Approximatives Textalignment

Sven Schrinner

14. Mai 2014

Inhalt

Inhalt:

- Global Alignment
 - Spezialfall des Needleman-Wunsch-Algorithmus
- Semiglobales Alignment

Inhalt

Inhalt:

- Global Alignment
 - Spezialfall des Needleman-Wunsch-Algorithmus
- 2 Semiglobales Alignment

Verwendete Quellen:

- Skript zu "Algorithmen auf Sequenzen (2013)"
- (((Wikipedia-Artikel zu "Needleman-Wunsch-Algorithmus")))

```
Gegeben Zwei Strings s,t\in\Sigma^*
Gesucht "Optimales" Alignment zwischen s und t
```

```
Gegeben Zwei Strings s,t\in\Sigma^*
Gesucht "Optimales" Alignment zwischen s und t
Was ist ein Alignment?
```

Gegeben Zwei Strings $s,t\in\Sigma^*$ Gesucht "Optimales" Alignment zwischen s und t

Was ist ein Alignment?

Definition (nach Def 4.8, Skript A.a.S.) Ein Alignment A von $s,t\in\Sigma^*$ ist ein String über $(\Sigma\cup\{-\})^2$ (ohne (-,-)), sodass die erste Komponente ohne "—" den String s ergibt und die zweite Komponente entsprechend t.

```
Gegeben Zwei Strings s,t\in\Sigma^*
Gesucht "Optimales" Alignment zwischen s und t
```

Was ist ein Alignment?

Definition (nach Def 4.8, Skript A.a.S.) Ein Alignment A von $s,t\in\Sigma^*$ ist ein String über $(\Sigma\cup\{-\})^2$ (ohne (-,-)), sodass die erste Komponente ohne "—" den String s ergibt und die zweite Komponente entsprechend t.

```
Beispiel s = RABABA, t = BARBARA
R A B A - B A - -
- - B A R B A R A
```

Was ist ein optimales Alignment?

Was ist ein optimales Alignment?

Jede Spalte (a, b) eines Alignment besitzt Kosten d(a, b), die wir über das gesamte Alignment minimieren wollen.

Was ist ein optimales Alignment?

Jede Spalte (a, b) eines Alignment besitzt Kosten d(a, b), die wir über das gesamte Alignment minimieren wollen.

Beispiel: Edit-Distanz / Levenshtein-Distanz R A B A - B A - - - B A R B A R A (Distanz 5) $d(a,b) = \left\{ \begin{array}{ll} 0, & \text{falls } a=b \\ 1, & \text{falls } a\neq b \end{array} \right.$

Was ist ein optimales Alignment?

Jede Spalte (a, b) eines Alignment besitzt Kosten d(a, b), die wir über das gesamte Alignment minimieren wollen.

Beispiel: Edit-Distanz / Levenshtein-Distanz

$$d(a,b) = \begin{cases} 0, & \text{falls } a = b \\ 1, & \text{falls } a \neq b \end{cases}$$

Optimales Alignment:

Globales Alignment mit dynamischer Programmierung

Definiere Matrix D mit |s| * |t| Einträgen:

$$D[i,j] := \text{Distanz zwischen } s[0,i-1] \text{ und } t[0,j-1]$$

Globales Alignment mit dynamischer Programmierung

Definiere Matrix D mit |s| * |t| Einträgen:

$$D[i,j] :=$$
 Distanz zwischen $s[0,i-1]$ und $t[0,j-1]$

Initialisierung:

$$D[i,0] = i$$
$$D[0,j] = j$$

Globales Alignment mit dynamischer Programmierung

Definiere Matrix D mit |s| * |t| Einträgen:

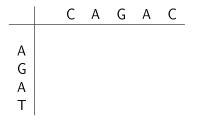
$$D[i,j] :=$$
 Distanz zwischen $s[0,i-1]$ und $t[0,j-1]$

Initialisierung:

$$D[i,0] = i$$
$$D[0,j] = j$$

Rekursion:

$$D[i,j] = \min \left\{ \begin{array}{ll} D[i-1,j-1] & +d(s[i-1],t[j-1]) \\ D[i-1,j] & +1 \\ D[i,j-1] & +1 \end{array} \right.$$



		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1					
A G A T	2					
Α	3					
Т	4					

			Α	G	Α	C
	0	1	2	G 3	4	5
Α	1	1				
G	0 1 2 3 4					
Α	3					
Т	4					

			Α		Α	C
	0	1	2	3	4	5
Α	1	1				
G	2	2				
Α	3					
Т	4			3		

			Α	G	Α	C
	0	1	2	3	4	5
Α	1	1				
G	2	2				
Α	3	3				
Т	4			3		

	C	Α	G	Α	C
0	1	2	3	4	5
1	1				
2	2				
3	3				
4	4				
	0 1 2 3 4	C 0 1 1 1 2 2 3 3 4 4	C A 0 1 2 1 1 2 2 3 3 4 4	C A G 0 1 2 3 1 1 2 2 3 3 4 4	C A G A 0 1 2 3 4 1 1 2 2 3 3 4 4

