Membership and finiteness problems for rational sets of regular languages

Sergey Afonin   Eelena Hazova

Lomonosov Moscow State University,
Institute of Mechanics
Moscow, Russian Federation

06.08.2005 / DLT2005

Outline

View-based query processing

- A query is a regular language.
- Query evaluation is a time consuming operation.
- Suppose that results of the queries $\mathcal{E} = \{E_1, \ldots, E_k\}$ are known. Can this data be used during evaluation of an arbitrary query R?
- Representation of the form

$$R = E_{i_1} E_{i_2} \ldots E_{i_n}$$

  may be effectively evaluated.
- The length of the decomposition is not bounded if E_i contains the empty word.

Language equations

Consider linear equation $R = EX$, where R and E are given regular languages and X is a variable.

- It is decidable whether or not this equation has a solution (Kari, 1994).
- The maximal solution is a regular language (Kari, 1994).
- In general there are infinitely many (regular) solutions:

$$\{a^{2n}, n \geqslant 1\}\ X = \{a^n, n \geqslant 1\}$$

- Rational sets could be a suitable formalism for solutions characterization.

Semigroups, languages and morphisms: the notation

Σ and Δ are finite disjoint alphabets. The *empty* word is denoted by ε.

A *semigroup* $(S, \cdot)$ is a set S equipped with associative binary product.

A semigroup *morphism* $\varphi : (S_1, \cdot) \to (S_2, \circ)$ is any function satisfying $\varphi(u \cdot v) = \varphi(u) \circ \varphi(v)$ for all $u, v \in S_1$.

The morphism $\varphi : (\Delta^+, \cdot) \to (2^{\Sigma^*}, \cdot)$ is called a *regular language substitution* if $\varphi(\delta)$ is a regular language over Σ for all $\delta \in \Delta$.

By $(\mathcal{S}, \cdot) = \langle \mathcal{E} \rangle$ we mean the semigroup, generated by $\mathcal{E}$ with concatenation as a semigroup product. $\mathcal{S}_\varphi = \langle \{\varphi(\delta) \mid \delta \in \Delta)\} \rangle$.

Rational set of languages

Definition (Rational set of languages)

A set $\mathcal{R}$ of regular languages over Σ is called *rational* if there exists a finite alphabet Δ, a regular language $K \subseteq \Delta^+$, and a regular language substitution $\varphi : \Delta^+ \to 2^{\Sigma^*}$, such that

$$\mathcal{R} = \{\varphi(w) \mid w \in K\}.$$

- The membership problem: Is it decidable whether or not a regular language R belongs to $\mathcal{R} = (K, \varphi)$?
- The finiteness problem: Is it decidable whether or not a rational set $\mathcal{R} = (K, \varphi)$ is finite?

Hashiguch's representation theorems

Theorem (Hashiguchi, 1983)

For any finite set $\mathcal{E}$ of regular languages over Σ, and a subset $T \subseteq \{\cdot, \cup, ^*\}$ of language operations it is decidable whether or not the language R may be constructed from elements of $\mathcal{E}$ using a finite number of operations from T.

Corollary

It is decidable whether or not R belongs to the semigroup $\mathcal{R} = (\Delta^*, \varphi)$.

Theorem (Calvanese et.al., 2002)

The maximal rewriting *of $R \subseteq \Sigma^*$ wrt φ, $M_\varphi(R) = \{w \in \Delta^+ \mid \varphi(w) \subseteq R\}$, is a regular language over Δ.*

Distance automata

A *distance Σ-automaton* is a tuple

$$\mathcal{A} = \langle \Sigma, Q, \rho, q_\circ, F, d \rangle$$

where $d : \rho \to \{0, 1\}$ is a distance function. The distance $d(\pi)$ of a path π is the sum of the distances of all edges in π. The distance $d(w)$ of a word $w \in L(\mathcal{A})$ is the minimum of $d(\pi)$ for all successful paths π labeled by w.

