Symbolic Solving of Extended Regular Expression Inequalities

Matthias Keil, Peter Thiemann
University of Freiburg,
Freiburg, Germany

December 15, 2014, IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science

Extended Regular Expressions

Definition

$$r, s, t \;:=\; \epsilon \mid A \mid r{+}s \mid r{\cdot}s \mid r^* \mid r\&s \mid !r$$

- Σ is a potentially infinite set of symbols
- $A, B, C \subseteq \Sigma$ range over sets of symbols
- $[\![r]\!] \subseteq \Sigma^*$ is the language of a regular expression r, where $[\![A]\!] = A$

Language Inclusion

UNI FREIBURG

Definition

Given two regular expressions r and s,

$$r \sqsubseteq s \Leftrightarrow [\![r]\!] \subseteq [\![s]\!]$$

- $[\![r]\!] \subseteq [\![s]\!]$ iff $[\![r]\!] \cap \overline{[\![s]\!]} = \emptyset$
- Decidable using standard techniques:
  Construct DFA for $r\&!s$ and check for emptiness
- Drawback is the expensive construction of the automaton
- PSPACE-complete

Antimirov's Algorithm

- Deciding containment for *basic regular expressions*
- Based on derivatives and expression rewriting
- Avoid the construction of an automaton
- $\partial_a(r)$ computes a regular expression for $a^{-1}[\![r]\!]$ (Brzozowski) with $u \in [\![r]\!]$ iff $\epsilon \in [\![\partial_u(r)]\!]$

Lemma

For regular expressions r and s,

$$r \sqsubseteq s \Leftrightarrow (\forall u \in \Sigma^*)\ \partial_u(r) \sqsubseteq \partial_u(s).$$

Antimirov's Algorithm (cont'd)

Lemma

$$r \sqsubseteq s \;\Leftrightarrow\; (\nu(r) \Rightarrow \nu(s)) \;\wedge\; (\forall a \in \Sigma)\; \partial_a(r) \sqsubseteq \partial_a(s)$$

CC-Disprove
$$\frac{\nu(r) \wedge \neg\nu(s)}{r \mathrel{\dot\sqsubseteq} s \vdash_{\mathcal{CC}} \mathit{false}}$$

CC-Unfold
$$\frac{\nu(r) \Rightarrow \nu(s)}{r \mathrel{\dot\sqsubseteq} s \vdash_{\mathcal{CC}} \{\partial_a(r) \mathrel{\dot\sqsubseteq} \partial_a(s) \mid a \in \Sigma\}}$$

- Choice of next step's inequality is nondeterministic
- An infinite alphabet requires to compute for infinitely many $a \in \Sigma$

First Symbols

Lemma

$$r \sqsubseteq s \;\Leftrightarrow\; (\nu(r) \Rightarrow \nu(s)) \;\wedge\; (\forall a \in \mathit{first}(r))\; \partial_a(r) \sqsubseteq \partial_a(s)$$

- Let first$(r) := \{a \mid aw \in [\![r]\!]\}$ be the set of first symbols
- Restrict symbols to first symbols of the left hand side
- CC-Unfold does not have to consider the entire alphabet
- For extended regular expressions, first(r) may still be an infinite set of symbols

Problems

- Antimirov's algorithm only works with basic regular expressions or requires a finite alphabet
- Extension of *partial derivatives* (Caron et al.) that computes an NFA from an extended regular expression
- Works on sets of sets of expressions
- Computing derivatives becomes more expensive

Goal

- Algorithm for deciding $[\![r]\!] \subseteq [\![s]\!]$ quickly
- Handle *extended regular expressions*
- Deal effectively with very large (or infinite) alphabets (e.g. Unicode character set)

Solution

- Require finitely many atoms, even if the alphabet is infinite
- Compute derivatives with respect to literals

Representing Sets of Symbols

A literal is a set of symbols $A \subseteq \Sigma$

Definition

A is an element of an *effective* boolean algebra $(U, \sqcup, \sqcap, \bar{\cdot}, \bot, \top)$ where $U \subseteq \wp(\Sigma)$ is closed under the boolean operations.

