МАТФ
Универзитет у Београду
Математички факултет

11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Алгоритми за решавање проблема најкраће заједничке надниске

студент: Милош Миковић 1050/2020
ментор: др Александар Картељ

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Увод

- ПНЗН је један од добро познатих NP-тешких проблема оптимизације у области анализе речи
- ПНЗН се може описати као проблем проналажења најкраће речи ω сачињене од слова задате коначне азбуке Σ, тако да су све речи из унапред задатог коначног скупа L садржане у речи ω
- Примене у многим областима информатике, укључујући компресију података, оптимизацију упита, анализу и поређење текста, биоинформатику

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Нотација

- Коначна азбука Σ
- Коначна реч $\omega = \omega(1)\omega(2)...\omega(n)$, $\omega(j) \in \Sigma$
- Празна реч ε
- Дужину речи ω означавамо са $|\omega| = n$, где $|\varepsilon| = 0$
- Надовезивање слова $\alpha \in \Sigma$ на почетак/крај речи ω означаваћемо са $\alpha\omega / \omega\alpha$
- Приступ слову/секвенци слова означаваћемо са $\omega[k] / \omega[a:b]$

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Проблем најкраће заједничке надниске

- Нека важи да $\omega_1, \omega_2 \in \Sigma^*$, за реч ω_1 кажемо да је надниска речи ω_2 у ознаци $\omega_1 \succ \omega_2$ ако важи следећа рекурзивна дефиниција:

  $\omega_1 \succ \varepsilon \triangleq$ Тачно

  $\varepsilon \succ \omega_2 \triangleq$ Нетачно, ако $\omega_2 \neq \varepsilon$

  $\alpha\omega_1 \succ \alpha\omega_2 \triangleq \omega_1 \succ \omega_2$

  $\alpha\omega_1 \succ \beta\omega_2 \triangleq \omega_1 \succ \beta\omega_2$, ако $\alpha \neq \beta$
- Симбол $\triangleq$ означава једнакост по дефиницији

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Пример

- За дату азбуку $\Sigma = \{a, c, t, g\}$, важи $agcatg \succ act$
- За инстанцу ПНЗН $I = (\{a, c, t, g\}, \{act, cta, aca\})$ најмања заједничка надниска је $acta$
- Најмања заједничка надниска произвољне инстанце проблема не мора бити јединствена

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Преглед досадашњих истраживања

- ДП – за k речи дужине максимално n у $O(n^k)$ прос. и врем. сложености
- Оптимизација
- Хеуристичке функције: алфабет, већинско спајање, тежинско већинско спајање
- Алгоритми: редукуј-прошири, AEL (апроксимира очекивану дужину случајно изабране надниске)
- Метахеуристички алгоритми: генеткси алгоритам, оптимизација колонијом мрава, претрага бима

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Алгоритам гранања са одсецањем

- Побољшава технику грубе силе
- Ефикасност зависи од критеријума на основу којих се врши одсецање
- Сложеност најгорег случаја остаје експоненцијална али пажљиво одабрани критеријуми одсецања могу одсећи јако велике делове претраге (који су често експоненцијалне величине у односу на димензије улазног проблема)

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Алгоритам гранања са одсецањем

```
1: d ← |ω|                                          ▷ d - dužina trenutne reči
2: if (d > maxD) ∨ (d ≥ nd) then
3:     return
4: end if
5: if (d ⩾ minD) ∧ ZajedničkaNadniska(ω) then
6:     iDaljeNadniska ← True
7:     pozicija ← 0
8:     while iDaljeNadniska do                      ▷ optimizacija
9:         if ZajedničkaNadniska(ω[pozicija + 1 :]) then
10:            pozicija ← pozicija + 1
11:        else
12:            iDaljeNadniska = False
13:        end if
14:    end while
15:    nn ← ω[pozicija :]                          ▷ nn - najbolja nadniska
16:    nd ← |nn|                                   ▷ nd - najbolja dužina
17: end if
18: for α ∈ Σ do                                   ▷ grananje po slovima azbuke
19:    GrananjeSaOdsecanjem(ωα)
20: end for
```

$$minD = \{min_{\omega \in \mathcal{L}} |\omega|\}$$

$$maxD = |\Sigma| \cdot L, \text{ gde je } L = \{max_{\omega \in \mathcal{L}} |\omega|\}$$

