卒業研究テーマ発表

05T4007T
江口　晃

- ベースとなる論文

『全ての部分文字列を考慮した文章分類』

岡野原大輔　辻井潤一

はじめに

- 文章分類

  与えられた文章に対して、その文章の意味に基づいたラベルを付与するタスク

- 現在の手法

  各文章に出現する単語により、ベクトルを作成する

手法の提案

- 現手法の問題点
  各単語の順序や位置は無視されてしまう
  例：固有表現(人名や商品名など)、映画のタイトル

- 手法の提案
  部分文字列を素性として利用して、文章分類をする

提案手法

- 従来の手法だと

  部分文字列の種類数はテキスト長の二乗に比例するため、計算量が膨大になってしまう。

- 提案手法

  テキスト長に比例する個数のみ存在する極大部分文字列を利用する

極大部分文字列（1）

- **極大部分文字列の定義1**

  二つの部分文字列q1、q2が、q1=αq2βで、P(q1)-|α|=P(q2)を満たすとき、q1=P q2と定義する。

  ※α、βは空文字を含む部分文字列

例：T=*abracadabra*の場合

ab=P *abr* , *bra*=P *abra* , *a*≠P *ab*

極大部分文字列(2)

- **極大部分文字列の定義2**

  全ての部分文字列を=Pの関係により、いくつかの集合に分類する。

  この各集合に属する部分文字列の中で、一番長い文字列を極大部分文字列と呼ぶ。

  出現回数が二回以上の極大部分文字列を扱うこととする。

例：T=*abracadabra*の場合

M_T={*a* , *abra*}

拡張接尾辞配列（１）

- 接尾辞配列の定義

  入力文字列がT[1,…,s]で、末尾はT[s]=$とする。

  Si=T[i,…,s](i=1,…,s)をTの接尾辞とていぎする。

  これらを辞書式順序の小さい順に並べた配列

  SA[1,…,s]をTの接尾辞配列と呼ぶ。

拡張接尾辞配列(2)

- 接尾辞配列の例

  入力T=*abracadabra*

  S1=*abracadabra$*

  S2=bracadabra$

  S3=racadabra$

  …

  S12=$

  辞書式順に並び替える

i	SA	suffix
1	12	$
2	11	a$
3	8	abra$
4	1	abracadabra$
5	4	acadabra$
6	6	adabra$
7	9	bra$
8	2	bracadabra$
9	5	cadabra$
10	7	dabra$
11	10	ra$
12	3	racadabra$

拡張接尾辞配列（３）

- 拡張接尾辞配列の定義

  接尾辞配列に最長共通接頭辞配列(LCP)を組み合わせたもの。

  LCP[1,s]はLCP[i]=Tsa[i]とTsa[i+1]の共通な接頭辞の長さと定義する。

  LCPと組み合わせることで、作業領域量は9Nbyteから9Nbitと少なくなる。

拡張接尾辞配列（4）

- **Burrows Wheeler's変換の定義**

  Burrows Wheeler's変換(BWT)は、B[i]=T[SA[i]-1]と定義する。ただし、SA[i]=1のときB[i]=T[s]である。

i	SA	L	B	suffix
1	12	0	a	$
2	11	1	r	a$
3	8	4	d	abra$
4	1	1	$	abracadabra$
5	4	1	r	acadabra$
6	6	0	c	adabra$
7	9	3	a	bra$
8	2	0	a	bracadabra$
9	5	0	a	cadabra$
10	7	0	a	dabra$
11	10	2	b	ra$
12	3	0	b	racadabra$