- For finite (small) alphabets:
  $U = \wp(\Sigma)$, $A \subseteq \Sigma$
- For infinite (or just too large) alphabets:
  $U = \{A \in \wp(\Sigma) \mid A \text{ finite} \vee \overline{A} \text{ finite}\}$
- Second-level regular expressions:
  $\Sigma \subseteq \wp(\Gamma^*)$ with $U = \{A \subseteq \wp(\Gamma^*) \mid A \text{ is regular}\}$
- Formulas drawn from a first-order theory over alphabets
  For example, `[a-z]` represented by $x \geq \texttt{'a'} \wedge x \leq \texttt{'z'}$

Derivatives with respect to Literals

- Definition for $\partial_A(r)$?
- $\partial_a(r)$ computes a regular expression for $a^{-1}[\![r]\!]$ (Brzozowski)

Desired property

$$[\![\partial_A(r)]\!] \stackrel{?}{=} A^{-1}[\![r]\!] = \bigcup_{a \in A} a^{-1}[\![r]\!] = \bigcup_{a \in A} [\![\partial_a(r)]\!]$$

Positive Derivatives on Literals

Definition

$$\delta_A^+(B) \quad := \quad \begin{cases} \epsilon, & B \sqcap A \neq \bot \\ \emptyset, & otherwise \end{cases}$$

Problem

With $A = \{a, b\}$ and $r = (a \cdot c)\&(b \cdot c)$,

$$\begin{aligned} \delta_A^+(r) &= \delta_A^+(a \cdot c)\&\delta_A^+(b \cdot c) \\ &= c\&c \\ &\sqsupseteq \emptyset \end{aligned}$$

Negative Derivatives on Literals

Definition

$$\delta_A^-(B) \quad := \quad \begin{cases} \epsilon, & \overline{B} \sqcap A = \bot \\ \emptyset, & \mathit{otherwise} \end{cases}$$

Problem

With $A = \{a, b\}$ and $r = (a \cdot c)+(b \cdot c)$,

$$\begin{aligned} \delta_A^-(r) &= \delta_A^-(a \cdot c)+\delta_A^-(b \cdot c) \\ &= \emptyset+\emptyset \\ &\sqsubseteq c \end{aligned}$$

Positive and Negative Derivatives

- Extends Brzozowski's derivative operator to sets of symbols.
- Defined by induction and flip on the complement operator

Definition

From $\partial_a(!s) = !\partial_a(s)$, define:

$$\delta_A^+(!r) := !\delta_A^-(r) \qquad \Big| \qquad \delta_A^-(!r) := !\delta_A^+(r)$$

Lemma

For any regular expression r and literal A,

$$[\![\delta_A^+(r)]\!] \supseteq \bigcup_{a \in A} [\![\partial_a(r)]\!] \qquad \Big| \qquad [\![\delta_A^-(r)]\!] \subseteq \bigcap_{a \in A} [\![\partial_a(r)]\!]$$

Literals of an Inequality

UNI FREIBURG

Lemma

$$r \sqsubseteq s \;\Leftrightarrow\; (\nu(r) \Rightarrow \nu(s)) \;\wedge\; (\forall a \in \mathit{first}(r))\; \partial_a(r) \sqsubseteq \partial_a(s)$$

- first(r) may still be an infinite set of symbols
- Use *first literals* as representatives of the *first symbols*

Example

1. Let $r = \{a, b, c, d\} \cdot d^*$, then $\{a, b, c, d\}$ is a first literal
2. Let $s = \{a, b, c\} \cdot c^* + \{b, c, d\} \cdot d^*$, then $\{a, b, c\}$ and $\{b, c, d\}$ are first literals

Literals of an Inequality (cont'd)

UNI FREIBURG

Problem

Let $r = \{a, b, c, d\} \cdot d^*$, $s = \{a, b, c\} \cdot c^* + \{b, c, d\} \cdot d^*$, and $A = \{a, b, c, d\}$, then

$$\delta^+_A(r) \;\dot{\sqsubseteq}\; \delta^+_A(s) \tag{1}$$

$$\delta^+_A(\{a, b, c, d\} \cdot d^*) \;\dot{\sqsubseteq}\; \delta^+_A(\{a, b, c\} \cdot c^*) + \delta^+_A(\{b, c, d\} \cdot d^*) \tag{2}$$

$$d^* \;\dot{\sqsubseteq}\; c^* + d^* \tag{3}$$

- Positive (negative) derivatives yield an upper (lower) approximation
- To obtain the precise information, we need to restrict these literals suitably to *next literals*, e.g. {{a},{b,c},{d}}

Next Literals

$$\begin{array}{lcl}
\mathsf{next}(\epsilon) & = & \{\emptyset\} \\
\mathsf{next}(A) & = & \{A\} \\
\mathsf{next}(r{+}s) & = & \mathsf{next}(r) \bowtie \mathsf{next}(s) \\
\mathsf{next}(r{\cdot}s) & = & \begin{cases} \mathsf{next}(r) \bowtie \mathsf{next}(s), & \nu(r) \\ \mathsf{next}(r), & \neg\nu(r) \end{cases} \\
\mathsf{next}(r^*) & = & \mathsf{next}(r) \\
\mathsf{next}(r\&s) & = & \mathsf{next}(r) \sqcap \mathsf{next}(s) \\
\mathsf{next}(!r) & = & \mathsf{next}(r) \cup \{\bigsqcap\{\overline{A} \mid A \in \mathsf{next}(r)\}\}
\end{array}$$

