が存在する文書に多く見られた．

4. おわりに

提案手法は語の関係性に着目して，文書をグラフとして捉えることによって重要語を抽出した．また，頻度情報によって抽出する tf に比べて，提案手法は tf と同様に文書単体から重要語を抽出できる上で，より高い再現率を示した．さらに，tf-idf と同様に idf を付加することによって，精度が向上することを確認した．しかし，文書内に同じ語が出現しない文書では，グラフ上で文間の接続ができないため，解析できない．今後の課題として，文間の関係に着目した解析があげられる．

参考文献

[1] Sergey Brin,Lawrence Page,The anatomy of a large-scale hypertextual Web search engine,Computer Networks and ISDN Systems,30,1-7, pp.107-117, 1998.

[2] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html，CD-毎日新聞データ集 95 版.

[3] MeCab，http://mecab.sourceforge.net/，2013.

[4] 今井智宏，望月久稔，語の関係性を抽出した特徴ベクトルによる文書分類の提案，情報処理学会第 76 回全国大会，第 2 分冊，pp.115-116，2014.