Regular Languages and Their Generating Functions: The Inverse Problem

Christoph Koutschan

Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany

Research Institute for Symbolic Computation
Johannes Kepler Universität Linz, Austria

Combinatorics Seminar
RISC, 18th January 2006

Short Overview

Short Overview

Short Overview

Short Overview

What is a Regular Language?

- Regular grammar (N, Σ, P, S)
- Alphabet (of terminals) Σ
- Set of nonterminal symbols N
- Production rules in P may have the form
  - $A \rightarrow a$
  - $A \rightarrow aB$
  - $A \rightarrow \lambda$
- Regular expression
- Accepted by a deterministic finite automaton

Examples

Consider the regular language given by the grammar
$G = (N, \Sigma, P, S)$ with
$N = \{A, S\}$,
$\Sigma = \{a, b, c\}$,
$P = \{S \rightarrow aS, S \rightarrow bA, A \rightarrow \lambda, A \rightarrow cA\}$.

- $L_G = \{b, ab, bc, aab, abc, bcc, aaab, aabc, \ldots\}$
- Regular expression: a^*bc^*

Connection to Power Series

The formal power series

$$S = \sum_{n=0}^{\infty} s_n x^n$$

is called the generating function (or characteristic series) of a formal language L, if

$$s_n = \Big|\{w \in L : |w| = n\}\Big|,$$

i.e., if the n^{th} coefficient of the series S gives the number of words in L having the length n.

Schützenberger Methodology

- Algorithm to obtain the generating function from a given grammar
- In order to compute the generating function for L_G, the morphism Θ is defined:

$$\Theta(a) = x, \quad \forall a \in \Sigma$$
$$\Theta(\lambda) = 1$$
$$\Theta(A) = A(x), \quad \forall A \in N$$

- Applying Θ to all elements of P yields a system of algebraic equations in $A(x), B(x), \ldots$
- Solving for $S(x)$ gives the generating function for L_G.

Our Goal

Last Chapter

Get the generating function from a language.

Now: The Inverse Problem

Given the characteristic series, find a regular expression for the corresponding language.

Question

Is this always possible?

Our Goal

Last Chapter

Get the generating function from a language.

Now: The Inverse Problem

Given the characteristic series, find a regular expression for the corresponding language.

Question

Is this always possible?

Our Goal

Last Chapter

Get the generating function from a language.

Now: The Inverse Problem

Given the characteristic series, find a regular expression for the corresponding language.

Question

Is this always possible?

Subgoals

Answer

The answer unfortunately is no!

This divides the problem into two subgoals:

- Check whether a corresponding regular language exists.
- Compute a regular expression for this.

Power Series over an Alphabet

Definition: Formal Power Series

Given an alphabet Σ and a semiring $\mathbb{K}$. A *formal power series* S is a function

$$S : \Sigma^* \to \mathbb{K}.$$

The image of a word w under S is the *coefficient* s_w.
S is written as a formal sum

$$S = \sum_{w \in \Sigma^*} s_w w.$$

The set of all formal power series over Σ^* with coefficients in $\mathbb{K}$ is denoted by $\mathbb{K}\langle\langle\Sigma^*\rangle\rangle$.

Quasiregularity and Star

Definition: Quasiregularity

A power series (especially a polynomial) $S \in \mathbb{K}\langle\langle\Sigma^*\rangle\rangle$, is called *quasiregular* if the coefficient of the neutral element of Σ^* vanishes, i.e., if $s_\lambda = 0$.

Definition: Star (Kleene closure)

$$S^* = \lim_{m \to \infty} \sum_{n=0}^{m} S^n$$

This limes exists only for quasiregular series!

Quasiregularity and Star

Definition: Quasiregularity

A power series (especially a polynomial) $S \in \mathbb{K}\langle\langle\Sigma^*\rangle\rangle$, is called *quasiregular* if the coefficient of the neutral element of Σ^* vanishes, i.e., if $s_\lambda = 0$.

Definition: Star (Kleene closure)

$$S^* = \lim_{m\to\infty} \sum_{n=0}^{m} S^n$$

This limes exists only for quasiregular series!