Definition

Let $\mathfrak{L}_1$ and $\mathfrak{L}_2$ be two sets of disjoint literals.

$$\mathfrak{L}_1 \bowtie \mathfrak{L}_2 \;:=$$

$$\{(A_1 \sqcap A_2), (A_1 \sqcap \overline{\bigsqcup \mathfrak{L}_2}), (\overline{\bigsqcup \mathfrak{L}_1} \sqcap A_2) \mid A_1 \in \mathfrak{L}_1, A_2 \in \mathfrak{L}_2\}$$

Next Literals (cont'd)

Example

Let $s = \{a, b, c\} \cdot c^* + \{b, c, d\} \cdot d^*$, then

$$\begin{aligned} \text{next}(s) &= \text{next}(\{a, b, c\} \cdot c^*) \bowtie \text{next}(\{b, c, d\} \cdot d^*) \\ &= \{\{a, b, c\}\} \bowtie \{\{b, c, d\}\} \\ &= \{\{a\}, \{b, c\}, \{d\}\} \end{aligned}$$

Lemma

For all r,

- $\bigcup next(r) \supseteq first(r)$
- $|next(r)|$ *is finite*
- $(\forall A, B \in next(r))\ A \sqcap B = \emptyset$

Coverage

Lemma

Let $\mathfrak{L} = next(r)$ *and* $A \in next(r) \setminus \{\emptyset\}$.

1. $(\forall a, b \in A)\ \partial_a(r) = \partial_b(r) \ \wedge\ \delta_A^+(r) = \delta_A^-(r) = \partial_a(r)$
2. $(\forall a \notin \bigcup \mathfrak{L})\ \partial_a(r) = \emptyset$

Definition

Let $A' \in \text{next}(r)$. For each $\emptyset \neq A \subseteq A'$ define $\partial_A(r) := \partial_a(r)$, where $a \in A$.

Next Literals of an Inequality

- *Next literal* of $\text{next}(r \mathrel{\dot{\sqsubseteq}} s)$
- Sound to join literals of both sides $\text{next}(r) \bowtie \text{next}(s)$
- Contains also symbols from s
- First symbols of r are sufficient to prove containment

Definition

Let $\mathfrak{L}_1$ and $\mathfrak{L}_2$ be two sets of disjoint literals.

$$\mathfrak{L}_1 \ltimes \mathfrak{L}_2 \ := \{(A_1 \sqcap A_2), (A_1 \sqcap \overline{\bigsqcup \mathfrak{L}_2}) \mid A_1 \in \mathfrak{L}_1, A_2 \in \mathfrak{L}_2\}$$

Left-based join corresponds to $\text{next}(r\&(!s))$.

Definition

Let $r \mathrel{\dot{\sqsubseteq}} s$ be an inequality, define: $\text{next}(r \mathrel{\dot{\sqsubseteq}} s) := \text{next}(r) \ltimes \text{next}(s)$

Solving Inequalities

Lemma

$$r \sqsubseteq s \;\Leftrightarrow\; (\nu(r) \Rightarrow \nu(s)) \;\wedge\; (\forall a \in \mathit{first}(r))\; \partial_a(r) \sqsubseteq \partial_a(s)$$

To determine a finite set of representatives

- select *one* symbol a from each equivalence class $A \in \text{next}(r)$
- calculate with $\delta^+_A(r)$ or $\delta^-_A(r)$ with $A \in \text{next}(r)$

Theorem (Containment)

$$r \sqsubseteq s \;\Leftrightarrow\; (\nu(r) \Rightarrow \nu(s)) \;\wedge\; (\forall \boldsymbol{A} \in \textbf{next}(\boldsymbol{r} \mathrel{\dot{\sqsubseteq}} \boldsymbol{s}))\; \partial_{\boldsymbol{A}}(r) \sqsubseteq \partial_{\boldsymbol{A}}(s)$$

Conclusion

- Generalize Brzozowski's derivative operator
- Extend Antimirov's algorithm for proving containment
- Provides a symbolic decision procedure that works with extended regular expressions on infinite alphabets
- Literals drawn from an effective boolean algebra
- Main contribution is to identify a finite set that covers all possibilities

