# キーワード周辺語のリンク先内容を考慮したキーワード関連付け手法

大野 和久　西川 侑吾　伊藤 直之　松本 征二　新堀 英二
大日本印刷株式会社 情報コミュニケーション研究開発センター
{Oono-K6, Nishikawa-Y5, Itou-N12, Matsumoto-S8, Shinbori-E}@mail.dnp.co.jp

## 1　はじめに

近年，書籍のデジタル化に伴い，書籍内に登場するキーワードの意味を知るために別の書籍を参照する需要が高まっている．ただし，このキーワードは，必ずしも一意にその意味が定まるわけではなく，同表記で多義性がある場合は，表記だけでは一意に意味を定めることができない．そのため，キーワードの意味を考慮して，キーワードと別の書籍を関連付ける技術の重要性が高まっている．

キーワードの意味を考慮して，文書内容に基づいた関連付けを行うためには，キーワードおよびキーワードの周辺に出現する周辺語の頻度を用いる手法が一般的にある [1][2]．これらの研究では，関連先の文書を決定し，リンクを生成するために，キーワードが出現する関連元の文書内において，同時に出現する語の共起および出現頻度を用いている．しかし，キーワードの周辺語を用いるだけでは，語の数が限られており，特定の周辺語に影響されるため，最適な関連付けを行うことができないといった課題がある．

また，Web ページ検索において，検索対象の Web ページの内容をより正確に表現するために，検索対象の Web ページから生成される特徴ベクトルを改良する手法についての研究がある [3]．この研究では，検索対象の Web ページからリンクしている Web ページおよび検索対象の Web ページへリンクしている Web ページに含まれる単語の頻度情報を，検索対象の Web ページに含まれる単語の頻度情報に加えることにより，検索対象の Web ページを特徴づけている．ただし，検索対象の Web ページをより正確に表現する一方で，検索語への特徴づけについては，検索語だけの頻度情報を用いている．前段落におけるキーワードの周辺語の数と同様に，検索語に含まれる語の数は限られているため，検索語に含まれる特定の語に影響されるといった課題がある．

そこで本研究では，関連元の文書において，キーワードの周辺語の頻度に加え，その周辺語の関連先文書に出現する語の頻度も用いることにより，キーワードを特徴づけるための語の種類数を増やす．このことにより，キーワードおよび周辺語に関連する語の重みを大きくする．同時に，特定の周辺語による影響を小さくすることにより，キーワードと文書との最適な関連付けを行い，関連付けの精度向上を行う．

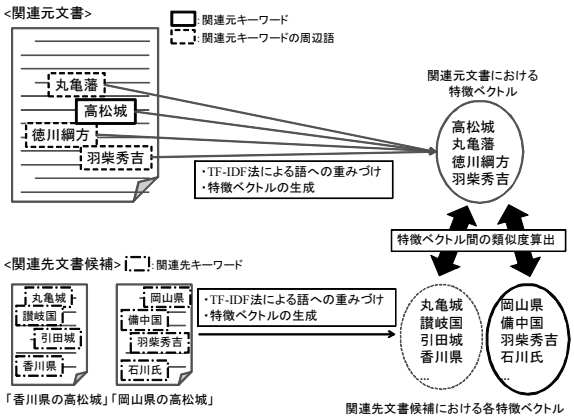


図 1　既存手法によるキーワードの関連付け

## 2　周辺語のリンク先内容を用いたキーワード関連付け

本研究では，正規表現と形態素解析処理を組み合わせたキーワード抽出処理によってキーワードを抽出した後，その抽出されたキーワードに対して文書を関連付ける．本稿の定義として，ある文書と関連付くキーワードが存在する文書を関連元文書とし，関連付け先にあたる文書を関連先文書とする．また，関連元文書に存在するキーワードを関連元キーワードとし，関連先文書に存在するキーワードを関連先キーワードとする．関連先文書を関連付ける際，関連元キーワードによっては，同表記の中で多義性を持っており，関連先文書の候補が複数存在する場合がある．

このとき，関連先文書を決定するために，一般的な既存手法では，関連元キーワードの周辺語を用いて，最適な関連先文書を算出する．既存手法によるキーワードの関連付けについて，図 1 に示す．ここで，本研究では，関連元キーワードの周辺として示す範囲を，関連元キーワードが存在する段落全体とし，周辺語として選択する対象の語を，キーワード抽出処理によって得られたキーワードとする．既存手法の具体的手法として，まず，関連元キーワードおよび関連元キーワードの各周辺語と，各関連先文書候補に出現する各関連先キーワードに対して，TF-IDF 法を用いて，関連元キーワードが存在する段落の文書内容および関連先文書候補群におけ