Rational Operations

- Rational operations:
  - Sum
  - (Cauchy-) Product
  - Star
- $M \subseteq \mathbb{K}\langle\langle\Sigma^*\rangle\rangle$ is *rationally closed* if it is closed w.r.t. the rational operations.
- $\mathbb{K}^{\mathrm{rat}}\langle\langle\Sigma^*\rangle\rangle$: Rational closure of $\mathbb{K}\langle\Sigma^*\rangle$
- S is called $\mathbb{K}$-*rational* if it is an element of $\mathbb{K}^{\mathrm{rat}}\langle\langle\Sigma^*\rangle\rangle$.

Theorem of Schützenberger

Definition: Recognizable

A formal series $S \in \mathbb{K}\langle\langle\Sigma^*\rangle\rangle$ is called *recognizable* if its coefficients can be written as follows:

$$s_w = \alpha \cdot \mu(w) \cdot \beta,$$

where $\alpha \in \mathbb{K}^{1,n}$, $\beta \in \mathbb{K}^{n,1}$, and $\mu : \Sigma^* \to \mathbb{K}^{n,n}$ $(n \geq 1)$ is a multiplicative homomorphism of monoids.

Theorem (Schützenberger)

A formal series $S \in \mathbb{K}\langle\langle\Sigma^*\rangle\rangle$ is $\mathbb{K}$-rational if and only if S is recognizable.

Connection to Regular Languages

Theorem

Let L be a regular language and $\mathbb{K}$ a semiring. Then the characteristic series of L is $\mathbb{K}$-rational.

Theorem

The support of any series $S \in \mathbb{N}^{\mathrm{rat}}\langle\langle\Sigma^*\rangle\rangle$ is a regular language.

Connection to Regular Languages

Theorem

Let L be a regular language and $\mathbb{K}$ a semiring. Then the characteristic series of L is $\mathbb{K}$-rational.

Theorem

The support of any series $S \in \mathbb{N}^{\mathrm{rat}}\langle\langle\Sigma^*\rangle\rangle$ is a regular language.

Basic Definitions

Definition: Poles and Roots

Let S be a rational power series and $f(x) = p(x)/q(x)$ its normalized generating function.
Then the roots of $q(x)$ are called *poles* of S.
The roots of the reciprocal polynomial $\bar{q}(x)$ are called *roots* of S.

Definition: Dominating Root

Let $\lambda_0, \ldots, \lambda_r$ be the roots of the rational power series S.
λ_0 is called *dominating root* if

$$\begin{aligned} &\lambda_0 \in \mathbb{R}_+ \text{ and} \\ &\lambda_0 > |\lambda_i|, 1 \leq i \leq r. \end{aligned}$$

Characterization of Rational Series in a Ring

$S \in \mathbb{K}^{\mathrm{rat}}\langle\langle x^* \rangle\rangle$ ($\mathbb{K}$ now a commutative ring)

$\iff$ S has generating function $\dfrac{p(x)}{1-q(x)}$ (q quasiregular)

$\iff$ $s_n = q_1 s_{n-1} + \cdots + q_k s_{n-k}$, $q_i \in \mathbb{K}$ (for large n).

Moreover, for infinite power series (i.e., not a polynomial):

$$S \in \mathbb{K}^{\mathrm{rat}}\langle\langle x^* \rangle\rangle \iff s_n = \sum_{i=0}^{r} P_i(n)\lambda_i^n \text{ (for large } n\text{)},$$

where

- $\lambda_0, \ldots, \lambda_r$: distinct roots with multiplicities $m_0, \ldots, m_r$
- P_i: complex nonzero polynomials with $\deg P_i = m_i - 1$

Introductory Example

From now on we are interested in positive series.
Consider the series A094423 from Sloane's Encyclopedia:

$$x + 4x^2 + x^3 + 144x^4 + 361x^5 + 484x^6 + 19321x^7 + 28224x^8 + \ldots$$

which is generated by the function

$$\frac{x + 5x^2}{1 + x - 5x^2 - 125x^3}.$$

Although all coefficients of this series are positive integers the series is not $\mathbb{N}$-rational. Later we will see why.

