Subwords and Stars

Tom Bourne
tom.bourne@st-andrews.ac.uk

School of Mathematics and Statistics
University of St Andrews

York Semigroup
5th October 2016

University of St Andrews | FOUNDED 1413

Regular Expressions

A – finite alphabet.

Define $\emptyset$, ε, and each $a \in A$ to be basic regular expressions.

Let E, F be regular expressions. Recursively define new regular expressions by:

- EF (concatenation)
- $E \cup F$ (set union)
- E^* (star)

Application: 'search and replace' in text.

Example

$a \cup ab^*c$ represents $\{a, ac, abc, abbc, abbbc, \dots\}$.

Regular Languages

Language – subset of free semigroup/monoid generated by A.

Any language that can be represented by a regular expression is regular.

Example

If $A = \{a, b\}$ then $A^*a = (a \cup b)^*a$ represents the regular language in which all words end with the letter a.

Simplest class of languages:

Regular $\subset$ context-free $\subset$ context-sensitive $\subset$ recursive $\subset$ recursively enumerable.

Star-Height

The star-height of a regular expression is defined recursively:

- $h(\emptyset) = h(\varepsilon) = h(a) = 0$, where $a \in A$;
- $h(EF) = h(E \cup F) = \max\{h(E), h(F)\}$;
- $h(E^*) = h(E) + 1$.

For a language L, define the star-height of L by

$$h(L) = \min\{h(E) \mid E \text{ is a regular expression for } L\}.$$

Star-height $\leftrightarrow$ minimum nesting-depth of stars.

Theorem (Eggan (1963))

There exist regular languages of star-height n for all $n \geq 0$.

Generali(s $\cup$ z)ed Extensions

Lemma

The class of regular languages is closed under complementation.

Can use generalised regular expressions (i.e. those with complementation included) without introducing non-regular languages.

Define $h(E^c) = h(E)$.

Generalised star-height of a language as in the restricted case.

De Morgan's laws allow use of $\cap$ and $\setminus$ too. It follows that

$$h(E \cap F) = h(E \setminus F) = \max\{h(E), h(F)\}.$$

Recognisability and Equivalencies

Automaton – machine with input, accepts or rejects.

Definition

A language L is recognised by a monoid M if $\exists$ a morphism $\varphi : A^* \to M$ such that $L = L\varphi\varphi^{-1}$.

Theorem

Let L be a language. TFAE:

- *L is regular;*
- *L is accepted by a finite state automaton;*
- *L is recognised by a finite monoid.*

Generalised Star-Height Problem

A language which has (generalised) star-height zero is star-free.

Theorem (Schützenberger (1965))

A regular language is star-free if and only if it is recognised by a finite aperiodic monoid.

Schützenberger $\Rightarrow$ can determine if a language is star-free.

Generalised Star-Height Problem

Does there exist an algorithm that determines the generalised star-height of a regular language? In particular, does there exist a language of generalised star-height greater than 1?

Counting Scattered Subwords

Definition

A word $w = a_1 a_2 \dots a_r$ is a scattered subword of a word v if v can be written as $v = v_0 a_1 v_1 a_2 \dots a_r v_r$ for some $v_0, \dots, v_r \in A^*$.

$\binom{v}{w}$ – number of times w appears as a scattered subword of v.

Define the language ScatModCount(w, k, n) by

$$\text{ScatModCount}(w, k, n) = \left\{ v \in A^* \mid \binom{v}{w} \equiv k \pmod{n} \right\}$$

$\forall w \in A^+, k \geq 0, n \geq 2$ such that $0 \leq k < n$.

Known Results and Motivation

Theorem (Thérien (1983))

Let L be a regular language. Then, L is recognised by a finite nilpotent group of class m if and only if L is a boolean combination of languages of the form $\mathrm{ScatModCount}(w, k, n)$*, where* $|w| \leq m$.

Theorem (Henneman (1971))

Every language recognised by a finite commutative group is of star-height at most 1.

Theorem (Pin, Straubing, Thérien (1989))

Every language recognised by a finite nilpotent group of class 2 is of star-height at most 1.

Class 3: partial result, difficult. Consider contiguous subwords...

Counting Contiguous Subwords

Let $u, w, x \in A^*$. If $v = uwx$ then u is a prefix of v, w is a (contiguous) subword of v, and x is a suffix of v.

$|v|_w$ – number of times w appears as a subword of v.

