反復文字列階層グラフによる文書からのキーワード自動抽出

Keyword Extraction from Documents Based on the String Repetition Acyclic Graph

白井 智◆ 鳥井 修◆ 金井 達徳◆

Satoshi SHIRAI Osamu TORII
Tatsunori KANAI

現在，Web 上や企業内には大量の文書データが存在する．それらの文書にキーワードを付与することは，検索効率を大きく向上させる．本稿では，辞書等の事前知識を利用せず，与えられた文書から自動的にキーワードを抽出する手法を提案する．抽出には，文書中に繰り返し現われる文字列を再帰的にくくりだし文書全体を非巡回グラフに再構成した反復文字列階層グラフと，グラフ中の文字列の生起確率を利用する．

Recently, there are many digital documents in companies and on the internet. If keywords in documents are extracted in advance, it becomes easier to search documents using these keywords. We have developed a method to extract keywords automatically from documents without utilizing any dictionaries. The occurrence probability of characters plays a significant role in our method, and our method takes advantage of the “String Repetition Acyclic Graph” constructed from the original document to choose appropriate keywords.

1. はじめに

インターネット上の Web ページや電子メール，企業内のデータ群，日々発行されるニュース等々，我々が利用する電子化された文書データは増加し続けている．一方で，蓄えられた文書から希望のものを取り出す検索技術の重要性も増している．

各文書にその内容を特徴付けるキーワードが付与されていれば，それを利用することで検索効率をあげることができる．しかしながら，キーワードが付与されていない文書も大量に蓄えられており，それらの文書に人手でキーワードを付与するには膨大なコストがかかる．そのため，キーワードを自動的に抽出し文書に付与する技術が望まれている．

我々は，良質の辞書や統計データ等の事前知識を必要とせず，解析しようとしている文書そのものだけを利用してキーワード自動抽出を行なう手法を開発した．本手法では，解析対象の文書を反復文字列階層グラフと呼ぶグラフ構造に変換し，その後このグラフの構造と文字の出現順の頻度を用いてキーワード抽出を行う．反復文字列階層グラフは解析対象の文書における文字の繰り返しに基づいたグラフであり，解析対象の文書そのものから得られる．また出現頻度も解析対象の文書そのものから得られる．従って形態素解析で用いられるような事前知識を利用することなく，解析対象の文書そのものだけからキーワードの自動抽出が可能である．

2. キーワード抽出方式の概要

本手法は図 1 に示すように大きく２つの段階の処理からなる．まず，文書からキーワード候補となる文字列群を生成するキーワード候補の生成処理を行い，次にキーワード候補の中からよりキーワードらしいものを抽出するキーワードの抽出処理を行う．１段階目ではキーワード候補を得るために，文書を反復文字列階層グラフに変換する．反復文字列階層グラフは，文書の文字列の繰り返しを抽出したグラフであり，グラフの各ノードがキーワード候補になる．２段階目で各キーワード候補に対して文字列の結合度及びグラフ中の参照数を用いた評価を行い，より評価の高いものをキーワードとして抽出する．

従来の日本文の処理では形態素解析[1]が利用されることが多い．形態素解析は解析対象の文書の文字列全体を品詞に対応した文字列に区切る手法であり，特にキーワード抽出処理の中では，抽出するべきキーワードの区切り位置を確定するために利用される．

形態素解析では辞書をはじめとした事前知識を利用して文書を解析する．解析文書中に辞書に登録されていない語句が出現した場合には解析ミスが発生し，精度のよい解析結果を得ることができない．精度よい解析結果を得るためには，解析対象の文書に最適な辞書が必要不可欠である．ところが，最適な辞書を準備することは不可能な場合や，準備に現実的でないコストがかかる場合がある．例えば日々新しく生まれる新語があらかじめ全て記載された辞書というものは存在しないし，各専門分野の用語に対応した辞書を用意するにはその分野に精通した人間が辞書を作成しなければならない．

したがって，Web 上の文書のように日々新しく生成される文書や特殊な専門分野の文書を対象とした場合，形態素解析を利用したキーワード抽出を行なうには解析ミスを考慮した工夫が必要になる．[2]は形態素解析を利用したキーワード抽出手法であるが，形態素解析の解析ミスにより細分化されてしまった専門用語を，連接頻度を利用して再構成している．