Crucial Property of N-rational Series

Theorem

Let $S \in \mathbb{N}^{\mathrm{rat}}\langle\langle x^* \rangle\rangle \setminus \mathbb{N}\langle x^* \rangle$ have the generating function $f(x)$ and the roots $\lambda_0, \ldots, \lambda_r$ and let $\varrho := \min_{0 \le i \le r} |\lambda_i^{-1}|$. Then the following statement holds:

ϱ is a pole of S (let m_ϱ be its multiplicity) and all other poles of modulus ϱ have the form $\varrho\vartheta$ and a multiplicity $\le m_\varrho$ (ϑ denotes a complex root of unity, i.e., $\exists\ p \in \mathbb{N} : \vartheta^p = 1$).

Decomposing and Merging

Definition: Decomposition and Merge

For any $p \in \mathbb{N}$ the list of series $S_0, \ldots, S_{p-1}$ is called a *decomposition* of S if

$$S_i = \sum_{n=0}^{\infty} s_{i+np} x^n .$$

On the other hand S is termed the *merge* of $S_0, \ldots, S_{p-1}$:

$$S(x) = \sum_{i=0}^{p-1} x^i S_i(x^p).$$

Decomposing and Merging

Example for p=3

$$S_0 = s_0 + s_3 x + s_6 x^2 + \dots$$
$$S_1 = s_1 + s_4 x + s_7 x^2 + \dots$$
$$S_2 = s_2 + s_5 x + s_8 x^2 + \dots$$

Rationality under Decomposition

Theorem

Let $\mathbb{K}$ be a semiring. $S \in \mathbb{K}\langle\langle x^* \rangle\rangle$ is $\mathbb{K}$-rational if and only if for any $p \in \mathbb{N}$ there exists a set of $\mathbb{K}$-rational power series $S_0, S_1, \ldots, S_{p-1}$ and their merge is S.

Remark: If $\mathbb{K}$ is commutative then the roots $\mu_0, \ldots, \mu_s$, $s \leq r$ of S_j are from the set $\{\lambda_0^p, \ldots, \lambda_r^p\}$, and any root μ_k of S_j has the multiplicity

$$m'_k \leq \max_{0 \leq i \leq r} \{m_i : \lambda_i^p = \mu_k\}.$$

Characterization of N-rational Series

Lemma

Let $S \in \mathbb{N}\langle\langle x^* \rangle\rangle$ be $\mathbb{Z}$-rational with dominating root λ_0.
Then S is $\mathbb{N}$-rational.

Theorem

A series $S \in \mathbb{N}\langle\langle x^* \rangle\rangle$ is $\mathbb{N}$-rational if and only if it is a merge of rational series each of them having a dominating root.

General Strategy

- Given a rational function
- Compute the roots
- Search for a dominating root
- In case of several roots with maximal modulus:
  - Compute decomposition
  - Search for a dominating root in each subseries
- Check whether all coefficients are nonnegative
- In case of $\mathbb{N}$-rationality: Compute a regular expression

Not so easy!

Problem

For roots λ_i and λ_j decide whether

$$|\lambda_i| > |\lambda_j|, \; |\lambda_i| = |\lambda_j|, \text{ or } |\lambda_i| < |\lambda_j|!$$

Maple is not capable to maintain this task by symbolic computation:

```
lambda1:= RootOf(x^5+2*x^4+3, index=1);
lambda2:= RootOf(x^5+2*x^4+3, index=5);
   evalb(abs(lambda1)=abs(lambda2));
```

gives false!

Estimate

Theorem (Gourdon, Salvy)

Let p be a polynomial with integer coefficients, $\alpha_1, \ldots, \alpha_n$ its roots and thus $\deg p = n > 0$ its degree. Define

$$\kappa(p) = \frac{\sqrt{3}}{2} \left(\frac{n(n+1)}{2} \right)^{-\left(\frac{1}{4}n(n+1)+1\right)} \cdot M(p)^{-\frac{1}{2}n\left(n^2+2n-1\right)},$$

then $|\alpha_i| \neq |\alpha_j| \implies \left| |\alpha_i| - |\alpha_j| \right| \geq \kappa(p)$ and $|\mathrm{Im}(\alpha_i)|$ is either 0 or larger than $\kappa(p)$. Herein $M(p)$ is defined by

$$M(p) := |p_n| \prod_{i=1}^{n} \max\{1, |\alpha_i|\}.$$

Caution!

