n-gram交代数を用いた半構造化データの不要部分削除

1M−1

山田泰寛＊　池田大輔†　廣川佐千男†
＊九州大学システム情報科学府　†九州大学情報基盤センター

1　はじめに

WWW上には類似したフォーマットを持つページが数多く存在する．例えば，検索エンジンの出力を含む機械的に生成されるページは，共通部分の多さに加え，そのようなページの数も非常に多いと考えられる．また，同じ著者やグループが生成したページを見ると，類似したフォーマットを持っていることが多い．

本論文では，このようなページを対象として，“出現頻度の高いものは全て不要である”と仮定し，入力データからこれらを全て削除する．そして，重要とは言えないまでも，不要でないものを抽出することを考える．この考えを実現するためにn-gram交代数という新たな概念を導入し，n-gramの長さと頻度で不要か非不要かの判断をする．この手法は，背景となる知識を一切使わず，言語にも依存せず，頑健性のある手法である．

この手法を用い，WWW上の新聞記事を入力とする実験を行ない，不要と思われる部分を削除することに成功した．また，入力データにノイズを含めた場合の実験も行ない，本論文で提案する手法の頑健性も示す．

2　n-gram交代数

文字列w上の長さがnである部分文字列のことを*n-gram*と呼ぶ．$V=\{v_1,v_2,\cdots,v_l\}$の各要素を文字列wに対するあるn-gramとするとき，文字列w上で$v_1,v_2,\cdots,v_l$の現われる領域をVの*n-gram*領域と呼ぶ．$P_w(V)$はwと同じ長さを持つ文字列であり，w上のi番目の文字がn-gram領域のときは，i番目の文字は1であり，それ以外は0である文字列と定義する．

次に，n-gram交代数を定義する．直感的に言えば，n-gram交代数は，文字列w上のn-gram領域とそうでない領域とが変化する回数である．

定義．文字列wと，wに対するn-gramの集合Vが与えられたとき，wにおけるVの*n-gram*交代数とは，$P_w(V)$における0と1の境界の数である．

例えば，$w=accbaacbc$と$V=\{cb,ba\}$が与えられたとき，$w=a\underline{ccba}\underline{acbc}$（下線部が$cb,ba$にあたる文字列），$P_w(V)=001110110$であるため，$n$-gram交代数は4である．

3　不要部分削除とアルゴリズム

対象とする類似したフォーマットを持つページの共通する部分に書かれてあることは，それぞれページが主張している情報とは関係ないことが多い．よって，共通部分を削除することにより，不要部分の削除を行なうことが可能になる．このとき，不要部分を出現頻度の高いn-gramのn-gram領域と定義する．

この出現頻度の高いn-gramのn-gram領域を決めるために[1]で提案されている*cut point*を用いる．cut pointとは，文字列の集合に対し各文字列から，長さnの全てのn-gramの出現頻度を数えたときのn-gramの長さnと，出現頻度の上位のa%のペア(n,a)のことを指す．

図1: n-gram交代数が大きい時（上），小さい時（下）

不要部分と非不要部分はある程度の長さを持つ文字列であると考える．もし，あるcut pointにおけるn-gram交代数が大きければ，図1の上図のように（この図では，帯全体が文字列を表し，薄い部分がn-gram領域を表す．）さまざまな場所でn-gram領域が現れ，それぞれのn-gram領域は小さいために，ある程度の長さを持つ文字列が得られない．n-gram交代数が小さければ，図1の下図のようにn-gram領域が大きくなる．よって，ある程度の長さを持つ文字列が得られる．

不要部分と非不要部分の理想的な分離とは，出現頻度の高いn-gramと不要部分のn-gramが一致することである．不要部分を共通部分と考えているので，不要部分のn-gramは長さnに依存せず，高い出現頻度を持つと考えてよい．しかし，非不要部分のn-gramは長さnが小さいとき，ある程度高い出現頻度を持つ可能性がある．よって，n-gramの長さをnを長くすることにより，非不要部分のn-gramの出現頻度が低くすることができる．

以上より，**不要部分削除**とは，複数の文字列が与え

られたときに，n-gram 交代数が極小になるような cut point を探す問題と定義する．ここで，極小の n-gram 交代数とは，cut point の n と a を増やしたときに，n-gram 交代数が増加しない範囲内で最小となる n-gram 交代数のことである．

不要部分削除を解くアルゴリズムは，HTML の文法や自然言語に関する知識，大文字と小文字の区別，全角と半角の区別などについて背景知識は用いない．そして，入力として与えられたデータから n-gram 交代数が極小になるような cut point を [1] で提案されているアルゴリズムを用いて求める．このアルゴリズムは cut point が (2, 1) を初期状態とし，n-gram 交代数が少ない cut point へ遷移していくことで，極小の n-gram 交代数をとる cut point を求める．