一方，キーワード抽出の前処理に形態素解析を利用しない手法もある．[3]では単語の出現頻度と出現集中という統計量に注目し，あらかじめ大量の文書から生成しておいた統計データをキーワード抽出の尺度として利用することで，形態素解析を用いずにキーワード抽出を実現している．

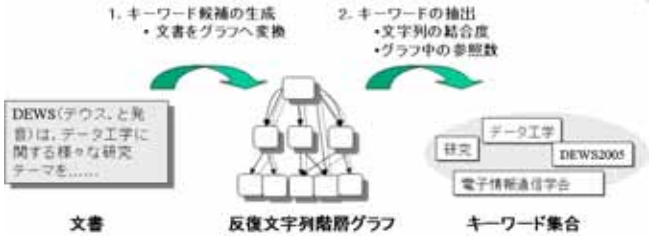

図 1 本手法の概要

Fig.1 Outline of our algorithm

◆正会員（株）東芝 研究開発センター
satoshi.shirai@toshiba.co.jp
osamu.torii@toshiba.co.jp
tatsunori.kanai@toshiba.co.jp

本手法では，[2]と異なり形態素解析を利用せず，また[3]とも異なりあらかじめ生成した統計データも利用しない．キーワード抽出には，解析対象の文書そのもののみを利用する．

本手法では「特徴的なキーワードはその文書中に 2 回以上繰返し出現する」及び「繰返し同じ順序で出現する頻度の高い文字列はキーワードを構成している」という仮定を置き，文書の文字列をグラフ化することで，柔軟に区切り位置を設定したキーワード候補を生成し，その後，それら候補からキーワードを抽出する手法をとった．この手法を用いると区切り位置を精度よく確定する必要はないため，膨大な統計データを用いず，解析しようとしている文書そのものだけを利用してキーワードを抽出することができる．

前者の仮定は特徴的なキーワードが文書中に一度しか出現しないことは稀だと考えられること，また，[3]の抽出結果を見ても，抽出されているキーワードのほとんどが文書中に２回以上出現している，という観察の結果である．

また，後者の仮定は文書を文字列と見た場合，キーワードを構成する部分の文字列の繋がりは，違う部分の文字列の繋がりより強い頻度で出現していると考えられること，また[2]では，これら結合度が単語を複合させた複合単語を抽出するのによい成果をあげている，という観察の結果である．

3. キーワード候補の生成

キーワード自動抽出の１段階目は文字列を反復文字列階層グラフに変換し，キーワード候補を得る処理である．本手法では，「特徴的なキーワードはその文書中に２回以上反復して出現する」という仮定に基づき，文書中に２回以上出現する文字列を，キーワード候補とする．

反復文字列階層グラフの構造を図 2 に示す(ノードを楕円でしめし，そのノードが対応する文字列を四角枠の並びで示した．ひとつの四角枠は他のノードへの参照を持つ文字列である)．

反復文字列階層グラフは，キーワード候補の文字列をグラフの各ノードに対応させ，ノードが対応する文字列の包含関係をノード間の参照関係として保持するように構成されたグラフであり，以下のような制約を保つように構成される．

1. 2 回以上繰返した文字列は, 必ずいずれかのノードに対応する文字列に含まれる
2. 同じ並び順の文字列の出現はひとつのノードにまとめられる
3. 各ノードの参照関係が, ノードが対応する文字列の包含関係を現わす

従って，『解析する』というノードの存在は，解析しようとしている文字列の中に『解析する』という文字列が少なくとも 2 回以上反復して出現していることを示している．『解析する』に対応するノードが参照しているノードは，『解析』と『する』でありどちらも『解析する』の部分文字列となっている．

ただ１つだけ，どのノードからも参照されていないノードが存在する．これを，ルートノードと呼ぶ．ルートノードには，他のどの文字列の部分文字列にもなりえない文字列，つまり解析対象の文書そのものが対応する．ルートノード以外の反復文字列階層グラフ中の各ノードは，2 回以上反復して出現する文字列に対応するので，先の延べた仮定に従うと，抽出すべきキーワードは反復文字列階層グラフ中のいずれかのノードに含まれる．

同じ文字列は同じノードとして表現されるので，ノードは複数のノードから参照される可能性がある．したがって，反復文字列階層グラフは木構造ではなく, グラフになる．また，参照先のノードに対応する文字列は必ず参照元のノードに対応する文字列より短くなるので，反復文字列階層グラフには巡回する部分は存在せず，非巡回グラフとなる．

