D-013

語の共起による文書グラフの構築と PageRank を導入した重要語抽出法

Construction of Document Graph with Term Co-ocurrences and Keyword Extraction Method with PageRank

今井 智宏 † Tomohiro Imai
望月 久稔 † Hisatoshi Mochizuki

1. はじめに

文書解析は，分類や要約，意味の抽出による意見分析など多種多様に利用できる [3]．これまでの研究では文書の属性を分析する手段として，語の頻度情報から文書の性質を表す語である重要語を抽出する方法があり，人手による解析よりも高速であるが，精度は人手より低く，特に語のつながりを解析する精度は低い．

提案手法は語の共起によるグラフを構築し，グラフ解析に PageRank を導入して，重要語を抽出する．

2.PageRank と共起を導入した重要語抽出

2.1.PageRank

PageRank は WWW におけるウェブページの検索ランキングを生成するアルゴリズムであり [2] [4]，ウェブページのリンク構造をマルコフ理論に帰着して，べき乗法により解析する．グラフ中の節点は，被リンクが多いほど重要であり，さらに重要な節点からリンクされるほど重要である．

左固有ベクトル $\pi^T(1,n)$ と以下の式 (1) に示す推移確率行列 $G(n,n)$ にべき乗法を用いて定常ベクトル $\pi^{(k)T}(1,n)$ を得る．n は節点数，H はハイパーリンク行列，A はぶら下がり節点からのテレポーテーション行列，α は H への依存度，e^T は全ての節点からのテレポーテーション行列を示す．パラメータ α と収束域 ϵ の大きさによって，精度と計算量が変化する．

$$G(n,n) = \alpha H + (\alpha A + 1 - \alpha)e^T/n \qquad (1)$$

最後に PageRank のアルゴリズムを Algorithm 1 に示す．ここで，δ はベクトル差の 1 ノルムを表す．

2.2. 重要語抽出の方法

まず，入力した文書を形態素解析 [5] して語の単位に切り分け，文中で重要語の候補として名詞，動詞，形容詞を抽出する．次に，語を節点，語の共起を辺と見なして，無向グラフを構築する．ここで，同じ文に出現した語同士は共起関係にあるとする．また，無向グラフであるため，ぶら下がり節点は存在しない．よって式 (1) の A は必要ない．最後に，構築したグラフを PageRank で解析し，定常ベクトルを得る．定常ベクトルの成分の値が大きい語ほど文書の重要語であるとする．

Algorithm 1 PageRank

Ensure: $\pi^{(k+1)T} = \pi^{(k)T}G$
  $k \Leftarrow 1$
  $\pi^{(1)T} \Leftarrow \frac{1}{n}e$
  $\delta \Leftarrow 1$
  while $\delta > \epsilon$ **do**
    $\pi^{(k+1)T} \Leftarrow \pi^{(k)T}G$
    $\pi^{(k+1)T} \Leftarrow \frac{\pi^{(k+1)T}}{|\pi^{(k+1)T}|}$
    $\delta \Leftarrow ||\pi^{(k+1)T} - \pi^{(k)T}||$
    $k \Leftarrow k + 1$
  end while

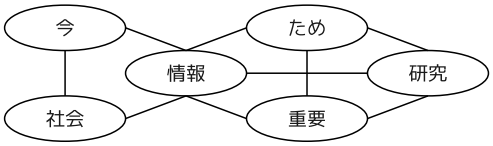

図 1: 共起から構築したグラフ構造

例: 『今は情報社会である．そのため情報の研究は重要だ.』を，まず形態素解析し，1 文目から {今，情報，社会} を，2 文目から {ため，情報，研究，重要} を抽出する．

次に語の共起からグラフを構築する．ここで，両方の文に出現した「情報」はグラフを構築する際に，異なる節点とせず 1 つの節点として扱う．以上により構築したグラフを図 1 に示し，行列を式 (2) に示す．A〜F はそれぞれ { 今，情報，社会，ため，研究，重要 } に対応する．また，行列の要素 (a,b) が 1 の場合，a から b へのリンクがあることを表し，0 はリンクなしを表す．例えば,「B」（情報）はそれ以外のすべての語とのリンクをもつため,「B」の行，列ともに「B」以外の要素すべてが 1 の値をもつ．

$$X = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \qquad (2)$$

最後に構築したハイパーリンク行列を $\alpha = 80, \epsilon = 0.00000001$ の条件下で解析すると式 (3) の定常ベクトルを得る．定常ベクトルの成分は各語の対象文書における重要度を示すため，もっとも大きな値を示す「B」が例文中でもっとも重要な語であることが分かる．