る，各周辺語および各関連先キーワードの重みを算出する．次に，その重みに基づき，関連元文書での特徴ベクトルおよび関連先文書候補での各特徴ベクトルを生成する．そして，その特徴ベクトルを用いて，コサイン尺度に基づき，特徴ベクトル間の類似度を算出し，最適な関連先文書を算出する．

ただし，周辺語を利用するだけでは，語の数が限られており，特定の周辺語に影響されるため，最適な関連先文書を算出できないといった課題がある．

例えば，図 1 に記述するように，関連元文書に “高松城” というキーワードが出現し，このキーワードの関連先文書候補として，「香川県の高松城」について書かれた文書と，「岡山県の高松城」について書かれた文書が存在する場合を考える．そして，関連先文書の正解を「香川県の高松城」とする．このとき，関連元文書において，“高松城” の周辺語として，“丸亀藩”，“徳川綱方”，“羽柴秀吉” の 3 種類の周辺語が存在すると，文脈上では，“丸亀藩” や “徳川綱方” といった周辺語から，「香川県の高松城」を算出するべきと考えられるが，既存手法により類似度を算出すると，“羽柴秀吉” といった特定の語に影響され，「岡山県の高松城」を最適な関連先文書として算出することが考えられる．

そこで本研究では，関連元キーワードの各周辺語に加え，関連元キーワードの周辺語に関連付けられた文書に現れるキーワードを，特徴ベクトルの要素として利用する．提案手法によるキーワードの関連付けについて，図 2 に示す．ここで，関連元キーワードの周辺語に関連付けられた文書を，周辺語関連先文書とする．また，関連先が一意に決まる周辺語関連先文書だけを利用し，それらの文書および該当する周辺語は，既に関連付けられているものとする．例として，前段落において述べた例と同じ例を用いる．特徴ベクトルの要素として，既存手法では，3 種類の周辺語だけであったが，これらの周辺語に加え，各周辺語関連先文書中の各キーワードも特徴ベクトルの要素へ加える．このことにより，他のキーワードを考慮し，キーワードの種類数を増やすことにより，“羽柴秀吉” による影響を小さくし，正解である「香川県の高松城」を最適な関連先文書として算出することができ，関連付けの精度を向上することができる．

### 2.1 キーワード部分の抽出

本手法の前処理として，関連元文書および関連先文書から，キーワードを抽出する．ここで，本研究では，各関連先文書の見出し語が記述された辞書があり，その辞書を利用してキーワードを抽出する場合を想定する．

見出し語を利用する場合では，一般的に，各見出し語と，各文書中の文章を照合することにより，キーワードを抽出することが考えられる．このときの照合方法として，形態素解析処理を用いる方法と，正規表現を用いる方法がある．

形態素解析処理を用いる方法では，形態素解析器によって，文章を形態素へ分割し，その形態素と各見出し語を照合する．しかし，複数の形態素によって構成されている見出し語に対しては，照合が不一致となる問題がある．

一方，正規表現を用いる方法では，複数の形態素に関係無く，見出し語そのものと一致することができる．また，最長一致により，照合することができる．しかし，表記で照合を行うため，見出し語が，関係の無い語の一部分に含まれる場合でも，一致するという問題がある．