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

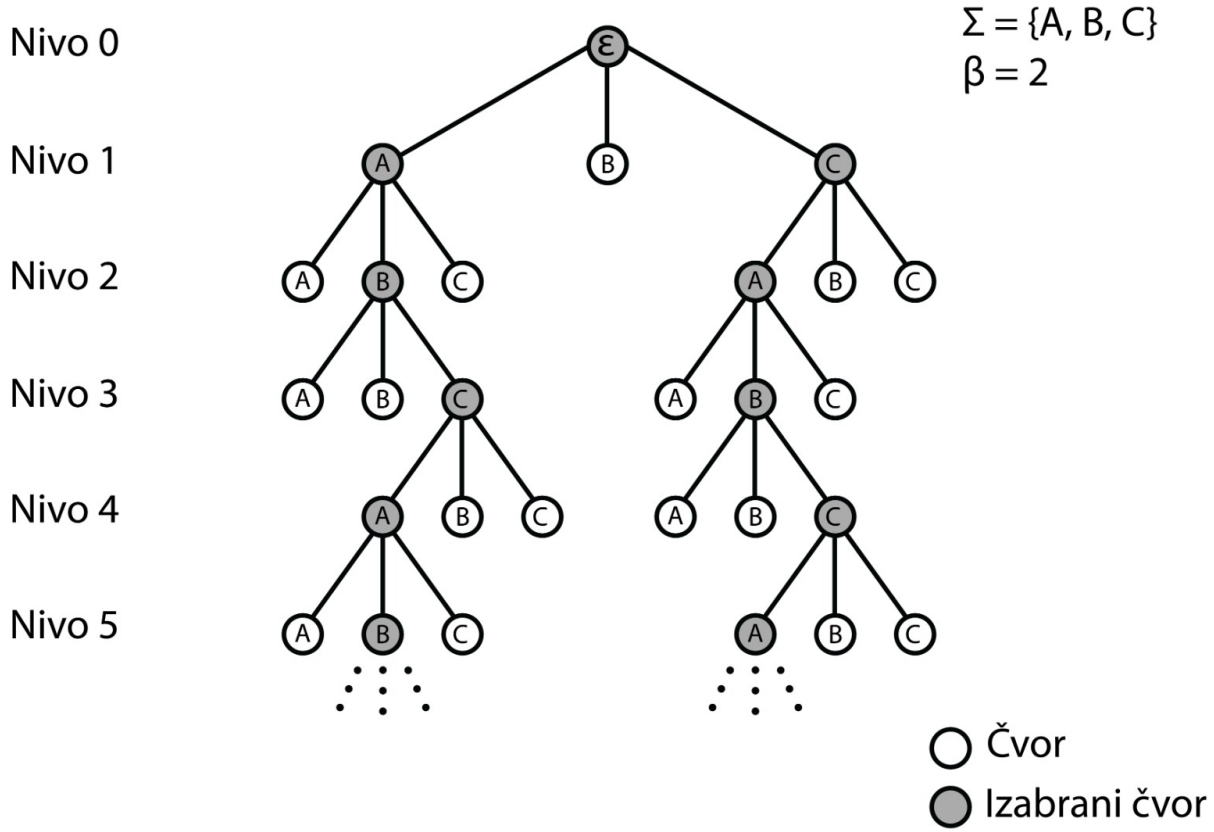
Алгоритам гранања са одсецањем

- У најгорем случају има експоненцијалну сложеност $O(\Sigma^{maxD})$
- Гарантује проналажење оптималног решења
- Практично не употребљив на већим инстанцама проблема

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Метахеуристика претрага бима

- Врста претраге графа у ширину
- Сложеност израчунавања држи у задатим границама
- Мета параметар β
- Не гарантује оптимално решење



МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Алгоритам претраге бима

- Захтева постојање хеуристичке функције ради оцене чворова на одређеном нивоу
- У пракси далеко ефикаснији од алгоритма гранања са одсецањем
- Полиномско време извршавања $O(\beta * maxD)$

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Алгоритам претраге бима

```
1: b1 ← {}
2: b2 ← {}
3: pronađenaNadniska ← False
4:
5: težine ← H([0,...,0])
6:
7: for α ∈ Σ do
8:     težinaSlova ← 0
9:     for t ∈ težine do
10:        if α == t.slovo then
11:            težinaSlova ← t.vrednost
12:        end if
13:    end for
14:    element ← {α, težinaSlova, [0,...,0]}
15:    b1.dodaj(element)
16: end for
17:
18: b1 ← RedukujBim(b1, pronađenaNadniska)
19: dubina ← 0
21: while ¬pronađenaNadniska ∧ dubina < maxD do
22:     for s ∈ b1 do
23:         težine ← H(s.pozicije)
24:         for α ∈ Σ do
25:             težinaSlova ← 0
26:             for t ∈ težine do
27:                 if α == t.slovo then
28:                     težinaSlova ← t.vrednost
29:                 end if
30:             end for
31:             element ← {(s.reč)α, s.hvrednost + težinaSlova, s.pozicije}
32:             b2.dodaj(element)
33:         end for
34:     end for
35:
36:     b1.obriši()
37:     b1 ← RedukujBim(b2, pronađenaNadniska)
38:     b2.obriši()
39:     dubina ← dubina + 1
40: end while
```

