

The Generalized Star-Height Problem

Jonah Sherman

May 8, 2007

1 Introduction

The minimal number of nested Kleene stars required in a regular expression representing a language provides a simple complexity measure on the regular languages. For restricted regular expressions in which complementation is not allowed, much is known about the measure. In particular, Eggan(1963) showed that for any n , there is a language of star-height n [1]. Hashiguchi(1983) then gave a method of computing the star-height of an arbitrary regular language[2].

However, when one allows complementation as a regular operation, one obtains the generalized regular expressions. In that case, not much is known about the problem. There is no known decision procedure to decide if a language has a particular star-height. In fact, it is not even known whether there is a language of star-height 2. We present some of the known results on the star-height problem, and describe some ideas that may lead the path to a solution.

2 Definitions

We begin by defining the *generalized regular expressions* over some alphabet Σ , recursively:

- \emptyset, ε are gres
- For all $\sigma \in \Sigma$, σ is a gre
- If A, B are gres, then $A \cup B$ and AB are gres
- If A is a gre, A^* and \overline{A} are gres

The complement operator is why these expressions are called “generalized”.

We then define the *star-height* of a gre, to be the number of nested Kleene stars it contains. Formally,

- $h(\emptyset) = h(\varepsilon) = h(\sigma) = 0$ for all $\sigma \in \Sigma$
- $h(A \cup B) = h(AB) = \max\{h(A), h(B)\}$
- $h(\overline{A}) = h(A)$

- $h(A^*) = h(A) + 1$.

Then, for each language A , define the star-height of the language A , $|A|$, to be the minimum over all regular expressions E representing A of $h(E)$.

$$|A| = \min_E h(E)$$

We can now state the actual problem. The *generalized star-height problem* is the question “is there an algorithm to determine the star-height of a regular language?”. The *star-height 2 problem* asks “is there a language of star-height 2?”. Since we are exclusively concerned with the generalized star-height, we shall refer to it simply as star-height.

3 Established Results

Nearly all work to date has been focused on determining the star-height of particular algebraic classes of languages. To describe some of these, we’ll first briefly describe the relevant concepts from algebraic automata theory.

The *syntactic congruence* of a language $L \subseteq \Sigma^*$ is the congruence \sim_L on Σ^* defined by,

$$x \sim_L y : \iff \forall u, v \in \Sigma^* [uxv \in L \iff uyv \in L]$$

Intuitively, in terms of regular languages, the relation corresponds to strings that are indistinguishable by L ’s minimal DFA. That is, $x \sim_L y$ iff for any given state q in L ’s minimal DFA, when the machine is in state q , reading x or y will bring the machine to the same state. As far as the machine is concerned, x and y are identical. Also, if $x \in L$ and $x \sim_L y$, then $y \in L$, so L itself is a union of equivalence classes of the congruence.

Then, Σ^* / \sim_L is the *syntactic monoid* of L . The syntactic congruence and monoid were introduced by Rabin and Scott[3].

Theorem (Rabin-Scott 1959). *Let $L \subseteq \Sigma^*$. Then, L is regular iff the syntactic monoid of L is finite.*

More detail, along with the proof, can be found in [3]. A much more thorough presentation of the algebraic theory of languages is given in [4], [5].

Schützenberger used this algebraic theory of regular languages to completely characterize the languages of star-height 0[6]. A monoid M is *aperiodic* iff for all $x \in M$, there exists $n \in \mathbb{N}^+$ such that $x^{n+1} = x^n$. If M is finite, one can take the largest n over all $x \in M$, and hence for a finite monoid, the same definition holds with the order of quantifiers reversed.

Theorem (Schützenberger 1965). *Let $L \subseteq \Sigma^*$ be a regular language. Then, L has star-height 0 iff the syntactic monoid of L is aperiodic.*

A much more concise proof of the forward direction is given in [4]. Note that the theorem trivially implies the hierarchy does not collapse at 0.