Regular Languages

The language $[\![r]\!] \subseteq \Sigma^*$ of a regular expression r is defined inductively by:

$$
\begin{aligned}
[\![\epsilon]\!] &= \{\epsilon\} \\
[\![A]\!] &= \{a \mid a \in A\} \\
[\![r{+}s]\!] &= [\![r]\!] \cup [\![s]\!] \\
[\![r{\cdot}s]\!] &= [\![r]\!] \cdot [\![s]\!] \\
[\![r^*]\!] &= [\![r]\!] \cdot [\![r^*]\!] \\
[\![r\&s]\!] &= [\![r]\!] \cap [\![s]\!] \\
[\![!r]\!] &= \overline{[\![r]\!]}
\end{aligned}
$$

Nullable

The *nullable* predicate $\nu(r)$ indicates whether $[\![r]\!]$ contains the empty word, that is, $\nu(r)$ iff $\epsilon \in [\![r]\!]$.

$$
\begin{array}{lcl}
\nu(\epsilon) & = & \mathit{true} \\
\nu(A) & = & \mathit{false} \\
\nu(r{+}s) & = & \nu(r) \vee \nu(s) \\
\nu(r{\cdot}s) & = & \nu(r) \wedge \nu(s) \\
\nu(r^*) & = & \mathit{true} \\
\nu(r\&s) & = & \nu(r) \wedge \nu(s) \\
\nu(!r) & = & \neg\nu(r)
\end{array}
$$

Brzozowski Derivatives

$\partial_a(r)$ computes a regular expression for the left quotient $a^{-1}[\![r]\!]$.

$$
\begin{aligned}
\partial_a(\epsilon) &= \emptyset \\
\partial_a(A) &= \begin{cases} \epsilon, & a \in A \\ \emptyset, & a \notin A \end{cases} \\
\partial_a(r{+}s) &= \partial_a(r){+}\partial_a(s) \\
\partial_a(r{\cdot}s) &= \begin{cases} \partial_a(r){\cdot}s{+}\partial_a(s), & \nu(r) \\ \partial_a(r){\cdot}s, & \neg\nu(r) \end{cases} \\
\partial_a(r^*) &= \partial_a(r){\cdot}r^* \\
\partial_a(r\&s) &= \partial_a(r)\&\partial_a(s) \\
\partial_a(!r) &= !\partial_a(r)
\end{aligned}
$$

First Symbols

Let $\text{first}(r) := \{a \mid aw \in [\![r]\!]\}$ be the set of first symbols derivable from regular expression r.

$$
\begin{array}{lcl}
\text{first}(\epsilon) & = & \emptyset \\
\text{first}(A) & = & A \\
\text{first}(r{+}s) & = & \text{first}(r) \cup \text{first}(s) \\
\text{first}(r{\cdot}s) & = & \begin{cases} \text{first}(r) \cup \text{first}(s), & \nu(r) \\ \text{first}(r), & \neg\nu(r) \end{cases} \\
\text{first}(r^*) & = & \text{first}(r) \\
\text{first}(r\&s) & = & \text{first}(r) \cap \text{first}(s) \\
\text{first}(!r) & = & \Sigma \setminus \{a \in \text{first}(r) \mid \partial_a(r) \neq \Sigma^*\}
\end{array}
$$

First Literals

Let $\text{first}(r) := \{a \mid aw \in [\![r]\!]\}$ be the set of first symbols derivable from regular expression r.

$$
\begin{aligned}
\text{literal}(\epsilon) &= \emptyset \\
\text{literal}(A) &= \{A\} \\
\text{literal}(r+s) &= \text{literal}(r) \cup \text{literal}(s) \\
\text{literal}(r \cdot s) &= \begin{cases} \text{literal}(r) \cup \text{literal}(s), & \nu(r) \\ \text{literal}(r), & \neg\nu(r) \end{cases} \\
\text{literal}(r^*) &= \text{literal}(r) \\
\text{literal}(r\&s) &= \text{literal}(r) \cap \text{literal}(s) \\
\text{literal}(!r) &= \Sigma \sqcap \overline{\bigsqcup\{A \in \text{literal}(r) \mid \partial_A(r) = \Sigma^*\}}
\end{aligned}
$$

Coverage

Lemma (Coverage)

For all a, u, and r it holds that:

$$u \in [\![\partial_a(r)]\!] \iff \exists A \in next(r) : a \in A \land u \in [\![\delta_A^+(r)]\!] \land u \in [\![\delta_A^-(r)]\!]$$

Termination

Theorem (Finiteness)

Let R be a finite set of regular inequalities. Define

$$F(R) = R \cup \{\partial_A(r \mathrel{\dot{\sqsubseteq}} s) \mid r \mathrel{\dot{\sqsubseteq}} s \in R, A \in next(r \mathrel{\dot{\sqsubseteq}} s)\}$$

For each r and s, the set $\bigcup_{i \in \mathbb{N}} F^{(i)}(\{r \sqsubseteq s\})$ is finite.