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Алгоритам претраге бима

```
1: if B.veličina() ⩽ β then
2:     for e ∈ B do
3:         AžurirajPozicije(e)
4:         if ZajedničkaNadniska(e.reč) then
5:             nn ← s.reč                    ▷ nn - najbolja nadniska
6:             nd ← |nn|                     ▷ nd - najbolja dužina
7:             pronađenaNadniska ← True
8:         end if
9:     end for
10:    return B
11: end if
12:
13: Promešaj(B)
14: Sortiraj(B)
15: Bp ← {}
16:
17: for i = 0 → β do
18:     AžurirajPozicije(B[i])
19:     if ZajedničkaNadniska(B[i].reč) then
20:         nn ← B[i].reč
21:         nd ← |nn|
22:         pronađenaNadniska ← True
23:     end if
24:     Bp.dodaj(B[i])
25: end for
26:
27: B.obriši()
28: return Bp
```

```
1: poslednjeDodatoSlovo ← element.reč[element.reč.dužina() − 1]
2:
3: for i = 0 → L.veličina() do
4:     if L[i][element.pozicije[i]] == poslednjeDodatoSlovo then
5:         element.pozicije[i] ← element.pozicije[i] + 1
6:     end if
7: end for
```

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Хеуристичке функције

1) Већинско спајање

- Инкрементално конструише надниску
- Одреди слово које се најчешће налази на почетку речи из скупа L, а затим се изабрано слово брише са почетка речи из L које га садрже
- Поступак се понавља док скуп L не постане празан
- Не препознаје глобалну структуру речи (слова са почетка краћих речи имају већу шансу да буду уклоњена)

2) Тежинско већинско спајање

- Тежина слова се одређује у односу на дужину преосталих речи у L које почињу тим словом
- Приоритизира скидање слова са почетка дужих речи
- Може да надмаши већинско спајање када нема структуираности унутар скупа L

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Тест подаци

- Тест инстанца је облика $I=(\Sigma, L)$
- Оптимално решење најчешће није јединствено а број оптималних решења расте са порастом броја речи у L
- Мањак тест података у литератури
- Проблем оптималног решења на случајно генерисаним инстанцама проблема
- Генератор тест инстанци

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Генератор тест инстанци

- Идеја је да се одабере реч gg која ће представљати горњу границу оптималности (параметри су дужина gg и Σ)
- Параметар γ представља вероватноћу уклањања слова из gg при генерисању речи у L
- Генерише m речи
- Фиксиран је $seed$ при генерисању тест инстанци
- Сви случајни бројеви узети су из униформне расподеле у одговарајућим опсезима

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Генератор тест инстанци

- Нека је нпр. дата азбука $\Sigma=\{a,b\}$
- Нека је $gg=abba$ дужине 4, нека $m=4$ број речи у скупу L и нека је вероватноћа уклањања $\gamma=0.2$
- Нека је добије скуп $L=\{abb,bba,abba,aba\}$
- Повратна вредност генератора је инстанца ПНЗН:

$$I=\{\{a,b\},\{abb,bba,abba,aba\}\}$$

- IPG $\left(ka\in\{2,4\},m\in\{8,16,32\},\gamma\in\{0.2,0.4\}\ ,|gg|_{|\Sigma|=2}\in\{20,24,28\},|gg|_{|\Sigma|=4}\in\{12,13,14\},\beta\in\{4,8,16\}\right)$
- IP $\left(ka\in\{2,4,16\},m\in\{10,20,40,80\},\gamma\in\{0.1,0.2,0.4\}\ ,|gg|\in\{50,100,500,2000\},\beta\in\{100,200,400\}\right)$

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Експериментални резултати

- $IP_{\gamma=0.1}{}^{\beta}$

β

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Експериментални резултати

$IP_{\gamma=0.1}$						
	$>gg$			$<=gg$		
BS	100	200	400	100	200	400
VS	0	0	0	48	0	0
TVS	0	0	0	48	0	0

$IP_{\gamma=0.2}$						
	$>gg$			$<=gg$		
BS	100	200	400	100	200	400
VS	0	2	1	43	1	1
TVS	1	0	2	44	0	1

$IP_{\gamma=0.4}$						
	$>gg$			$<=gg$		
BS	100	200	400	100	200	400
VS	1	4	5	24	8	6
TVS	1	4	4	22	10	7

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Дискусија резултата

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Закључак и правци даљег рада

- Look-ahead верзије представљених хеур. функција:
  1. Инкрементална изградња и рангирање слова (лошији резултати)
  2. Инкрементална изградња без рангирања (лошији резултати)
  3. Без инкременталне изградње (дуже време извршавања)
- Селекција при редукцији бима
- Технике редукције надниске
- Хеуристичке функције које процењују очекивану дужину надниске (AEL)
- Генератор тест инстанци

МАТФ
Универзитет у Београду
Математички факултет
11001 БЕОГРАД
Студентски трг 16
П.П. 550
Телефон: 011 20 27 801
Факс: 011 26 30 151
matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Хвала на пажњи

- Питања?