Corollary. *There is a language of star-height 1.*

Proof. Let $L = (bb)^*$ over $\{a, b\}$. The syntactic monoid of L is not aperiodic, since for any $n \in \mathbb{N}$, $b^n \in L \iff b^{n+1} \notin L$, so $[b]^n \neq [b]^{n+1}$. Hence, L has star-height 1. \square

All further results on the problem have been classifications of certain algebraic families of languages as having star-height 1. Most of the existing results are summarized in [7]. In particular, enough families of groups have been characterized to know that all languages whose syntactic monoids are groups of order ≤ 12 have star-height ≤ 1 .

As an example of such a classification, I managed to prove that all languages whose syntactic monoids are commutative have star-height ≤ 1 . However, it turned out this was already known[7], though I could not locate the cited reference (it seems to be contained in an MIT PhD thesis).

Theorem. *Let $L \subseteq \Sigma^*$ be regular. If the syntactic monoid of L is commutative, then L has star-height ≤ 1 .*

Proof. Let $\Sigma = \{\sigma_1, \dots, \sigma_k\}$. Let $M = \Sigma^* / \sim_L$, and assume M is commutative. Let $\varphi m = [m]$. Since L is a union of equivalence classes and M is finite, it suffices to show $\varphi^{-1}(\{m\})$ has star-height ≤ 1 for each $m \in M$, as then $L = \bigcup_{m \in \varphi(L)} \varphi^{-1}(\{m\})$.

Let $m \in M$ be given. Then, for all $x = x_1 \dots x_n \in \Sigma^*$, we have, $x \in \varphi^{-1}(\{m\}) \iff \varphi(x) = m$. Then by commutativity,

$$\varphi(x) = \varphi(x_1 \dots x_n) = \prod_{i=1}^n \varphi(x_i) = \prod_{i=1}^k \varphi(\sigma_i)^{|x|_{\sigma_i}}$$

For $q \in M$, let t_q, p_q be the transient and period of q (that is, the least $t_q \geq 0$ and $p_q > 0$ such that $q^{t_q} = q^{t_q + p_q}$). For $i, j \in \mathbb{N}$, say $i \equiv j \pmod{t, p}$ iff $i = j$ or $i, j \geq t$ and $i \equiv j \pmod{p}$. Then, $q^i = q^j$ iff $i \equiv j \pmod{t_q, p_q}$.

Let $L(\sigma, i, t, p) = \{x \in \Sigma^* : |x|_{\sigma} \equiv i \pmod{t, p}\}$. We'll show $L(\sigma, i, t, p)$ has star-height ≤ 1 . Let $S = \Sigma \setminus \{\sigma\}$. Then, $L(\sigma, i, t, p)$ is defined by the expression $A^*(\sigma A^*)^i$ for $i < t$, and $A^*(\sigma A^*)^i((\sigma A^*)^p)^*$ for $i \geq t$, both of which have star-height ≤ 1 .

Let

$$T = \left\{ (\alpha_1, \dots, \alpha_k) \in (t_{\varphi\sigma_1} + p_{\varphi\sigma_1}) \times \dots \times (t_{\varphi\sigma_k} + p_{\varphi\sigma_k}) : \prod_{i=1}^k \varphi(\sigma_i)^{\alpha_i} = m \right\}$$

Then,

$$\varphi^{-1}(\{m\}) = \bigcup_{(\alpha_1, \dots, \alpha_k) \in T} \bigcap_{i=1}^k L(\sigma_i, \alpha_i, t_{\varphi\sigma_i}, p_{\varphi\sigma_i})$$

\square

Unfortunately, it seems such algebraic characterizations are ultimately a dead-end, and algebraic theory alone cannot characterize all languages, unless all languages have star-height ≤ 1 .

Theorem (J.E. Pin). *Let $L \subseteq A^*$. Then, there is a finite set B , regular language $K \subseteq B^*$ of star-height ≤ 1 , and homomorphism $\varphi : A^* \rightarrow B^*$ such that $L = \varphi^{-1}K$.*

The theorem shows no algebraic characterization can be given for those languages of star-height n , unless all languages have star-height ≤ 1 . Pin believes the latter statement to be false.

4 Further Ideas

I will now discuss some ideas I came up while thinking about the problem. I believe there is a language of star-height 2, so I spent almost all of my time trying to find a language of star-height 2, rather than trying to solve the star-height problem itself. But before describing those ideas, I'll briefly describe my thoughts on the actual star-height problem.