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1			
G	2	2				
Α	3	3				
Т	4	4		3		

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1			
G	2	2	2			
Α	3	3				
Т	4	4		3		

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1			
G	2	2	2			
Α	3	3	2			
Т	4	4		3		

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1			
G	2	2	2			
Α	3	3	2			
Т	4	4	3	3		

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1	2		
G	2	2	2	3 2		
Α	3	3	2			
Т	4	4	3			

		C	Α	G	Α	
	0	1	2	3	4	5
Α	1	1	1	2		
G	2	2	2	1		
Α	3	3	2			
Τ	4	4	3	3 2 1		

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1	2		
G	2	2	2	1		
Α	3	3	2	2		
Т	4	4	3		4	

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1	2		
G	2	2	2	1		
Α	3	3	2	2	4	
Т	4	4	3	3		

					Α	
	0	1	2	3	4	5
Α	1	1	1	2	3	
G	2	2	2	1	2	
Α	3	3	2	2	4 3 2 1 2	
Т	4	4	3	3	2	

		C	Α	G	Α	C
	0	1	2	3	4	5
Α	1	1	1	2	3	4
G	2	2	2	1	2	3
Α	3	3	2	3 2 1 2 3	1	2
Т	4	4	3	3	2	2

Globales Alignment (Analyse)

Laufzeit und Speicherbedarf:

- Konstante Laufzeit pro Eintrag $o \mathcal{O}(mn)$
- Spaltenweise Berechnung: $\mathcal{O}(m+n)$ Speicherbedarf

Globales Alignment (Analyse)

Laufzeit und Speicherbedarf:

- Konstante Laufzeit pro Eintrag $o \mathcal{O}(mn)$
- Spaltenweise Berechnung: $\mathcal{O}(m+n)$ Speicherbedarf

Berechnung des tatsächlichen Alignments:

- Zusatzmatrix, um gewählten Rekursionszweig zu speichern
- Speicherbedarf steigt auf $\mathcal{O}(mn)$

Gegeben Text T und Muster P

Gesucht "Optimales" Alignment zwischen P und einem Substring von T

Gegeben Text T und Muster P

Gesucht "Optimales" Alignment zwischen P und einem Substring von T

Idee: Passe Algorithmus für globales Alignment an:

Gegeben Text T und Muster P

Gesucht "Optimales" Alignment zwischen P und einem
Substring von T

Idee: Passe Algorithmus für globales Alignment an:

- Erlaube eine kostenlose, beliebig lange Deletion zu Beginn des Textes
- Initialisiere erste Zeile mit Nullen

Gegeben Text T und Muster P

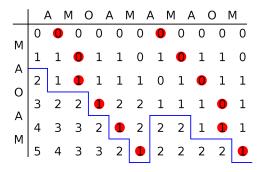
Gesucht "Optimales" Alignment zwischen P und einem Substring von T

Idee: Passe Algorithmus für globales Alignment an:

- Erlaube eine kostenlose, beliebig lange Deletion zu Beginn des Textes
- Initialisiere erste Zeile mit Nullen
- → Komplexität bleibt gleich

Zusätzliche Optimierung:

- Bei vorgegebener Fehlerschranke nicht komplette Tabelle berechnen
- Senkt die erwartete Laufzeit auf etwa O(nk)



Allgemeine Distanzmaße

Needleman-Wunsch-Algorithmus:

- Verallgemeinerung des DP-Algorithmus auf allgemeinere Distanzmaße möglich
- Unterschiedliche Gewichtung von Mismatches und Gaps
- Positionsabhänige Gewichtung für Gaps möglich

Allgemeine Distanzmaße

Needleman-Wunsch-Algorithmus:

- Verallgemeinerung des DP-Algorithmus auf allgemeinere Distanzmaße möglich
- Unterschiedliche Gewichtung von Mismatches und Gaps
- Positionsabhänige Gewichtung für Gaps möglich

Konsequenzen:

- Rekursionsgleichung muss angepasst werden (bei positionsabhängigen Gaps muss ganze Spalte/Zeile pro Zelle berücksichtigt werden)
- Bei lezterem Fall: Verschlechterung der Laufzeit auf $\mathcal{O}(mn(m+n))$
- → Edit-Distanz für unsere Zwecke wahrscheinlich ausreichend!