Be careful: Evil example yields $\kappa(q) \doteq 2.159917528 \cdot 10^{-287579}$, although the dominating root differs already in the second digit!!!

Strategy: First numerical computation with few digits, and if necessary, in a second step high precision.

Identifying Roots of Unity

- Define the symmetrical polynomial $R(x) := \prod_{\substack{0 \leq i,j \leq r \\ i \neq j}} (\lambda_i - \lambda_j x)$.
- R has integral coefficients.
- R has the roots λ_i / λ_j, $0 \leq i, j \leq r$
- Assume $\lambda_i = \varrho\vartheta$ for some root of unity ϑ, $\varrho \in \mathbb{R}_+$
- If the series is $\mathbb{N}$-rational then all roots of unity $\vartheta_0, \ldots, \vartheta_{n-1}$ are roots of R.
- $R(x)$ must be divisible by the n^{th} cyclotomic polynomial $\Phi_n(x)$.

Identifying Roots of Unity

- Compute R via `resultant`
- Factor R
- Check if among the factors are some cyclotomic polynomials (use `invphi`)
- The least common multiple of the orders gives the number of subseries!
- Result: In the decomposed series we have no (multiples of) roots of unity any more.

Computing the Decomposition

Theorem

Given a series S by its generating function $f(x)$, and an integer p (number of subseries). Then

$$f_i(x) = \frac{1}{px^{i/p}} \sum_{j=1}^{p} s^{p-ij} f(s^j x^{1/p}), \quad s = e^{2\pi i/p}$$

is the generating function for the subseries S_i.

Take care!

$\rightarrow$ This formula leads to vast computations!

Tricks for improving:

- Substitute $x^{1/p}$ by a new variable y
- Recall: $\mathbb{Q}[\vartheta] \cong \mathbb{Q}\langle x^* \rangle / \Phi_p(x)$, where ϑ is a p^{th} primitive root of unity
- Introduce a new variable s which represents $e^{2\pi i/p}$, and compute modulo $\Phi_p(s)$

Check Nonnegativity of Coefficients

- Recall: We can write the coefficients by means of the exponential polynomial

$$s_n = \sum_{i=0}^{r} P_i(n)\lambda_i^n$$

- Compute a boundary n_0 such that $s_n \geq 0$ for $n > n_0$.
- Check $s_0, \ldots, s_{n_0}$ by hand!

Regular Expression

For sake of completeness: Here is the formula for computing a regular expression:

$$S = \frac{1}{R^{(p)}}\left(T^{[h]} + \gamma_k s_h x^{h+k}(cx)^* + z(x)\right) + c s_h x^{h+1}(cx)^* + \sum_{n=0}^{h} s_n x^n.$$

- Recursion over the multiplicity of the dominating root
- The integer constant c must fulfill $\lambda_0 > c > \max_{1 \le i \le r} |\lambda_i|$ (and some other conditions).
- Further decomposition may be necessary!

Hofstadter's MIU-System

- From the book "Gödel, Escher, Bach"
- $\Sigma = \{\mathsf{M}, \mathsf{I}, \mathsf{U}\}$
- Start with MI

Rules

1. $w\mathsf{I} \longrightarrow w\mathsf{IU}$
2. $\mathsf{M}w \longrightarrow \mathsf{M}ww$
3. $\mathsf{III} \longrightarrow \mathsf{U}$
4. $\mathsf{UU} \longrightarrow \lambda$

Question: Does MU belong to the language?

Hofstadter's MIU-System

The generating function for this language is

$$x \mapsto \frac{x^2}{1 - 3x + 3x^2 - 2x^3}$$

and the corresponding power series is

$$x^2 + 3x^3 + 6x^4 + 11x^5 + 21x^6 + 42x^7 + 85x^8 + \dots$$

The "regular expression" computed by our program is

$$(x^2)^* \left(x^2 \left(2 + 5x^2 + 9x^4(x^2)^*\right)\right)^* x^2(2x+1)(x^2+x+1)$$

Look and Say!

The sequence is obtained by looking and saying:

$$1, 11, 21, 1211, 111221, 312211, 13112221, 1113213211, \ldots$$

- John Conway's "Cosmological Theorem"
- 92 strings build up the sequence
- Each of them develops without influencing the others
- "Audioactive Decay"

Look and Say!