Limitedness property of a distance automaton

A distance automaton $\mathcal{A}$ is called *limited* if there exists a constant M such that $d(w) < M$ for all words $w \in L(\mathcal{A})$.

Theorem (Hashiguchi, 1982; Leung, Podolskiy 2004)

Limitedness property of a distance automaton is decidable.

Finite section problem

Let $M_1, \ldots, M_d$ be a set of $N \times N$ matrices over a semiring $(K, +, \cdot)$. Define $V_{ij} \subseteq K$ as $\{M_{ij} \mid M \in \langle\{M_1, \ldots, M_d\}\rangle\}$. Is it decidable whether V_{ij} is finite or not?

The limitedness property of a distance automaton may be considered as the finite section problem for the corresponding set of matrices over the *tropical semiring* $\mathcal{N} = (\mathbb{N} \cup \{\infty\}, \min, +)$.

Finite section problem for square matrices over the semiring of regular languages over an unary alphabet is decidable (Mascle, 1986).

Language factorization

- Krohn-Rhodes decomposition theorem for transformation semigroups (Arbib, 1968)
- Language factorization into a finite number of stars and primes (Brzozowski, 1969)
- Finite language substitution problem (Kirsten, 2004; Bala, 2004)

Factors are arbitrary languages of certain classes (i.e. stars and primes or finite languages).

Theorem

For any rational set $\mathcal{R} = (K, \varphi)$ of regular languages over Σ and any regular language $R \subseteq \Sigma^*$ is it decidable whether $R \in \mathcal{R}$ or not.

Proof.

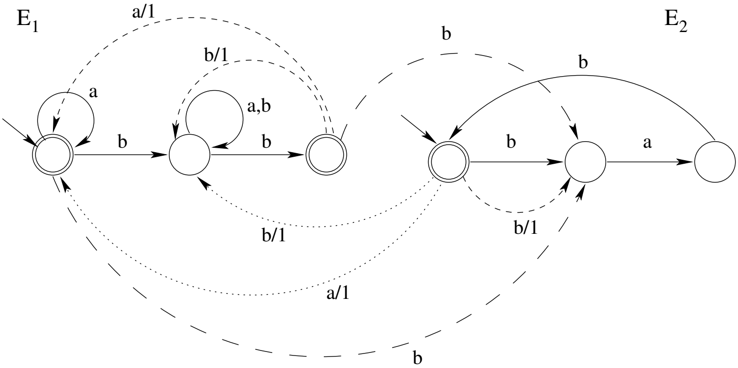
- $R \in \mathcal{R}$ only if $M = M_\varphi(R) \cap K$ is *exact rewriting*:
  $\forall u \in R \ \exists w \in M : u \in \varphi(w)$.
- Assume that M is a union-free language and $\varepsilon \in \varphi(\delta)$ for all $\delta \in \Delta$.

  For any pair of words $u, v \in M$ there exists a word $w \in M$, such that $\varphi(u) \subseteq \varphi(w)$ and $\varphi(v) \subseteq \varphi(w)$.
- $R \in \mathcal{R}$ if the corresponding distance $\Sigma-$automaton $\mathcal{A}_M$ (see next slide) is limited.

□

Example

$$E_1 = \varphi(\delta_1) = a^* + a^* b(a+b)^* b \qquad E_2 = \varphi(\delta_2) = (bab)^*$$
$$R = a^* + a^* b a^* b(a+b)^* \qquad M_\varphi(R) = (\delta_1 + \delta_2)^* = (\delta_1^* \delta_2^*)^*$$



$$\varphi(\delta_1\delta_2\delta_2\delta_1\delta_1) = R$$

The finiteness problem

Let $\mathcal{R} = (K, \varphi)$ be a rational set of regular languages. It is decidable whether $\mathcal{R}$ is finite or not?

Specific cases of the finiteness problem:

- If $K = \Delta^+$ and $|\Delta| = 1$ then we have the *finite power property* problem of regular languages.
- When $K = \Delta^+$, the finiteness problem for $\mathcal{R}$ is equivalent to the finiteness problem of the finitely generated semigroup $\mathcal{S}_\varphi$.
- Finiteness of a rational set.