4 実験

入力として “Yomiuri On-Line”(YOMUIRI), “The Washington Post” (WPOST), “Berliner Morgenpost” (BERLIN) の 3 つのページからリンクされている新聞記事を扱った．これらの HTML ファイルを不要部分削除を解くアルゴリズムの入力として与え，出力された cut point の n-gram 領域を削除する．

まず単一フォーマットの HTML ファイルを入力とした実験を 3 つ行った．具体的には実験 1 は YOMIURI(164 ファイル，2.6MByte)，実験 2 は WPOST(131 ファイル，6.0MByte)，実験 3 は BERLIN(104 ファイル，3.5MByte) を入力として与えた．

実験 1 について，不要部分削除を解くアルゴリズムによって出力された cut point は (4, 5) であった．図 2 は，入力として与えた HTML ファイルのうち，ある 1 ファイルについて削除した部分を薄く表示した出力結果である．図中の “□□□” に囲まれた数字は，省略した文字の数を表す．

```
<html> <head> <title> Yomiuri On-Line/社 会
</title> □□□ 2782 □□□ <font size="-1" class="101"
color="#000000">12:30/ 2001.6.6(水) 更新 </font></td>
<!- update=nc=end -> □□□ 2485 □□□ <font
size="+2"><b> 米ＦＤＡ、クローン動物の食肉利用を規制
へ </b></font><br> <img src="/g/d.gif" width="1"
height="5"><br> <img src="/g/db.gif" width="465"
height="1"><br><img src="/g/d.gif" width="1"
height="15"><br> <!- photo start -> <!- NO PHOTO ->
<!- photo end -> <!- honbun start -> <p> 【ワシントン5
日＝館林牧子】米食品医薬品局（ＦＤＡ）は五日、牛や豚、羊などクロー
ン技術で作られた家畜の、食肉や乳製品などへの利用を規制する方針を
固めた。ＦＤＡでは、クローン動物が人間の健康や環境に与える影響を
科学的に評価した上で、規制のための指針作りに入る。 </p> (6 月 6
日 12:30)<br> <!- honbun end -> <div align="right"><img
src="/gif6/arw.gif" width="11" height="11"> □□□ 5943
□□□ <LAYER SRC="/srcfiles/specials.htm" VISIBIL-
ITY=hidden ONLOAD="moveToAbsolute(specials.pageX,
specials.pageY); visibility=true;"></LAYER> </body>
</html>
```

図 2: 実験 1 について削除部分を薄く表示したもの

図 2 より，見出し，本文にあたる部分は語頭，語尾の数文字を除けば削除されなかったことが分かる．また，見出し，記事以外の部分はほとんど削除された．実験 2，実験 3 の実験結果についても実験 1 の実験結果と同様の結果が得られた．

次に，実験 1〜3 で用いた新聞記事それぞれに対して，本文が同じ言語で書かれている新聞記事，違う言語で書かれている新聞記事をノイズとして含め実験を行った．その結果から，どの実験においても単一フォーマットでの実験とほぼ同様の結果が得られた．これより，データにノイズが含まれたとしても結果に影響を与えないことが分かった．

5 まとめ

本論文では，不要部分削除という問題に対し，出現頻度の高い n-gram を削除するという単純なアルゴリズムを用いることにより，不要部分の大部分を削除することに成功した．このアルゴリズムは，一切の背景知識を使っておらず，言語に依存しない．実験により，ノイズが含まれたデータに対しても，頑健性があることが分かった．一方、精度をあげることなどは今後の課題として上げられる．

本論文で提案するアルゴリズムは，不要部分の削除だけでなく，テキストマイニングの一種であるレコード抽出にも役立つ．本論文では，特にレコードを定義せず n-gram 交代数が極小となる cut point の n-gram 領域を削除することで，おおまかにレコードのある場所の特定をしている．ただし，このアルゴリズムでは，各レコードに対し，フィールドに分割することは不可能であるし，レコードを抽出することも不可能であることが分かる．[2] では，このアルゴリズムを用いて行う処理をレコード抽出のための前処理と位置づけ，実際にレコード抽出を行っている．

参考文献

[1] D. Ikeda, Y. Yamada and S. Hirokawa, Eliminating Useless Parts in Semi-structured Documents using Alternation Counts. Discovery Science, 2001. (to appear)

[2] Y. Yamada, D. Ikeda, and S. Hirokawa, SCOOP: a Record Extractor without Knowledge on Input. Discovery Science, 2001. (to appear)