We do not consider the Look and Say Sequence itself, but the lengths of its words.

$$1, 11, 21, 1211, 111221, 312211, 13112221, 1113213211, \ldots$$

$$1 + 2x + 2x^2 + 4x^3 + 6x^4 + 6x^5 + 8x^6 + 10x^7 + \ldots$$

This sequence is generated by the following monstruous rational function:

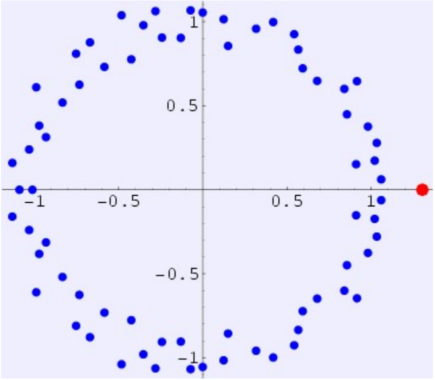
Numerator

$$p(x) = -12x^{78} + 18x^{77} - 18x^{76} + 18x^{75} - 18x^{74} + 20x^{73} + 22x^{72} - 31x^{71} - 15x^{70} + 4x^{69} + 4x^{68} + 19x^{67} - 62x^{66} + 50x^{65} + 21x^{64} + 11x^{63} - 41x^{62} - 54x^{61} + 56x^{60} + 44x^{59} - 15x^{58} + 27x^{57} + 15x^{56} - 45x^{55} + 8x^{54} - 89x^{53} + 64x^{52} + 66x^{51} + 25x^{50} - 38x^{49} - 126x^{48} + 39x^{47} + 32x^{46} + 33x^{45} + 65x^{44} - 107x^{43} - 14x^{42} - 16x^{41} + 13x^{40} + 79x^{39} - 7x^{38} - 42x^{37} - 12x^{36} - 8x^{35} + 26x^{34} + 9x^{33} - 35x^{32} + 23x^{31} + 20x^{30} + 30x^{29} - 34x^{28} - 58x^{27} + x^{26} + 20x^{25} + 36x^{24} + 6x^{23} - 13x^{22} - 8x^{21} - 6x^{20} - 3x^{19} + x^{18} + 4x^{17} + x^{16} + 4x^{15} + 5x^{14} + x^{13} - 8x^{12} - 6x^{11} + 6x^{9} + 4x^{8} - x^{7} - x^{5} - x^{4} - x^{3} - x^{2} + x + 1$$

Denominator

$$q(x) = 6x^{72} - 9x^{71} + 9x^{70} - 18x^{69} + 16x^{68} - 11x^{67} + 14x^{66} - 8x^{65} + x^{64} - 5x^{63} + 7x^{62} + 2x^{61} + 8x^{60} - 14x^{59} - 5x^{58} - 5x^{57} + 19x^{56} + 3x^{55} - 6x^{54} - 7x^{53} - 6x^{52} + 16x^{51} - 7x^{50} + 8x^{49} - 22x^{48} + 17x^{47} - 12x^{46} + 7x^{45} + 5x^{44} + 7x^{43} - 8x^{42} + 4x^{41} - 7x^{40} - 9x^{39} + 13x^{38} - 4x^{37} - 6x^{36} + 14x^{35} - 14x^{34} + 19x^{33} - 7x^{32} - 13x^{31} + 2x^{30} - 4x^{29} + 18x^{28} - x^{26} - 4x^{25} - 12x^{24} + 8x^{23} - 5x^{22} + 8x^{20} + x^{19} + 7x^{18} - 8x^{17} - 5x^{16} - 2x^{15} + 3x^{14} + 3x^{13} - 2x^{8} - x^{7} + 3x^{5} + x^{4} - x^{3} - x^{2} - x + 1$$

The Roots



Result

- Computation takes a few hours, but it works!
- Check the result by replacing $*$ by $x \mapsto 1/(1-x)$
- Regular expression is several pages long (not cited here)...

Acknowledgement

Thanks to

- Volker Strehl for advising my diploma thesis,
- Peter Paule for inviting me to RISC,
- Natee Tongsiri for explaining the Beamer package,
- The audience for attention and patience!

Final remark: Thesis, Maple worksheet, and also these slides can be found on my RISC personal homepage!