One can ignore the search problem and think about the star-height problem purely as a decision problem: is there an algorithm that, given a language A and number n , decides if A has star-height n . Clearly a solution to the search problem yields a solution to the decision problem. Conversely, if one has a solution to the decision problem, one can simply enumerate $n = 0, 1, 2, \dots$. Since every regular language has some regular expression representing it, for some n the decision procedure must answer “yes”. Hence, one really need only be concerned with the decision problem.

To establish decidability, it would suffice to prove a computable bound on the length of a regular expression of a given language. That is, it would suffice to find some f such that if a language A has any regular expression of star-height n , then it has one no longer than $f(A, n)$. However, it may be the case that proving a regular expression is minimal in terms of length is just as hard (or harder) than proving a regular expression is minimal in terms of star-height. If not, though, bounds on the minimal length of regular expressions would immediately yield a solution to the star-height problem. If it turns out to be decidable, of course, the next question will be its complexity. The issue of complexity of the star-height-0 decision problem could even be looked at already.

One issue that becomes apparent with the problem is it is not very robust. For example, the language $(bb)^*$ has star-height 1 over the alphabet $\{a, b\}$, but has star-height 0 over the alphabet $\{aa, ab, ba, bb\}$. Hence, it is possible that the star-height of a language may vary greatly depending on the particular alphabet used. This suggests perhaps a more general problem. Given some monoid M and subset M' , one can define the regular sets generated by M' as the least set containing all subsets of M' and closed under the regular operations. Perhaps strings of words with concatenation are not the right monoid

to study the problem with, and the answer will become clear by studying another monoid. Unfortunately, that is quite difficult, as nearly all of the other “common” monoids fall into one of the classes already characterized to have star-height ≤ 1 .

I personally believe there is a language of star-height 2. In particular, I conjecture the language over $\{a, b\}$ defined by,

$$((aa \cup ab \cup ba)^* bb (aa \cup ab \cup ba)^* bb)^*$$

has star-height 2. My intuition is that this language must keep track of two things simultaneously: it must count that the number of occurrences of bb that are evenly aligned is even. However, I’ve been unable to prove it.

The most recent idea I’ve had, and the one that looks to be promising is to study the kernel relation of the Kleene star. That is, the equivalence relation on regular languages defined by,

$$A \sim B : \iff A^* = B^*$$

If a regular expression contains some expression E^* , then one can replace E with any E' for which the languages of E and E' are equivalent. I hope to study the star-height problem by studying properties of the equivalence classes of this relation. The rest of this report will be devoted to investigating interesting properties of the relation.

It would be desirable to find some kind of simple, canonical representative element from each equivalence class. Since the Kleene star is not injective, of course one cannot obtain A from A^* . However, we can obtain some information back about the original set. Define $\bar{\varepsilon}A$ to be $A \setminus \{\varepsilon\}$, and define the *core* of a language A by,

$$\text{core } A := (\bar{\varepsilon}A) \setminus (\bar{\varepsilon}A)(\bar{\varepsilon}A)^*$$

Intuitively, the core of A are those nonempty strings in A that cannot be formed by concatenation of shorter strings in A . That is, $\text{core } A$ gives us a simplified language equivalent to the original language.

Theorem. $A \sim \text{core } A$.

Proof. First, $\text{core } A \subseteq \bar{\varepsilon}A \subseteq A$, so by isotonicity, $(\text{core } A)^* \subseteq A^*$. For the reverse inclusion, let $x \in A^*$. Choose $x_1, \dots, x_n \in \bar{\varepsilon}A$ such that $x = x_1 \cdots x_n$, and n is maximized.

If $n = 0$, then $x = \varepsilon \in (\text{core } A)^*$. Otherwise, we argue each $x_i \in \text{core } A$. Suppose not; that is, suppose some $x_i \in (\bar{\varepsilon}A) \setminus \text{core } A$. Note that,

$$\text{core } A = (\bar{\varepsilon}A) \setminus \bigcup_{k=2}^{\infty} (\bar{\varepsilon}A)^k$$

Thus, we may choose $k \geq 2$ and y_1, \dots, y_k such that $x_i = y_1 \cdots y_k$. But then $x = x_1 \cdots x_{i-1} y_1 \cdots y_k x_{i+1} \cdots x_n$, contradicting the maximality of n .