The finite power property

A regular language L is said to have the *finite power property* if there exists a natural number $k \in \mathbb{N}$ such that $L^* = L^k$.

Examples

$$L_1 = a^* \qquad k = 1$$
$$L_2 = \varepsilon + a^2(a^3)^* \qquad k = 2$$
$$L_3 = (\varepsilon + b^*a)(b + ab^*ab^*a)^* \qquad k = 3$$

We shall write $fpp(L) = p$ or $fpp(L) < \infty$ if $L^* = L^p$, and $fpp(L) = \infty$ if no such number exists.

FPP characterization

Theorem (c.f. Salomaa, 1981)

Let L be a regular language. Then

1. *if fpp*$(L) < \infty$*, then either L is infinite and* $\varepsilon \in L$*, or* $L = \{\varepsilon\}$*,*
2. *fpp*$(L) < \infty$ *if and only if there exist* $k, n \in \mathbb{N}$*, such that* $L^k = L^{k+n+1}$*,*
3. *fpp*$(L) = \infty$ *if and only if there exists* $w \in L^*$*, such that* $w^n \notin L^n$ *for all* $n \in \mathbb{N}$*.*

Theorem (Hashiguchi, 1982)

Finite power property is decidable for regular languages.

Finiteness of a free semigroup

Proposition
Let $(\mathcal{S}, \cdot) = \langle \mathcal{E} \rangle$ be a finitely generated semigroup of regular languages over Σ. The semigroup $\mathcal{S}$ is finite only if $fpp(E) < \infty$ for all $E \in \mathcal{E}$.

If and only if? No.

$E_1 = a^*$ and $E_2 = b^*$, the semigroup is infinite because $(a^* b^*)^* = (a + b)^*$, but for any $p \in \mathbb{N}$ the language $(a^* b^*)^p$ consists of all words with at most p "*b* to *a* letter switches".

The semigroup given by presentation $S = \langle \Delta \mid x^3 = x^2 \text{ for all } x \in \Delta^+ \rangle$ is infinite even if $|\Delta| = 2$ (Brzozowski et.al., 1971).

Finiteness of a free semigroup (cont)

Theorem (A. De Luca and A. Restivo, 1984)

A finitely generated semigroup S is finite if and only if there exists a positive number m such that, for any sequence $s_1, \ldots, s_m$ *of elements of S, there exist integers* $j, k,\ 1 \leqslant j \leqslant k \leqslant m$*, such that*

$$s_1 \ldots s_k = s_1 \ldots s_{j-1}(s_j \ldots s_k)^2.$$

This condition, however, does not provide an upper bound for *m* and can not be used as a decision procedure.

The finiteness condition of $\mathcal{S}_\varphi$

Theorem

Let $\varphi : \Delta \to 2^{\Sigma^*}$ be a regular language substitution. The semigroup $\mathcal{S}_\varphi$ is finite if and only if for any m-subset $\{\delta_1, \ldots, \delta_m\} \subseteq \Delta$ ($m = 1, \ldots, |\Delta|$)

$$fpp(\varphi(\delta_1 \delta_2 \ldots \delta_m)) < \infty.$$

Proof.

- Denote $\varphi(\delta_i)$ by E_i. $\varepsilon \in E_i$ for all E_i.
- $(E_{\sigma_1} E_{\sigma_2} \ldots E_{\sigma_k})^* = (E_1 + E_2 + \ldots + E_k)^*$.
- $fpp(E_1 E_2 \ldots E_k) < \infty \Rightarrow fpp(E_{\sigma_1} E_{\sigma_2} \ldots E_{\sigma_k}) < \infty$.
- By induction on $|\Delta|$. Let $|\Delta| = n$ and $fpp(\delta_1 \ldots \delta_n) = p$.
  $w = w_{11} \delta_1 w_{12} \delta_2 \ldots w_{1n} \delta_n w_{21} \delta_1 w_{22} \delta_2 \ldots w_{2n} \delta_n \ldots \ldots w_{pn} \delta_n \tilde{w}$
  $\varphi(w) \supseteq \varphi((\delta_1 \delta_2 \ldots \delta_n)^p) \supseteq \varphi(w)$

□

Finite semigroup example

There exist non-trivial finite semigroups of regular languages.