接尾辞木（1）

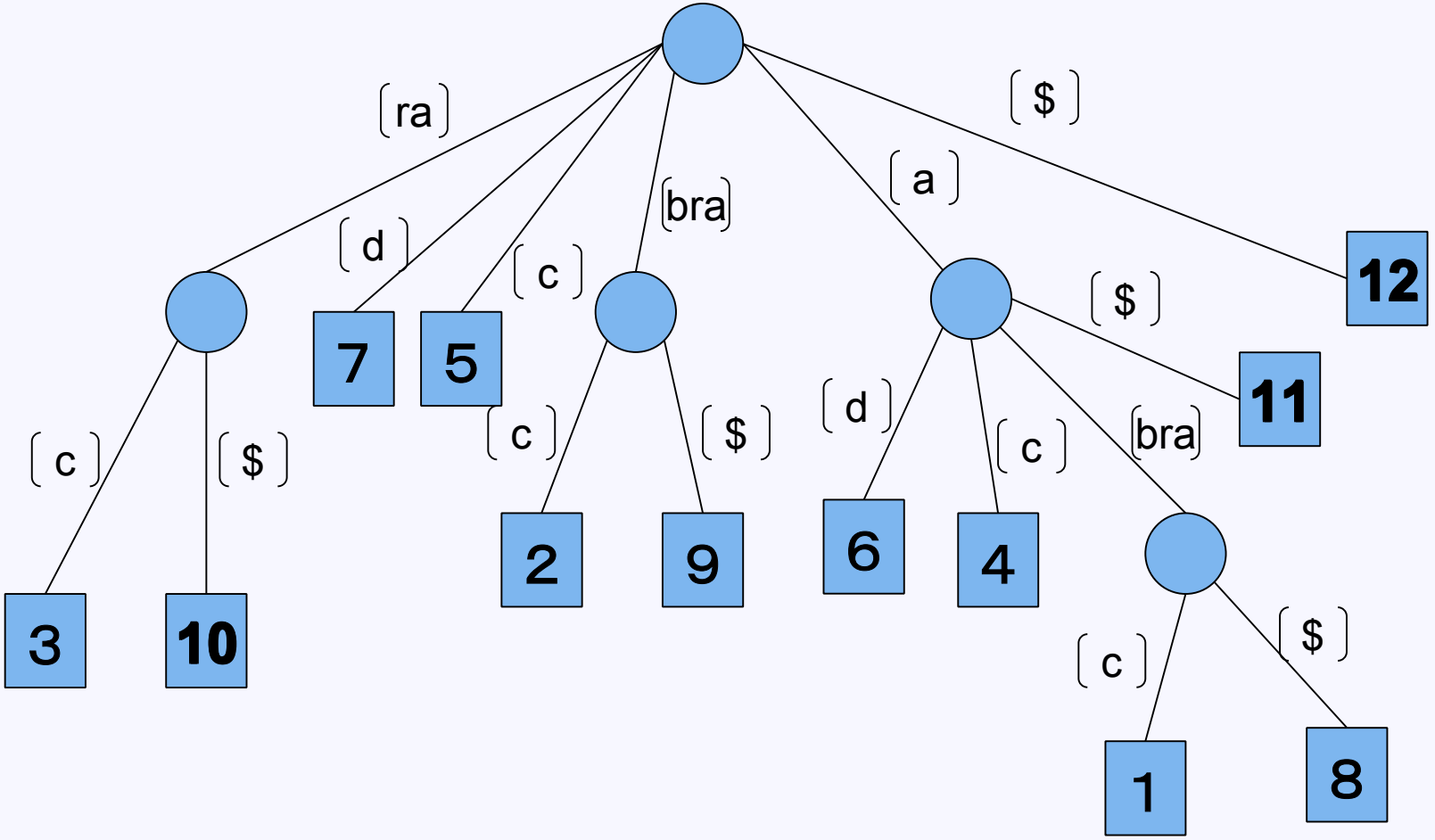
- 拡張接辞配列を用いた極大文字列の列挙

  極大文字列は以下が成り立っている

  -接辞木の内部接点に対応している

  -接辞木の接点に対応するBWT配列が二種類以上の異なる文字を持つ

接尾辞木(2)



研究テーマ

- 極大部分文字列の手法を理解し、言語間の関係などを使って、より簡単な手法に置き換えられないか検討してみる。