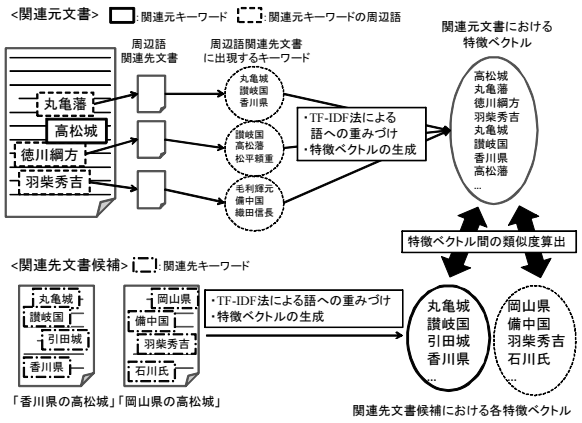


図 2　提案手法によるキーワードの関連付け

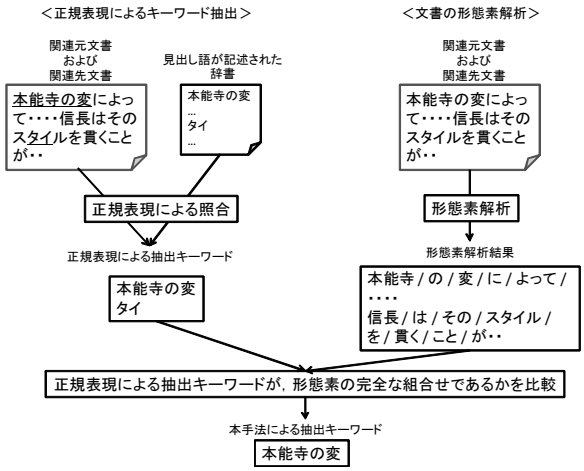


図 3　正規表現と形態素解析によるキーワード抽出

本研究では，正規表現と形態素解析による方法を組み合わせて，キーワードを抽出する．本研究におけるキーワード抽出について，図 3 に示す．まず，辞書に存在する見出し語と各文書を正規表現によって照合し，見出し語が一致する部分を抽出する．次に，各文書に対して形態素解析処理を行い，各文章を形態素単位へ分割する．そして，正規表現による抽出キーワードと，分割された形態素を比較する．このとき，正規表現による抽出キーワードが，形態素の完全な組合せであれば，キーワードとして抽出する．ここで，形態素のうち，形態素自身が意味を持たないと考えられる形態素については，キーワードの要素として除外することを考える．この抽出処理により，例えば，“タイ” は “スタイル” の一部分となるため，抽出しない．一方，“本能寺の変 “は，“本能寺”，“の”，“変” といった各形態素の完全な組合せとなるため，キーワードとして抽出することができる．このことにより，複数の形態素によって構成される見出し語を抽出

し，かつ，表記による照合の失敗を防ぐことができる．

### 2.2 周辺語のリンク先内容を用いた関連付け

本研究では，2.1 節によって抽出した関連元キーワードに対し，関連先文書を関連付ける．文書の関連付け手法として，本研究では，ベクトル空間モデルを用いる．ベクトル空間モデルでは，関連元キーワードの周辺語および関連先キーワードから，各特徴ベクトルを生成し，特徴ベクトルの類似度を算出することにより，最適な文書を関連付ける．

特徴ベクトルの重み付け手法として，TF-IDF 法を用いる．一般的な TF-IDF 法では，出現キーワードをベクトルの要素とし，各キーワードへ重みを与える．本研究では，関連元キーワードおよび周辺語の重みに加え，周辺語関連先文書に出現するキーワードも重みとして用いる．このことより，関連元文書においてキーワードの種類数を増やし，特定の語が重みへ与える影響を小さくする．

まず，関連先文書における関連先キーワードの重みを算出する式を (1) 式に示す．ここで，$r$ 個の関連先文書 $D = D_1, D_2, \quad , D_r$ に $m$ 個のキーワード $w_1, w_2, \quad , w_m$ が出現するとし，$l$ 番目 $(l = 1, 2, \quad , r)$ の文書 $D_l$ における $k$ 番目 $(k = 1, 2, \quad , m)$ のキーワード $w_k$ の出現頻度を $t_{kl}$，$w_k$ を含む文書数を $p_k$，$D_l$ における $w_k$ の重みを $d_{kl}$ とする．

$$d_{kl} = t_{kl}(\log \frac{r}{p_k} + 1) \tag{1}$$

次に，関連元文書におけるキーワードの重みを算出する式を (2) 式に示す．ここで，関連元キーワードおよび周辺語における $w_k$ の出現頻度を $u_k$，周辺語関連先文書における $w_k$ の出現頻度を $v_k$，関連元キーワードおよび周辺語の重みを $q_k$ とする．ただし，周辺語の中で，演算対象とする $w_k$ は，関連先文書が既に一意に決定されているものだけとする．

$$q_k = u_k + v_k \tag{2}$$

そして，これらの重みを要素とした特徴ベクトルを生成し，関連元文書と関連先文書との間において，各特徴ベクトル間類似度を算出する．本研究では，コサイン尺度を特徴ベクトル間類似度の定義とする．関連先文書の特徴ベクトル $\boldsymbol{d}_l = [d_{1l}\ d_{2l} \quad d_{ml}]$ と関連元文書におけるキーワードによる特徴ベクトル $\boldsymbol{q} = [q_1\ q_2 \quad q_m]$ 間のコサイン尺度を算出する式を，(3) 式に示す．

$$\cos(\boldsymbol{d}_l, \boldsymbol{q}) = \frac{\sum_{k=1}^{m} d_{kl} q_k}{\sqrt{\sum_{k=1}^{m} d_{kl}^2} \sqrt{\sum_{k=1}^{m} q_k^2}} \tag{3}$$

