RSS フィード作成のためのニュース記事タイトルの抽出手法

藤村元彦
茨城大学工学部
システム工学科

新納浩幸
茨城大学工学部
システム工学科

佐々木稔
茨城大学工学部
情報工学科

1 はじめに

本論文では Web ニュース記事から記事タイトルを自動抽出する手法を提案する。本手法はある程度広範囲のサイトに汎用的に使える。また目的を RSS フィードの作成に特化させているので、正確な抽出に失敗したとしても、RSS フィードの目的を達することができるという特長をもつ。

RSS とは Web サイトの見出しや更新日時などのメタデータを構造化して記述するための XML ベースのフォーマットである。このフォーマットに従って記述された文書を RSS フィードという。RSS は現在、主に、blog の更新情報を発信するために利用されているが、本来はニュース記事のヘッドラインの配信が目的で作成されたものである。ニュースサイトから発信される RSS フィードを読むことで、そのニュースサイトを訪れることなしに、新着記事の有無や新着記事のタイトル等が確認できるため、RSS によって効率的な情報収集が可能になる。

RSS は利用者には恩恵が大きいが、発信者側には明確な利点はないために、現在、RSS を発信しているニュースサイトは限られている。このためニュースサイトの記事を独自に解析し、RSS を作成して提供する組織もいくつか存在する。

本論文でもニュースサイトの記事を解析し、RSS を自動作成することを行なう。この場合、対象とするニュースサイトを固定すれば、そのサイト内の記事は基本的に同じフォーマットで記述されるので、RSS フィードの作成に必要な情報（記事タイトルや記事の本文）を、記事のページ（HTML ファイル）から取り出すことは困難ではない。しかしサイトが変更された場合には、抽出規則を新たに作成する必要があり、手間がかかる。このため、ある程度汎用的に利用できる抽出規則が望まれる。

本論文で提案する手法は、概略、タイトルの候補をニュース記事から取りだし、次にニュース記事の本文と思われる部分を取りだし、その本文で使われている単語の様子からタイトルを決定するアプローチをとる。この手法は比較的頑健であり、様々なニュースサイトで適用可能である。しかも RSS の性格を考えれば、正確にタイトルを抽出する必要はなく、記事本文の内容を想像できる文章であれば RSS の目的は達せられるので、本手法が誤って選択したタイトルであっても、記事本文の様子から選ばれた単語を使っている文が選ばれるため、選択が誤ったとしても大きな損失はない。

2 記事ページからのタイトルの抽出

2.1 抽出の対象

自動的に記事のタイトルと要約の情報を RSS として記録するには、記事のタイトルと本文を HTML ファイルから抽出する必要がある。HTML には、明確にタイトルと本文を示す規則はないので、抽出の手法を考える必要がある[1]。

抽出対象とする HTML ファイルは、一般的なニュースサイトの記事にあるような、一つのページに一つの記事 (本文) と一つのタイトルがあるものとしている。これは抽出精度の問題が最も大きくかかわってくるからでもあるが、RSS の形式が一つの link に一つの title のようになっていることもある。

2.2 本文の抽出

本論文で行った実験プログラムでは、まず、本文部分とタイトル部分を文字列として抽出するため、table タグ等の、タイトルや本文の区切り部分に使われている可能性が高いいくつかの HTML タグで区切りを判断し、その区切りごとに文字を抽出して、タイトル・本文の候補となる文字列を抽出している。

[1]特にニュースサイトなどでは、title タグには記事のタイトルではなく、カテゴリ的な内容となっている場合が多い。

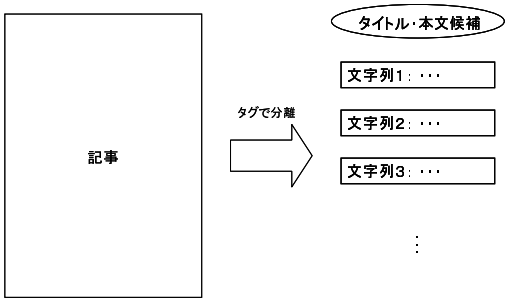

図 1: タイトル・本文の候補の抽出

その抽出された候補の中から、本文と見なされる文字列を抽出するため、その一つの条件として、句点を含むかどうかを見ている。本文は文章であるため、単語単位で使われる可能性のあるタグで分離しない限りは、候補の文字列に句点が含まれるはずであるからである。

また、候補の文字列の中に、どれだけの割合でリンクの文字が存在しているのかもみている。これは特にニュースサイトでは、本文中にリンクの文字はそれほど多くはなく、逆に本文でないものは、他ページへのリンク文字である場合がほとんどであるからである。

本文中にある写真の説明などの文章は、リンク文字はほとんど含まない場合が多い。したがって、句点やリンク文字の割合だけでは、本文かどうかの区別をすることができない。この写真と文章は、ほとんどの場合テーブルによって配置されている。よって、この文章を取り除くために、本文部分の中身にテーブルがある場合には、そのテーブルの中身を参照しないようにしていることで対処している。しかし、この方法では、本文中にあるテーブルの中に、さらに本文である文章があるような構成の場合、その文章は無視されてしまう、といったデメリットも出でくる。

