Theoretische Informatik

Teil 2
Reguläre Ausdrücke

Frühlingssemester 2019

L. Di Caro
D. Flumini
O. Stern

Reguläre Ausdrücke: Einführung

Viele Probleme in der Informatik beinhalten die **Prüfung, ob gewisse Wörter zu einer gegebenen Sprache gehören.**

- Ist eine Benutzereingabe sinnvoll?
- Gehört ein Wort zu einem Wörterbuch (Rechtschreibprüfung)?
- Bestimmte Folgen von “Events” sollen bestimmte Reaktionen auslösen (z. B. Getränkeautomat).

Bemerkung

Da IT-Systeme nur über **endliche Speicherressourcen** verfügen, ist es daher wichtig Sprachen **endlich repräsentieren** zu können.

Unendlich grosse Sprachen werden so einer **maschinellen Bearbeitung** überhaupt erst zugänglich gemacht.

Reguläre Ausdrücke: Einführung

Reguläre Ausdrücke sind Wörter, die Sprachen beschreiben, also eine Möglichkeit (gewisse) Sprachen endlich zu repräsentieren.

- Die **Syntax der regulären Ausdrücke** befasst sich mit der Frage, welche **Form** diese Wörter haben.
- In der **Semantik der regulären Ausdrücke** wird erklärt, wie man reguläre Ausdrücke als Sprachen **interpretiert**.

Reguläre Ausdrücke: Beispiel

Beispiel

Ein regulärer Ausdruck, der die Sprache aller Binärwörter der Länge 4 beschreibt:

$$\underbrace{(0|1)}_{0 \text{ oder } 1}\ \underbrace{(0|1)}_{\text{nochmals}}\ \underbrace{(0|1)}_{\text{dreimal}}\ \underbrace{(0|1)}_{\text{genug}}$$

Ein passender regulärer Ausdruck ist also

$$(0|1)(0|1)(0|1)(0|1)$$

Reguläre Ausdrücke: Beispiel

Beispiel

Ein regulärer Ausdruck für die Sprache der Binärwörter, die das Teilwort 00 enthalten:

$$\underbrace{(0|1)^*}_{0 \text{ oder } 1 \text{ beliebig oft}} \quad \underbrace{00}_{\text{das Teilwort}} \quad \underbrace{(0|1)^*}_{0 \text{ oder } 1 \text{ beliebig oft}}$$

Ein passender regulärer Ausdruck ist also

$$(0|1)^*00(0|1)^*$$

Reguläre Ausdrücke: Syntax

Definition (Reguläre Ausdrücke)

Es sei Σ ein beliebiges Alphabet. Die Sprache RA_Σ der **regulären Ausdrücke** über Σ ist wie folgt definiert:

- $\varnothing, \epsilon \in \mathsf{RA}_\Sigma$
- $\Sigma \subset \mathsf{RA}_\Sigma$
- $R \in \mathsf{RA}_\Sigma \Rightarrow (R^*) \in \mathsf{RA}_\Sigma$
- $R, S \in \mathsf{RA}_\Sigma \Rightarrow (RS) \in \mathsf{RA}_\Sigma$
- $R, S \in \mathsf{RA}_\Sigma \Rightarrow (R|S) \in \mathsf{RA}_\Sigma$

Reguläre Ausdrücke: Erläuterungen

Erläuterungen zur Definition

- Die Sonderzeichen ϵ und $\varnothing$ sind reguläre Ausdrücke.
- Jedes Symbol aus dem Alphabet Σ ist auch ein regulärer Ausdruck über Σ.
- Ist R ein regulärer Ausdruck über Σ, dann ist auch (R^*) ein regulärer Ausdruck über Σ.
- Sind R und S reguläre Ausdrücke über Σ, dann auch (RS) und $(R|S)$.

Reguläre Ausdrücke: Erweiterte Syntax

Bemerkung

Es gibt unzählige Erweiterungen der Sprache der regulären Ausdrücke. Einige verbreitete abkürzende Schreibweisen sind:

- Ist R ein regulärer Ausdruck, dann steht (R^+) für $R(R^*)$.
- Ist R ein regulärer Ausdruck, dann steht $(R?)$ für $(R|\epsilon)$.
- Sind $R_1, \ldots, R_k$ reguläre Ausdrücke, dann steht $[R_1, \ldots, R_k]$ für $R_1|R_2|\ldots|R_k$[1].