コサイン尺度の値が大きいほど，各特徴ベクトル間の類似度は高く，関連元キーワードと関連先文書の内容が類似しているといえる．そこで，関連先文書候補群において，コサイン尺度が最大となる関連先文書を，関連元キーワードの最適な関連先文書とする．

## 3 評価実験

### 3.1 実験方法

提案手法による文書関連付けの正解率と，既存手法による文書関連付けの正解率を比較し，提案手法の評価を行う．既存手法として，関連元キーワードおよび周辺語だけに重みを与え，その重みを要素とした特徴ベクトルを生成し，関連元文書と関連先文書間において，特徴ベクトル間類似度を算出する手法を用いる．このことにより，周辺語数が少ない場合においても，最適な関連付けを行うという点において，提案手法が既存手法に比べて有用であることを実証する．

本実験では，まず，キーワード抽出処理において，形態素解析器として，「茶筌」[*1]を用いた．また，キーワードの要素から除外する形態素を，「助詞」，「助動詞」，「接頭詞」，「接続詞」，「フィラー」，「感動詞」，「名詞-代名詞，数，非自立，動詞非自立的，接続詞的，引用文字列，特殊，接尾」，「動詞-非自立」，「記号」，「その他」とした．

関連元文書としては，新字新仮名の青空文庫[*2]150 件を用い，関連付け先文書としては日本語版 Wikipedia[*3] を用いた．キーワード選定として，まず，Wikipedia の見出しの中から，日本語で始まる見出しを選んだ．そして，その中から，ひらがなだけで構成されている見出しを除いた．その結果，対象となるキーワードとして，456116 種類のキーワードを用いた．正解判定については，人手によって，関連付け結果をもとに判定した．

### 3.2 実験結果

本実験では，人手によって正解判定を行った結果のうち，既存手法もしくは提案手法の一方の手法だけが正解であったキーワード 224 件から正解率を算出した．提案手法および既存手法による関連付けの正解率を，表 1 に示す．提案手法は，61.6% の正解率であり，既存手法による 38.4% の正解率を上回っており，関連付け精度が向上しているといえる．

また，224 件のキーワードについて，周辺語の数による正解数の違いを表 2 に示し，正解率の違いを図 4 に示す．表 2 から，周辺語の数が 30 個以下にあたるキーワードが 91% を占めており，それらのキーワードにおいて，提案手法は 60.7% の正解率であり，既存手法は 39.3% であった．このことより，周辺語の数が少ない場合において，提案手法の関連付け精度が既存手法の関連付け精度を上回っており，提

表 1　提案手法および既存手法による正解率

| | 提案手法 | 既存手法 |
|---|---|---|
| キーワード数 | 224 | |
| 正解数 | 138 | 86 |
| 正解率 | 61.6% | 38.4% |

[*1] http://chasen-legacy.sourceforge.jp/
[*2] http://www.aozora.gr.jp/
[*3] http://ja.wikipedia.org/

案手法が既存手法より有用であるといえる．

提案手法により最適な関連付けを行った例として，関連元キーワードが “渤海” であり，関連先文書が，「国家の渤海」であるか，「海域の渤海」であるかを決定する場合を述べる．ここで，関連元キーワードは，『蕃書を読むことが出来なければ、 (中略) “渤海” の奴らに笑われるだろう。彼奴ら兵を起こすかもしれない。国境を犯すに相違ない。誰か読め誰か読め！』[*4]といった文章の中に出現した．また，抽出された周辺語は，“国境”，“相違”，“読む” であった．この場合，関連先文書としては，「国家の渤海」が最適である．しかし，既存手法では，関連先文書として「海域の渤海」を提示した．その理由として，三つの周辺語と関連元キーワードだけを特徴ベクトルの要素として用いており，“国境” というキーワードが出現するにもかかわらず，関連元キーワード自身である “渤海” に影響されたためと考える．一方，提案手法では，関連先文書として，正解文書である「国家の渤海」を提示した．その理由として，“国境” の Wikipedia 記事には，「国家の渤海」にも含まれるキーワードである，“領土”，“領域”，“少数民族” といったキーワードが含まれており，関連元キーワード自身の影響を小さくし，「国家の渤海」に関連するキーワードを増やしたためと考える．