Hence, each $x_i \in \text{core } A$, so $x = x_1 \cdots x_n \in (\text{core } A)^*$. Thus, $A^* = (\text{core } A)^*$, and so $A \sim \text{core } A$. \square

Theorem. $\text{core } A = \text{core } A^*$.

Proof. First, let $x \in \text{core } A$. Then, $x \in \bar{\varepsilon} A$, and for all $k \geq 2$, $x \notin (\bar{\varepsilon} A)^k$. Then, $x \in \bar{\varepsilon} A^*$. We argue $x \notin \bigcup_{k=2}^{\infty} (\bar{\varepsilon} A^*)^k$. Suppose not. Then, we may choose $k \geq 2$ and $x_1, \dots, x_k \in \bar{\varepsilon} A^*$ such that $x = x_1 \cdots x_k$. Then, for each $i \in [k]$, we may choose $m_i \geq 1$ and $y_{i,1}, \dots, y_{i,m_i} \in \bar{\varepsilon} A$ such that $x_i = y_{i,1} \cdots y_{i,m_i}$. But then,

$$x = x_1 \cdots x_k = y_{1,1} \cdots y_{1,m_1} y_{2,1} \cdots y_{2,m_2} \cdots y_{k,1} \cdots y_{k,m_k} \in (\bar{\varepsilon} A)^{m_1+m_2+\cdots+m_k}$$

a contradiction. Hence, $x \in \text{core } A^*$.

Conversely, let $x \in \text{core } A^*$. Then, $x \in \bar{\varepsilon} A^*$, and for all $k \geq 2$, $x \notin (\bar{\varepsilon} A^*)^k$. Thus, we may choose $n \geq 1$ and $x_1, \dots, x_n \in \bar{\varepsilon} A$ such that $x = x_1 \cdots x_n$. Note that $x_i \in \bar{\varepsilon} A \subseteq \bar{\varepsilon} A^*$, so $x = x_1 \cdots x_n \in (\bar{\varepsilon} A^*)^n$. Thus, we must have $n = 1$. Then, $x = x_1 \in \bar{\varepsilon} A$. Also, $\bar{\varepsilon} A \subseteq \bar{\varepsilon} A^*$, so $x \notin (\bar{\varepsilon} A)^k$ for any $k \geq 2$. Thus, $x \in \text{core } A$. \square

Corollary. core is idempotent.

Proof. By the previous two theorems, for each $A \subseteq \Sigma^*$,

$$\text{core}(\text{core } A) = \text{core}(\text{core } A)^* = \text{core } A^* = \text{core } A$$

\square

Corollary. $A \sim B \iff \text{core } A = \text{core } B$.

Proof. Let $A, B \subseteq \Sigma^*$. If $A \sim B$ then,

$$\text{core } A = \text{core } A^* = \text{core } B^* = \text{core } B$$

Also, if $\text{core } A = \text{core } B$, then,

$$A^* = (\text{core } A)^* = (\text{core } B)^* = B^*$$

so $A \sim B$. \square

From these results, we see the core is the most reasonable candidate for an “un- $*$ ”. The goal is that the core, along with the relation, will allow for more properties of the equivalence classes to be determined.

References

- [1] L. C. Eggan. Transition graphs and the star-height of regular events. *Michigan Math. J.* **10**, 385-397.
- [2] K. Hashiguchi. Representation theorems on regular languages. *J. Comput. System Sci.* **27**, (1983), 101-115

- [3] M. O. Rabin, D. Scott. Finite Automata and Their Decision Problems. *IBM Journal of Research and Development* **3**. (1959) 114-125.
- [4] Mateescu, Solomaa. *Formal Languages*.
- [5] J.-E. Pin. *Syntactic Semigroups*.
- [6] M. P. Schützenberger. On finite monoids having only trivial subgroups. *Information and Control* **8**. (1965) 190-194.
- [7] J.-E. Pin, H. Straubing et D. Thérien. Some results on the generalized star-height problem. *Information and Computation* **101**. (1992). 219-250