Example

Let $\Sigma = \{a, b\}$, $E_1 = b^*a + (b^*ab^*ab^*)^*$, and $E_2 = (b + ab^*a)^*$. The languages E_1 and E_2 satisfy the finite power property and do not commute ($E_1E_2 \neq E_2E_1$). One can verify that $(E_1E_2)^* = (E_1E_2)^2$.

Finiteness of a rational set

Let $K = K_1 S_1^* K_2 S_2^* \dots K_m S_m^* K_{m+1}$, where $K_i \in \Delta^*$ and $S_i \in \Delta^+$.

v_1	v_2	v_3	$\dots$	v_m	$\mathcal{A}_{\gamma(\mathbf{v})}$ limit	
*	*	*		*	p_0	
p_0	*	*		*	p_1^1	
*	p_0	*		*	p_1^2	
		$\dots$			$\dots$	
*	*	*		p_0	p_1^m	$p_1 = \max\{p_1^1, \dots, p_1^m\}$
			$\dots$			
*	p_{m-1}	p_{m-1}		p_{m-1}	p_m^1	
p_{m-1}	*	p_{m-1}	$\dots$	p_{m-1}	p_m^2	
		$\dots$			$\dots$	
p_{m-1}	p_{m-1}	p_{m-1}	$\dots$	*	p_m^m	$p_m = \max\{p_m^1, \dots, p_m^m\}$

Rational set (K, φ) is finite if $p_m < \infty$.

Complexity

- Limitedness property of a distance automaton has exponential complexity (wrt the number of states of the automaton)
- Exponential lower bound for maximal rewriting construction is proved (Calvanese, 2002).
- Only exponential time complexity factorization algorithms are known for finite languages (Salomaa, Yu, 1999).
- A possible solution is a reduction to the membership problem over unary alphabet using a language morphism $f : \Sigma \to \{a\}$. If corresponding problem for one-letter alphabet has no solution then the original problem has no solution either.

Equivalence and intersection

For rational sets $\mathcal{R}_1 = (K_1, \varphi_1)$ and $\mathcal{R}_2 = (K_2, \varphi_2)$

- Is it decidable whether or not $\mathcal{R}_1 = \mathcal{R}_2$, and
- Is the intersection $\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_2$ a rational set?

Example

$\varphi(\delta_1) = \{a\}$ $\varphi(\delta_2) = \{ba\}$ $\varphi(\delta_3) = \{bb^*\}$	$\varphi(\gamma_1) = \{ab\}$ $\varphi(\gamma_2) = \{b^*\}$
$K_1 = \delta_1\delta_2^*\delta_3$	$K_2 = \gamma_1^*\gamma_1\gamma_2$

$$(K_1, \varphi_1) = (K_2, \varphi_2)$$

Δ- and K-minimality

Definitions:

- We say that a representation (K, φ) of a rational set $\mathcal{R}$ is *K-minimal* if $(K', \varphi) \subset \mathcal{R}$ for any regular language $K' \subset K$.
- Similarly, we say that a representation (K, φ) is *Δ-minimal* if $\mathcal{R}$ has no representation (K', φ') in a smaller alphabet Δ'.

Questions:

- Is it decidable whether a given representation is Δ- or K-minimal?
- Is it possible to compute Δ- or K-minimal representation for a given rational set?

Summary

- The membership and finiteness problems for rational set of regular languages are decidable.
- The proof is based on limitedness property of a distance automaton.
- Open problems include:
  - Equivalence and intersection.
  - $K-$ and Δ-minimality of a representation.