$$\pi^T = \{0.125, 0.263, 0.125, 0.161, 0.161, 0.161\} \qquad (3)$$

† 大阪教育大学, Osaka Kyoiku University

3. 文書分類による精度評価

3.1. 実験方法

重要語を用いた文書分類によって重要語抽出の精度を評価する．提案手法によって得た定常ベクトルは文書構造を表す指標であるため，その比較によって文書を分類する．

分類方法として，はじめに 2.2 を用いて得た各文書の定常ベクトルをその文書における文書スコアとし，データベースへ登録する．ここで，登録する文書は分類済みであるため，そのカテゴリ情報も登録する．次に分類する対象としての別文書の定常ベクトルとデータベースに登録した定常ベクトルを比較し，スコアの差を式 (4) で求める．A と B は文書を表し，それぞれの定常ベクトルを $A[i], B[i]$ で表す．最後に，差が最も小さい値を示した文書のカテゴリを対象文書のカテゴリとする．

$$D(A,B) = \sum_{i=1}^{n} |A[i] - B[i]| \tag{4}$$

実験は Intel Core i7-920 2.67GHz,Memory 12GB, Fedora15 上で行った．データベースに登録するデータは朝日新聞 [6] の 5 カテゴリ，スポーツ，社会，経済，政治，国際のデータを 700 件ずつと，ITpro [7] の 3 カテゴリ，management，network，security の 2000 件ずつを用いる．また，分類対象として用いるデータはデータベースに登録するものとは別に，前者は各カテゴリ 30 件ずつ，後者は 50 件ずつを使用する．精度は分類結果から F 値を求め評価する．最後に，比較手法として TF-iDF を用いる．TF-iDF の値は文書中の語の重要度を示すため，提案手法と同様の形式で精度評価ができる．

3.2. 評価

F 値についての結果を図 2 に示す．朝日新聞の分類では 0.023〜0.084，ITpro の分類では 0.040〜0.107 の値だけ提案手法が上回った．したがって，両データの分類は比較手法に比べ提案手法の方が精度は高い．

比較手法は数え上げによって頻度情報を生成する．それに比べて，提案手法は文書構造をグラフとして捉え，語同士の関係性を解析することで，より精度の高い重要語抽出ができた．これは提案手法が頻度情報に加えて，語のつながりを解析しているためである．

比較手法は非常に単純なアルゴリズムであるため高速である．それに比べて，提案手法は行列を扱う上に収束するまで繰り返し計算するため時間計算量は大きくなる．したがって図 3 に示すように処理時間は比較手法のほうが優れた結果となった．処理時間の差は朝日新聞の場合 13〜38 秒，ITPro の場合 23〜159 秒となった．パラメータ α，ϵ の変更によって，提案手法の処理時間を削減することが可能であるが，精度とトレードオフである．しかし，比較手法は大量の文書を解析する必要がある [1] ことに比べて，提案手法は対象文書単体を解析することで重要語の抽出が可能である．よって，提案手法の文書解析における時間計算量は扱う文書数に依存しない．

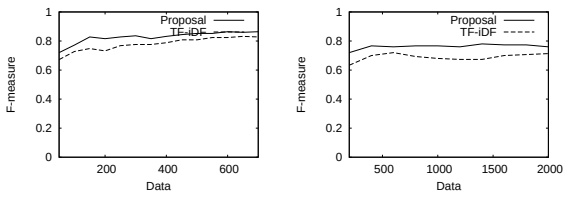

図 2: F 値（左:朝日新聞，右:ITpro）

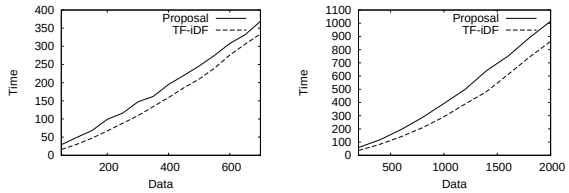

図 3: 処理時間（左:朝日新聞，右:ITpro）

4. おわりに

提案手法は語のつながりに着目し，グラフ解析アルゴリズムである PageRank によって重要語を抽出した．文書からの意味抽出に対して提案手法が適用できれば，頻度情報を用いた手法に比べ，より精度の高い結果が期待できる．しかし，時間計算量は従来の解析の方が優れていることから，精度と時間計算量はトレードオフの関係である．したがって，提案手法は静的なデータベースの構築利用に対して有益である．

参考文献

[1] 松尾豊，石塚満:語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学会論文誌，Vol.17，No.3，pp.217-223，2002.

[2] Lawrence Page:Improved Text Searching in Hypertext Systems，IN THE UNITED STATES PATENT AND TRADEMARK OFFICE 60/035,205，1997/03/28.

[3] 金明哲:テキストデータの統計科学入門，岩波書店，2009.

[4] Amy N.Langville，Carl D.Meyer:GOOGLE'S PAGERANK AND BEYOND，Princeton University Press，2006.

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

[6] 朝日新聞，http://www.asahi.com/，2012.

[7] ITpro，http://itpro.nikkeibp.co.jp/，2012.