また、こういった本文以外の文は、大抵文字列長は短く、逆に本文から抽出された候補は、他に比べて長い文字列となる傾向がある。そのため本文として判断するための、最低文字列長を定めている。

2.3 タイトルの抽出

2.3.1 名詞による類似度判定

本文の抽出を終えたら、そのデータを基に次はタイトル抽出行う。タイトルの抽出は、まず本文を形態素解析し名詞部分を取り出す。そして取り出した名詞が、タイトル候補の文字の長さに対して、どれだけの割合で存在してるかを計算し、その割合を基に判断するという方法で行っている。タイトルとは、本文の内容が一目で分かるように書かれた文字列であるため、本文の言葉が使われる可能性の高い文字列であるタイトルに対して、高いヒット率を得る事が出来ることが予測できる。

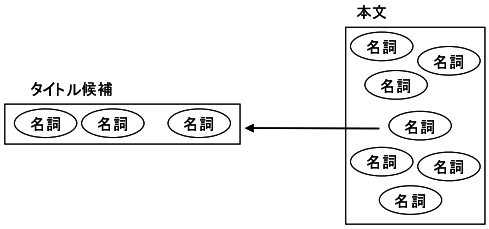

図 2: 名詞による類似度判定

しかし、この方法のみでは、タイトル候補である文字列が短い時、偶然のヒットで高いヒット率となり、タイトルとして誤認されてしまう可能性がある。さらに、タイトル候補が長いほど名詞以外の語句が増えるので、必然的にヒット率が下がってしまう傾向になってしまう。タイトルの長さとは、長すぎず短すぎずのある程度の長さであるのが一般的である。そこで、そのある程度の基準を決め、その長さから離れているもの程、タイトルの候補としての優先度を下げる、というような方法で対処している。

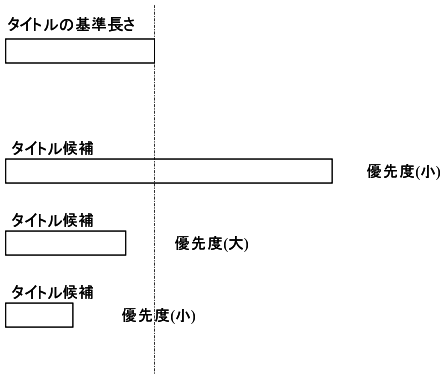

図 3: タイトル長による優先度変化

2.3.2 記事の構成による判断情報

その他にも実験で利用したプログラムでは、精度を上げるために HTML のタグや構成状態から得られる

情報を基に、優先度が変化するようになっている。

まず、大きな見出し文字として扱われた文字列はタイトルである可能性がかなり高い。よって見出しタグが使われた候補を優先させるようにすることで精度を上げることができる。本論文で行った実験では、h1、h2、h3 のタグが使用されていた候補に対して、大きく優先させるようにしている。

title タグにある候補に対しても、ある程度優先させている。高いヒット率を得たタイトル候補が得られなかった場合でも、title タグにある候補がタイトルに来やすくなり、まったくおかしな候補がタイトルとなる可能性が低くなるわけである。

2.4 問題点と特徴

色々と条件をつけていく事で精度を上げているが、基本的には同じ名詞があるかないかで判断しているわけであり、本文ではあまり利用していない言葉・言いまわしでタイトルが設定されている時には、間違ったタイトル候補がタイトルと判断されてしまう事が発生してしまう場合がある。ただ、間違ったタイトル候補が選ばれたとしても、本文の内容に近く、タイトルとしてもおかしくないような言葉が選ばれやすいという特徴がある。

3 RSS フィードの作成

RSS は、HTML と同じようにタグを使うことによって情報を記録している。必須な要素としては、まず大きくサイトの情報を示す channel 要素と、サイトにある記事 (リソース) の情報を示す item 要素がある。rdf:about 属性で channel 要素はチャンネルのリンク（通常この RSS 自身の URI)、item 要素には記事のリンク (URI) が記述される。

さらに channel 要素の中には、title、link、description、items という要素があり、title 要素にはサイトのタイトル情報を、link にはサイトのリンクを、description にはサイトの概要を記述する。items 要素は、item 要素に記録されている記事情報の目次に相当し、items 要素の中に、rdf:li 要素の rdf:resource 属性で記事のリンクを示し、item 要素と対応させる。

item 要素の中には、title、link、そしてオプションとして description 要素があり、それぞれ記事のタイトル、記事のリンク、記事の要約を記述する。なお、description 要素については古いバージョンとの互換性から、出来れば 500 バイト以内であることが好ましい。

RSS の基本的な構成要素による全体の構成を図 4 に示す。[1][2]