Decision Procedure for Containment

(DISPROVE)

$$\frac{\nu(r) \qquad \neg\nu(s)}{\Gamma \ \vdash \ r \mathrel{\dot\sqsubseteq} s \ : \ \mathit{false}}$$

(CYCLE)

$$\frac{r \mathrel{\dot\sqsubseteq} s \in \Gamma}{\Gamma \ \vdash \ r \mathrel{\dot\sqsubseteq} s \ : \ \mathit{true}}$$

(UNFOLD-TRUE)

$$\frac{\begin{array}{c} r \mathrel{\dot\sqsubseteq} s \notin \Gamma \qquad \nu(r) \Rightarrow \nu(s) \\ \forall A \in \mathrm{next}(r \mathrel{\dot\sqsubseteq} s) : \ \Gamma \cup \{r \mathrel{\dot\sqsubseteq} s\} \ \vdash \ \partial_A(r) \mathrel{\dot\sqsubseteq} \partial_A(s) \ : \ \mathit{true} \end{array}}{\Gamma \ \vdash \ r \mathrel{\dot\sqsubseteq} s \ : \ \mathit{true}}$$

(UNFOLD-FALSE)

$$\frac{\begin{array}{c} r \mathrel{\dot\sqsubseteq} s \notin \Gamma \qquad \nu(r) \Rightarrow \nu(s) \\ \exists A \in \mathrm{next}(r \mathrel{\dot\sqsubseteq} s) : \ \Gamma \cup \{r \mathrel{\dot\sqsubseteq} s\} \ \vdash \ \partial_A(r) \mathrel{\dot\sqsubseteq} \partial_A(s) \ : \ \mathit{false} \end{array}}{\Gamma \ \vdash \ r \mathrel{\dot\sqsubseteq} s \ : \ \mathit{false}}$$

Prove and Disprove Axioms

(PROVE-IDENTITY)

$$\Gamma \vdash r \sqsubseteq r : \mathit{true}$$

(PROVE-EMPTY)

$$\Gamma \vdash \emptyset \sqsubseteq s : \mathit{true}$$

(PROVE-NULLABLE)

$$\frac{\nu(s)}{\Gamma \vdash \epsilon \sqsubseteq s : \mathit{true}}$$

(DISPROVE-EMPTY)

$$\frac{\exists A \in \mathsf{next}(r) : A \neq \emptyset}{\Gamma \vdash r \sqsubseteq \emptyset : \mathit{false}}$$

Soundness

Theorem (Soundness)

For all regular expression r and s:

$$\emptyset \vdash r \mathrel{\dot{\sqsubseteq}} s : \top \Leftrightarrow r \sqsubseteq s$$

Negative Derivatives

Counterexample

Let $r = \{a, b, c, d\} \cdot d^*$, $s = \{a, b, c\} \cdot d^* + \{b, c, d\} \cdot d^*$, and $A = \{a, b, c, d\}$, then

$$\delta_A^-(r) \;\dot{\sqsubseteq}\; \delta_A^+(s) \tag{4}$$

$$\delta_A^-(\{a, b, c, d\} \cdot d^*) \;\dot{\sqsubseteq}\; \delta_A^-(\{a, b, c\} \cdot d^*) + \delta_A^-(\{b, c, d\} \cdot d^*) \tag{5}$$

$$d^* \;\dot{\sqsubseteq}\; \emptyset + \emptyset \tag{6}$$

Next Literals of an Inequality

Example

Let $r = \{a, b, c, d\}\cdot d^*$, $s = \{a, b, c\}\cdot c^* + \{b, c, d\}\cdot d^*$ then

$$\begin{aligned}\mathrm{next}(r \mathrel{\dot\sqsubseteq} s) &= \mathrm{next}(\{a, b, c, d\}\cdot d^*) \ltimes \mathrm{next}(\{a, b, c\}\cdot d^* + \{b, c, d\}\cdot d^*) \\ &= \{\{a\}, \{b, c\}, \{d\}\}\end{aligned}$$

Incomplete Containment

Conjecture

$$r \sqsubseteq s \;\Leftarrow\; (\nu(r) \Rightarrow \nu(s)) \;\wedge\; (\forall \boldsymbol{A} \in \mathit{literal}(\boldsymbol{r}))\; \delta^{+}_{\boldsymbol{A}}(r) \sqsubseteq \delta^{-}_{\boldsymbol{A}}(s)$$