反復文字列階層グラフへの変換の前処理には，データ圧縮アルゴリズムである SEQUITUR アルゴリズム[4]を用いた．SEQUITUR は, 圧縮するデータの規則性を階層的に抽出し，圧縮するデータ列のみを生成する文脈自由文法を構成するアルゴリズムである．SEQUITUR アルゴリズムでは反復文字列階層グラフに要求される(1)の制約が満たされないケースがあるので，まず文字列を SEQUITUR アルゴリズムで前処理を行い，(1)の制約を満たさない部分を探し，我々が欲する性質を満たすように再構成し，反復文字列階層グラフを生成している．

SEQUITUR アルゴリズムは，複数回出現した 2 文字以上の文字列をノードとしてくくり出す処理を再帰的に行なうことで文法を生成する．その結果，反復文字列階層グラフにも解析する文字列全体の再帰的な反復構造が反映される．本手法では，再帰的な反復構造を利用することで，評価すべきノードの数を削減するとともに，階層的な参照数を用いて，キーワードの評価を行っている．

4. キーワードの抽出

キーワード自動抽出の２段階目は得られた各キーワード候補にキーワードらしさの評価点をつけ，上位いくつかの候補，又は，ある得点以上を獲得した候補をキーワードとして抽出する．本手法では，「繰返し同じ順序で出現する頻度の高い文字列はキーワードを構成している」という仮定に基づき，グラフ上のトポロジから得られる参照数を用いた評価と，文字列の結合度を用いた評価を用いる．

4.1 グラフ上の参照数を用いた評価

キーワードはそれ自体のみで独立して意味を持ち，様々な文脈で利用される．従って，文字列の並びという観点で見ると，キーワードの前後に出現する文字列の種類が多くなる．本手法では「隣りあうノードのバリエーションの多いノードがよりキーワードらしい」という仮定をキーワードの評価に用いた．

反復文字列階層グラフでは，ノードが他のノードから参照されるのは，ノードに対応する文字列が他のノードに対応する文字列に含まれる場合である．図 2 では『解析』ノードが『解析する』および『形態素解析』ノードから参照されている．ルートノードからは『形態素解析』ノードが 3 回，『解析する』ノードが２回参照されている．『解析』という文字列の出現は全体としては 5 回あるが，反復文字列階層グラフでは，『解析』の後に『する』が出現するパターンが『解析する』に対応するノードにまとめられ，『解析』の前に『形態素』が出現するパターンが『形態素解析』に対応するノードにまとめられて，『解析』への直接の参照数は 2 になっている．つまり，反復文字列階層グラフでは，そのノードに隣接するノードのバリエーションの数がそのまま参照の数となって表現される．従って，参照数をキーワード評価の尺度に用いることで，キーワードの前後に出現する文字列の種類と同等の尺度が得られる．

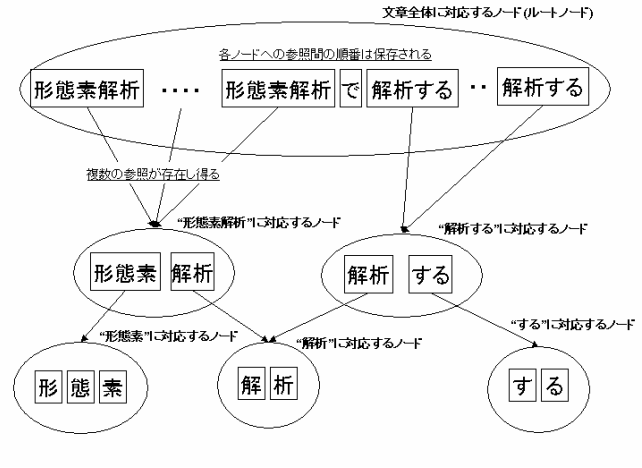

図 2 反復文字列階層グラフの構造

Fig.2 Data structure of the string repetition acyclic graph

単純に繰り返し文字列だけを抽出しキーワード候補にし，それぞれの文字列の頻度を調べるだけでは，このような文字列が使われるバリエーションの情報を得ることができない．例えば，『形態素解析』への参照が非常に多く，『解析』への参照が少ない場合には，『解析』より『形態素解析』のほうがキーワードとしてふさわしい．しかしながら，グラフ構造を持たず，キーワード候補だけで『解析』及び『形態素解析』の出現頻度を比較すると，『解析』のほうの頻度が大きくなる．出現頻度が高いものほど重要，という尺度を用いると，『形態素解析』より『解析』の評価のほうが高い評価になる．