Diese abkürzenden Schreibweisen (R^+, $R?$ und $R_1|R_2|\ldots|R_k$) können in der Folge verwendet werden.

[1] Streng genommen müsste man hier $R_1|(R_2|(\ldots|R_k)\ldots)$ schreiben.

Reguläre Ausdrücke: Beispiele

Beispiele

Einige reguläre Ausdrücke über dem Alphabet $\{a, b\}$.

- $((((aa)^*)(b^*))(a(ba)))$
- $((a|(ab))^*)$
- $((ab)|(ba))$
- $(a(b(ba)))$

Reguläre Ausdrücke: Erste Eigenschaften

Eigenschaften und Konventionen:

- Die Menge RA_Σ der regulären Ausdrücke über dem Alphabet Σ ist eine Sprache über dem Alphabet $\{\varnothing, \epsilon, {}^*, (,), |\} \cup \Sigma$.
- Der Lesbarkeit halber werden “überflüssige” Klammern weggelassen.
- Damit reguläre Ausdrücke auch mit (teilweise) weggelassenen Klammen eindeutig lesbar bleiben, gilt folgende Rangfolge der Operatoren:
  - “*” vor “Konkatenation” und
  - “Konkatenation” vor “$|$”.

  Der Ausdruck $ab^*|c$ wird beispielsweise als $((a(b^*))|c)$ gelesen.

Reguläre Ausdrücke: Semantik

Definition (Die Sprache von regulären Ausdrücken)

Es sei Σ ein beliebiges Alphabet. Für jeden regulären Ausdruck $R \in \mathrm{RA}_\Sigma$ definieren wir die **Sprache** $L(R)$ **von** R wie folgt:

- $L(\varnothing) = \varnothing$
- $L(\epsilon) = \{\varepsilon\}$
- $L(a) = \{a\}$ für $a \in \Sigma$
- $L(R^*) = L(R)^*$
- $L(R|S) = L(R) \cup L(S)$
- $L(RS) = L(R)L(S)$

Reguläre Ausdrücke: Semantik

Erläuterungen zur Definition

- $\varnothing$ beschreibt die leere Sprache.
- ϵ beschreibt die Sprache $\{\varepsilon\}$.
- Jedes Symbol $a \in \Sigma$ beschreibt die Sprache $\{a\}$.
- (R^*) beschreibt alle durch Konkatenation kombinierten Wörter, die von R beschrieben werden.
- $(R|S)$ beschreibt alle Wörter, die entweder von R oder von S beschrieben werden.
- (RS) beschreibt die Wörter, die durch Konkatenation aus einem von R beschriebenen Wort gefolgt von einem durch S beschriebenen Wort entstehen.

Reguläre Sprachen

Definition

Eine Sprache A über dem Alphabet Σ heisst *regulär*, falls $A = L(R)$ für einen regulären Ausdruck $R \in \text{RA}_\Sigma$ gilt.

Beispiele

- Für $R = 0|\,(-?)[1,\ldots,9][0,\ldots,9]^*$ gilt

$$L(R) = \text{ Menge der ganzen Zahlen in Dezimaldarstellung}$$

- Für $R = (0?)(10)^*(1?)$ oder $R = (10)^*|(01)^*|(10)^*1|(01)^*0$ gilt

$$L(R) = \begin{array}{l}\text{Menge der Binärwörter mit}\\ \text{abwechselnd Nullen und Einsen}\end{array}$$

Reguläre Sprachen: Erste Eigenschaften

Lemma (Rechenregeln für reguläre Ausdrücke)

Für jedes Alphabet und alle regulären Ausdrücke $R, S, T \in \mathrm{RA}_\Sigma$ gelten folgende Identitäten.

- $L(R|S) = L(S|R)$
- $L(R(ST)) = L((RS)T)$
- $L(R|(S|T)) = L((R|S)|T)$
- $L(R(S|T)) = L(RS|RT)$
- $L((R^*)^*) = L(R^*)$
- $L(R|R) = L(R)$

Beweis.

Elementare Mengenumformungen. □

Reguläre Ausdrücke: Anwendungen

Anwendungen von regulären Ausdrücken:

- Mustersuche in Texten
- Lexikalische Analyse (in Compilern); Erkennung von Schlüsselwörtern (“Token”)
- Syntax Test (bei einer einfachen Syntax)