ただし，周辺語の数が比較的多い場合では，既存手法が正解となり，提案手法が不正解となる場合があった．その例として，関連元キーワードが “石見” であり，関連先文書が，「地域名の石見」であるか，もしくは，「戦艦の石見」であるかを決定する場合について述べる．ここで，関連元キーワードは，『出雲の名物は人狐で (中略) “石見” にては土瓶とも (後略)』[*5]といった文章の中に出現した．この場合，最適な関連先文書としては，「地域名の石見」であり，既存手法が「地域名の石見」を関連先文書として提示した一方で，提案手法は「戦艦の石見」を関連先文書として提示した．その原因として，抽出された周辺語が 46 語であり，既存手法では，それらの周辺語による特徴ベクトルの要素だけで，最適な関連先文書を提示するために必要なキーワードの頻度情報を確保できたにもかかわらず，提案手法では，周辺語関連先文書におけるキーワードが過剰に増えることとなり，類似度計算にとってノイズとなる語が増えたことが考えられる．

表 2 周辺語の数による正解数の違い

| 周辺語数 | 提案手法の正解数 | 既存手法の正解数 | キーワード件数 |
|---|---|---|---|
| 1〜10 | 58 | 47 | 105(46.9%) |
| 11〜20 | 52 | 26 | 78(34.8%) |
| 21〜30 | 15 | 8 | 23(10.3%) |
| 31〜40 | 5 | 2 | 7(3.1%) |
| 41〜50 | 7 | 2 | 9(4.0%) |
| 51 以上 | 1 | 1 | 2(0.9%) |
| 計 | 138 | 86 | 224(100%) |

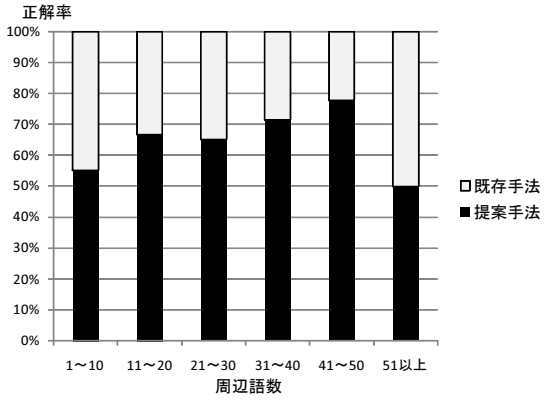


図 4 周辺語の数による正解率の違い

## 4 おわりに

本研究では，関連元キーワードと関連先文書との関連付け精度を向上するために，関連元キーワードの周辺語頻度に加えて，周辺語関連先文書に出現するキーワードも関連元キーワードの重みとして用いることにより，特定の周辺語に影響されず，最適な関連先文書を提示する手法を提案した．評価実験により，周辺語が比較的少ない場合において，提案手法の精度が既存手法の精度を上回っていることを確認した．

今後の課題として，2 点の課題を考える．1 点目として，周辺語関連先文書のキーワードを用いる際に，類似度計算にとってのノイズとならないように，利用するキーワードの制限を行う手法について検討する．

2 点目として，本研究では，周辺語の範囲として一つの段落を用いたが，文書によって段落の長さは様々であり，また，複数の段落を合わせることにより，文書内容を読み取れる場合もある．そこで，キーワードが関連する周辺語の適切な範囲について検討する必要があると考える．

## 参考文献

[1] 石田和生, 市山俊治. 複数文書間のハイパーリンク自動生成とメンテナンス. 情報処理学会研究報告. DD, [デジタル・ドキュメント], Vol. 99, No. 25, pp. 33–40, 1999.

[2] 立石健二, 細見格, 久寿居大. 課題解決型ハイパーリンク自動生成方式の開発とコンタクトセンターへの適用. 日本データベース学会論文誌, Vol. 8, No. 3, pp. 19–25, 2009.

[3] 杉山一成, 波多野賢治, 吉川正俊, 植村俊亮. ハイパリンクで結ばれた隣接ページの内容に基づく Web ページのための TF-IDF 法の改良. 電子情報通信学会論文誌. D-I, Vol. 87, No. 2, pp. 113–125, 2004.

[*4] 国枝史郎，岷山の隠士，
http://www.aozora.gr.jp/cards/000255/ les/43773_19492.html

[*5] 井上円了，迷信解，
http://www.aozora.gr.jp/cards/001021/ les/49373_39852.html