Define the languages $\text{Count}(w, k)$ and $\text{ModCount}(w, k, n)$ by

$$\text{Count}(w, k) = \{v \in A^* \mid |v|_w = k\}$$

and

$$\text{ModCount}(w, k, n) = \{v \in A^* \mid |v|_w \equiv k \pmod{n}\}$$

$\forall w \in A^+, k \geq 0, n \geq 2$ such that $0 \leq k < n$.

Main Result

Theorem (TB, Ruškuc (in preparation))

Let A be a finite alphabet. Then,

$$h(\mathrm{Count}(w, k)) = 0$$

and

$$h(\mathrm{ModCount}(w, k, n)) \leq 1$$

$\forall w \in A^+, k \geq 0, n \geq 2$ *such that* $0 \leq k < n$.

Overlapping Subwords

Occurrences of w might (and in many cases, do) overlap!

Definition

A prefix of a word that is also a suffix of that word is a border.

Example

If $v = aabaabaa$ then $\{\varepsilon, a, aa, aabaa, aabaabaa\}$ is the set of borders of v.

First, restrict attention to

$$\mathrm{CountWithBorder}(w, k) = wA^* \cap \mathrm{Count}(w, k) \cap A^* w.$$

Notation

Let

$$B = \{b \in A^+ \mid w = bx \text{ and } w = yb \text{ for some } x, y \in A^+\},$$

the set of all proper, non-empty borders of w;

$$P = \{p \in A^+ \mid w = pb \text{ for some } b \in B\},$$

the set of prefixes of w after each border is removed as a suffix; and,

$$S = \{s \in A^+ \mid w = bs \text{ for some } b \in B\},$$

the set of suffices of w after each border is removed as a prefix.

A Problem?

Consider CountWithBorder($aabaabaa, k$).

$B = \{aabaa, aa, a\}$.

$S = \{baa, baabaa, abaabaa\}$.

Now, $aabaabaa \cdot baabaa$ contains 3 occurrences of $aabaabaa$.

Easier if each appended suffix adds on 1 new occurrence.

Introduce

$$\bar{S} = \{s \in S \mid \nexists s' \in S \text{ such that } s = s'x \text{ for some } x \in A^+\}.$$

A Proposition

Let

$$F = (A^* w A^* \cup S A^* \cup A^* P \cup \{x \in A^* \mid w = b_1 x b_2 \text{ for some } b_1, b_2 \in B\})^c.$$

Proposition

CountWithBorder$(w, k) =$

$$\bigcup_{j=1}^{k} \bigcup_{\substack{k_1, k_2, \dots, k_j \geq 0 \\ k_1 + k_2 + \cdots + k_j = k - j}} w \bar{S}^{k_1} F w \bar{S}^{k_2} F \dots F w \bar{S}^{k_j}.$$

This is a star-free expression.

Back to the Theorem

Theorem (TB, Ruškuc (in preparation))

Let A be a finite alphabet. Then,

$$h(\mathrm{Count}(w,k)) = 0$$

and

$$h(\mathrm{ModCount}(w,k,n)) \leq 1$$

$\forall w \in A^+, k \geq 0, n \geq 2$ *such that* $0 \leq k < n$.

Proof.

Write $\mathrm{Count}(w,k)$ as

$$(\emptyset^c w \emptyset^c \cup \emptyset^c P)^c \cdot \mathrm{CountWithBorder}(w,k) \cdot (S\emptyset^c \cup \emptyset^c w \emptyset^c)^c.$$

Similar idea for $\mathrm{ModCount}(w,k,n)$. □

Algebraic Applications

$S = M^0[G; I, \Lambda; P]$ – Rees zero-matrix semigroup over a group G.

Using our result with words of length two aids in the proof of:

Theorem (TB, Ruškuc (to appear))

Regular languages recognised by Rees zero-matrix semigroups over commutative groups are of generalised star-height at most 1.

Rees' Theorem

Finite semigroup zero-simple $\Leftrightarrow$ isomorphic to Rees zero-matrix semigroup over group.

First step towards characterisation of languages recognised by finite simple semigroups.

Future Work

- What effect does replacing ‘scattered subwords’ with ‘contiguous subwords’ have on Thérien (1983)?
- What is the generalised star-height of a language recognised by a Rees zero-matrix semigroup over a nilpotent group of class 2? (Conjecture: 1.)
- Filling in the gaps for counting scattered subwords of length 3.

Thank you!