一方，階層グラフでは，『形態素解析』中に『解析』が含まれているという情報に加え，『解析』は『解析する』と『形態素解析』の 2 つの文字列に出現し，参照数が 2 であること，また『形態素解析』は参照数が 3 であるという情報が得られるため，相対的に『解析』という文字列より『形態素解析』という文字列のほうがキーワードとして評価が高いという判断を下すことができる．

4.2 文字列の結合度を用いた評価

キーワードと見なせる文字列は，文書全体の文字列中に繰り返し出現している．したがって，キーワードを個々の文字の繋がりと捉えると，キーワードを構成する文字の繋がりは，キーワードを構成しない文字間の繋がりより相対的に強いはずである．したがって，文字間の繋がりの強さでキーワードらしさを評価することができる．

文字間の繋がりの強さを測るために，結合度という概念を利用した．結合度は，ある文字列の各々の文字が，その順番で文書に出現する頻度が高ければ高いほど，高くなる値である．結合度の高い文字列ほど単語としてまとまっていると考えられる．結合度は解析対象の文書から得られる連接頻度から算出できる．

連接頻度は解析対象の文書に含まれる全ての文字に対して，その文字の前後にはどの文字がどのような割合で出現するかを示した頻度である．例えば，『あ』という文字の出現総数が 10 であり，『あ』の直後に出現する『い』の総数が 5 であれば，『あ』の直後に『い』が出現する連接頻度は 0.5 となる．

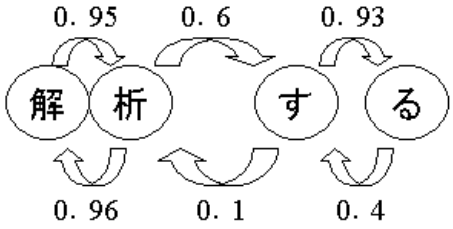

図 3 『解析する』の連接頻度

Fig. 3 Concatenation frequency

ある文字列の結合度は，その文字列を構成する各文字の連接頻度を平均したものである．図 3 のように『解析する』という文字列に関する連接頻度が与えられたとする．図中の上段の数字が直後の文字への連接頻度，下段が直前の文字への連接頻度である．結合度を算出するには，『解』の直後に『析』が出現する頻度，『析』の直前に『解』が出現する頻度，『析』の直後に『す』が出現する頻度，というように文字列間の連接頻度を順に求め，足し合わせたものの平均をとる．これが『解析する』という文字列の結合度となる．したがって『解析する』の結合度は(0.95+0.96+0.6+0.1+0.93+0.4)/6 となり，0.66 となる．

『解析』という単語を含む文書を例にとると，文書中には『解析する』，『解析した』，『解析後』というように，『解析』+ある文字という形の文字列が多く出現する．『解析』という文字列と『解析する』という文字列の結合度を比較すると，『解析』の『析』の後にくる文字の種類が多くなるため，『析』と『す』の間では，連接頻度が小さくなる[1]．したがって，『解析』の結合度 0.96((0.95+0.96)/2)より，『解析する』の結合度 0.66 のほうが小さくなり，『解析する』よりも『解析』のほうがより単語らしいといえる．

このような連接頻度を用いた評価は，[2]の一部でも用いられている評価法である．[2]では形態素解析で細分化されてしまった専門用語を再構成するために形態素解析後の文字列間の連接頻度を利用している．一方，本手法では，形態素解析を使わず，文字の並びから単語の区切りを判定するために文字間の連接頻度を利用している．

4.3 総合評価

階層グラフの各ノードに対して，以上 2 つの評価方法を利用してそれぞれに得点を算出する．結合度による評価で算出される点数は，0 から 1 の間の数値をとり，参照数で算出される点数は，2 以上の自然数となる．

これら 2 つの得点を総合的な評価点に反映させる計算式は複数考えられるが，今回は参照数の log をとったものと，結合度の積を最終的な評価点として採用した．積を採用したのは，最終的な得点に双方の値を反映させるためであり，log を採用したのは，値域の幅の大きい参照数の影響を抑えるためである．