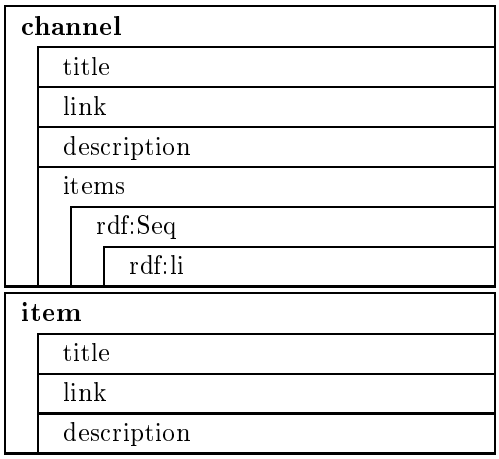

図 4: RSS の基本的な構成要素による全体の構成

4 実験

本論文で提案した抽出方法を用いたプログラムで、タイトルについての抽出精度の実験を行った。実験に使用したプログラムは、URL を指定する事で、その URL のページからタイトルと要約を抽出し、RSS で出力するものである。今回の実験に使用したサイトは、以下の表に示す 12 のサイトである。タイトルがきちんと出力されているものは○、タイトルのようなものが RSS に出力されていれば△、それ以外の出力結果の内容がまったく出来ていないものは×として、各サイトから任意に 6 記事ずつ、抽出精度についての実験を行った。

表 1: タイトルの抽出精度

	○	△	×
毎日新聞	6	0	0
読売新聞	4	2	0
産経新聞	5	1	0
Gendai.net	6	0	0
四国新聞社	6	0	0
東京新聞	5	1	0
河北新報	6	0	0
SANSPO.COM	5	1	0
CNN.co.jp	0	6	0
ロイター	6	0	0
PC Watch	6	0	0
Japan.internet.com	3	0	3

5 考察

いくつか記事ではサブタイトルや小見出しなどが、タイトルとして誤認される事があった。タイトルよりもサブタイトルのほうが、本文の言葉を用いていたり、文頭により近い位置にあったりする事があるためと考えられる。

CNN.CO.JP のサイトの記事からは、○に該当する結果を一つも得ることは出来なかった。どの記事も、タイトルの文字が要約の要素に出力されるという結果になってしまったが、これはタイトル部分と本文部分の区切り方として、BR タグや P タグで行われていたため、区切り部分が判断することが出来ず、タイトルと本文が一つの候補として扱われてしまったためである。この問題は当初から可能性があることを予測してはいたが、BR タグや P タグを区切りとして判断させてしまうと、本文候補の文字列が短くなり、特に BR タグは文の途中でも使われる可能性があり、句点がない本文の文字列であっても句点の判定で引っかかる可能性も出てきてしまう。したがって、部分部分で本文の文章が抜け落ちる可能性があるため、今まで区切りとして判断させてはいなかった。

japan.internet.com のサイトでは、本文と関連記事のリンクとの区切れが判断されずまとめて本文と判断されてしまった。関連記事はほとんどリンク文字であるため、本文の長さが短い場合には、リンク文字の割合が大きくなり本文として認識することが出来ず、処理が中断される結果となった。この場合も本文と関連記事の区切れとして使われたタグは BR タグであった。

読売新聞サイトでは、社説にある記事のタイトルが、日付の後に読売社説という言葉が付いているだけの、カテゴリ的な内容になっている。そのため、高いヒット率を得る事が出来ず、title タグにある候補がタイトルとして選ばれる結果となった。

また、タイトルはきちんと出力されていたが、写真の説明が本文の前に入ってしまい、要約部分が写真の説明になってしまった記事が一つあった。これは写真の説明が HTML ソースの位置的に、本文の前にあったため、写真説明のあるテーブルを判断できなかったためであると考えられる。

今回の実験から、タイトルや本文、関連リンクの境目にあるタグが BR タグや P タグのみ、という記事形式のニュースサイトはいくつかあり、これらの記事については大きく抽出精度が下がる、という結果を得る事ができた。特にタイトルについては、この分離が行う事ができない限り、完全なタイトルを得ることはできない。この BR タグや P タグについても区切りとして判定する必要があるだろう。ただ、本文判定が現状のまま判断を行うと、特に句点が付かない本文中の小見出し部分や、短い文などが抽出されなくなったり等、本文の細かい抜け落ちの可能性があると考えられる。

6 おわりに

本論文のタイトルを抽出する手法は、ある程度広範囲のニュースサイトに汎用的に使える。また抽出した本文から最初の 2 文を要約として、RSS フィードへの出力も行っている。しかし現状においては、まだニュースサイトの記事形式によって大きく精度を損なう場合が発生する。より汎用性を高め、多くのサイトで抽出精度の高い RSS フィードの出力を行えるための問題点の改良と、サイト単位での、RSS フィード自動更新処理の機能を追加が今後の課題である。

参考文献

[1] 神崎正英 , “RSS -- サイト情報の要約と公開”, http://www.kanzaki.com/docs/sw/rss.html

[2] 伊藤直也 , “RSS の技術的概要と最新動向”, UNIX USER 2004 年 4 月号, pp.80-90.