[1] 図 3 では結合度が高くより単語らしい文字ほど近くに図示した

5. キーワード抽出実験

本手法を用いたキーワード抽出システムを実装し，本稿の予備原稿をサンプルとして抽出実験を行った．図 1 に上位 20 位以内までをリストアップしたものを示す．

本手法でのキーワード抽出結果を見ると，『評価』,『解析』,『抽出』といった単純な単語は，文字列の結合度も高く，またいろいろな繋がりの中で出現するため，参照数も高くなっており，上位に抽出される．また，頻出する単語に加えて，本稿での特徴的な複合キーワードである『反復文字列階層グラフ』が抽出できていることがわかる．本手法では，階層グラフ中の参照数というグラフのトポロジを用いた尺度を利用しており，連続した名詞からなる複合単語の組合せの中から，独立して意味を持ち様々な文脈で利用される部分だけがキーワードと高く評価されるからである．

また，『する』,『から』,『という』という結合度も高く，いろいろな使われ方をする助詞や接続詞や『を利用』のようにたまたま文中で多く使われた名詞以外の語句など，キーワードの区切りとしてはおかしいものも抽出されている．本手法の場合，品詞情報を一切利用していないので，明らかに助詞や接続詞とわかるノイズも多く抽出されている．ノイズを抽出しないような評価軸を加え，キーワード抽出の精度を改善するのは今後の課題である．

表 1 本手法での抽出結果

Table 1 Keyword extraction

順位	結合度	参照数	キーワード
1	0.912162	24	文字列
2	0.965116	19	反復文字列階層グラフ
3	1	17	評価
4	0.951807	18	抽出
5	1	13	参照
6	0.962701	14	解析
7	0.899061	16	文字
8	0.715802	32	する
9	0.914286	13	結合度
10	0.916667	11	を行なう
11	0.761905	16	という
12	0.954185	9	グラフ
13	0.641901	23	から
14	0.655002	19	キーワード
15	0.70811	13	形態素解析
16	1	6	候補
17	1	6	検索
18	1	6	情報
19	0.958333	6	処理
20	0.880952	7	利用

6. まとめ

本稿では，与えられた文書から自動的にキーワードを抽出する手法を説明した．本手法では，文書そのものから求められる文字間の結合度及び繰り返しによる階層構造を利用してキーワードの抽出を行なう．形態素解析等の前処理は必要ではなく，形態素解析時に必要とされる辞書および統計データを必要としない．

したがって，専門分野に応じた辞書の修正，言葉の移りかわりによる辞書のアップデート等のコストをかけずに単一の文書からキーワード抽出を行なうことができる．

ただ，本手法ではキーワードらしくないものも抽出されており，完全に適正なキーワードのみに高い評価を与えることができるとはいいがたい．改善の方針としては，事前知識を利用しないという前提を維持したまま，キーワードを抽出する評価軸に工夫を加えることによって，キーワード抽出の精度を上げることが考えられる．将来的には，形態素解析の結果を利用して，不要な語を候補から外す処理を加えるといったように，部分的に事前知識を利用することも考えられる．

[文献]

[1] 永田昌明: “ 形態素解析 ”, 言語と心理の統計, pp.62-73, 岩波書店, 2003.

[2] 中川裕志, 森辰則, 湯本紘彰: “ 出現頻度と連接頻度に基づく専門用語抽出 ”, 自然言語処理, Vol.10, No.1, pp.27-45, 2003.

[3] 梅村恭司: “ 未踏テキスト情報中のキーワードの抽出システム開発 ”, 未踏ソフトウェア創造事業, 2000.

[4] Nevill-Manning, C.G. and Witten, I.H.: “Identifying Hierarchical Structure in Sequences : A linear-time algorithm”, Journal of Artificial Intelligence Research, Vol.7, pp.67-82, 1997.

白井 智 Satoshi SHIRAI

株式会社東芝 研究開発センター コンピュータ・ネットワークラボラトリー．2001 京都大学大学院情報学研究科知能情報学専攻修士課程修了．メタデータ生成/検索アルゴリズムの研究・開発に従事．日本データベース学会会員．

鳥井 修 Osamu TORII

株式会社東芝 研究開発センター コンピュータ・ネットワークラボラトリー．1995 東京大学大学院工学系研究科計数工学専攻修士課程修了．データベース検索アルゴリズムの研究・開発に従事．情報処理学会会員．

金井 達徳 Tatsunori KANAI

株式会社東芝 研究開発センター コンピュータ・ネットワークラボラトリー．1989 京都大学大学院工学研究科情報工学専攻博士後期課程研究指導認定退学．データベースとデザインオートメーションの研究・開発に従事．情報処理学会，電子情報通信学会会員．