Orbits of Language Operations: Finiteness and Upper Bounds

Émilie Charlier[1] Mike Domaratzki[2] Tero Harju[3] Jeffrey Shallit[1]

[1]University of Waterloo [2]University of Manitoba [3]University of Turku

Theory Seminar
Toronto, August 12, 2011

Closure operations

Let $x : 2^{\Sigma^*} \to 2^{\Sigma^*}$ be an operation on languages. Suppose x satisfies the following three properties:

1. $L \subseteq x(L)$ (expanding);
2. If $L \subseteq M$ then $x(L) \subseteq x(M)$ (inclusion-preserving);
3. $x(x(L)) = x(L)$ (idempotent).

Then we say that x is a closure operation.

Example

Kleene closure, positive closure, prefix, suffix, factor, subword.

Some notation and a first result

If $x(L) = y(L)$ for all languages L, then we write $x \equiv y$.

We write $\epsilon(L) = L$ and $xy = x \circ y$, that is, $xy(L) = x(y(L))$.

Define c to be the complementation: $c(L) = \Sigma^* - L$. In particular, we have $cc \equiv \epsilon$.

Theorem

Let x, y be closure operations. Then $xcycxcy \equiv xcy$.

Proof of the previous result

$\forall L,\ xcycxcy(L) \subseteq xcy(L)$:

We have: $\forall L,\ L \subseteq y(L)$.
Then: $\forall L,\ cy(L) \subseteq c(L)$.
Then: $\forall L,\ xcy(L) \subseteq xc(L)$.
Then: $\forall L,\ xcy(cxcy(L)) \subseteq xc(cxcy(L)) = xcy(L)$.

$\forall L,\ xcy(L) \subseteq xcycxcy(L)$:

We have: $\forall L,\ L \subseteq x(L)$.
Then: $\forall L,\ cy(L) \subseteq x(cy(L))$.
Then: $\forall L,\ cxcy(L) \subseteq ccy(L) = y(L)$.
Then: $\forall L,\ ycxcy(L) \subseteq yy(L) = y(L)$.
Then: $\forall L,\ cy(L) \subseteq cycxcy(L)$.
Finally: $\forall L,\ xcy(L) \subseteq xcycxcy(L)$.

Corollary (Peleg 1984, Brzozowski-Grant-Shallit 2009)

Let x be any closure operation and L be any language.
If $S = \{x, c\}$, then the orbit $\mathcal{O}_S(L) = \{y(L)\colon y \in S^*\}$ contains at most 14 languages, which are given by the images of L under the 14 operations

$$\epsilon,\ x,\ c,\ xc,\ cx,\ xcx,\ cxc,\ xcxc,\ cxcx,$$
$$xcxcx,\ cxcxc,\ xcxcxc,\ cxcxcx,\ cxcxcxc.$$

NB: This result is the analogous for languages of Kuratowski-14 sets-theorem for topological spaces.

Orbits of languages

Given a set S of operations, we consider the orbit of languages $\mathcal{O}_S(L) = \{x(L) \ : \ x \in S^*\}$ under the monoid generated by S.

So compositions of operations in S are considered as “words over the alphabet S”.

We are interested in the following questions:

- When is the monoid $S^*/_\equiv$ finite?
- Is the cardinality of $\mathcal{O}_S(L)$ bounded, independently of L?

Operations with infinite orbit

The orbit of L under an arbitrary operation need not be finite.

Example

Consider the operation q defined by

$$q(L) = \{x \in \Sigma^* \colon x \text{ is a proper prefix of some } y \in L\}.$$

Let $M = \{a^n b^n \colon n \geq 1\}$. Then the orbit

$$\mathcal{O}_{\{q\}}(M) = \{M, q(M), q^2(M), q^3(M), \ldots\}$$

is infinite, since we have

$$a^{i+1}b \in q^i(M) \quad \text{and} \quad a^{i+1}b \notin q^j(M) \quad \text{for} \quad j > i.$$

The situation is somewhat different if L is regular:

Theorem

Let L be a regular language accepted by a DFA of n states. Then $|\mathcal{O}_{\{q\}}(L)| \leq n$, and this bound is tight.

To see that the bound is tight, consider the language $L_n = \{\epsilon, a, a^2, \ldots, a^{n-2}\}$, which is accepted by a n state DFA. Then $q(L_n) = L_{n-1}$, so this shows $|\mathcal{O}_{\{q\}}(L_n)| = n$.

It is possible for the orbit under a single operation to be infinite even if the operation is expanding and inclusion-preserving.

Example

Consider the operation of fractional exponentiation, defined by

$$n(L) = \{x^\alpha \colon x \in L \text{ and } \alpha \geq 1 \text{ rational}\} = \bigcup_{x \in L} x^+ p(\{x\}).$$

Let $M = \{ab\}$. Then the orbit

$$\mathcal{O}_{\{n\}}(M) = \{M, n(M), n^2(M), n^3(M), \ldots\}$$

is infinite, since we have

$$aba^i \in n^i(M) \;\text{ and }\; aba^i \notin n^j(M) \text{ for } j < i.$$

Some notation and definitions

If t, x, y, z are words with $t = xyz$, we say

- x is a prefix of t;
- z is a suffix of t; and
- y is a factor of t.

If $t = x_1 y_1 x_2 y_2 \cdots x_n y_n x_{n+1}$ for some words x_i and y_j, we say

- $y_1 \cdots y_n$ is a subword of t.

Thus a factor is a contiguous block, while a subword can be "scattered".

Further, x^R denotes the reverse of the word x.

8 natural operations on languages

$$k\colon L \mapsto L^* \qquad s\colon L \to \mathrm{suff}(L)$$
$$e\colon L \mapsto L^+ \qquad f\colon L \to \mathrm{fact}(L)$$
$$c\colon L \mapsto \overline{L} = \Sigma^* - L \qquad w\colon L \to \mathrm{subw}(L)$$
$$p\colon L \mapsto \mathrm{pref}(L) \qquad r\colon L \to L^R$$

where

$$L^* = \cup_{n\geq 0} L^n \quad \text{and} \quad L^+ = \cup_{n\geq 1} L^n$$
$$\mathrm{pref}(L) = \{x \in \Sigma^* \colon x \text{ is a prefix of some } y \in L\}$$
$$\mathrm{suff}(L) = \{x \in \Sigma^* \colon x \text{ is a suffix of some } y \in L\}$$
$$\mathrm{fact}(L) = \{x \in \Sigma^* \colon x \text{ is a factor of some } y \in L\}$$
$$\mathrm{subw}(L) = \{x \in \Sigma^* \colon x \text{ is a subword of some } y \in L\}$$
$$L^R = \{x \in \Sigma^* \colon x^R \in L\}$$

Kuratowski identities

We now consider the set $S = \{k, e, c, p, s, f, w, r\}$.

Lemma

The 14 operations k, e, p, s, f, w, kp, ks, kf, kw, ep, es, ef, and ew are closure operations.

Theorem (mentioned above)

Let x, y be closure operations. Then $xcycxcy \equiv xcy$.

Together, these two results thus give $196 = 14^2$ separate identities.

Further identities

Lemma

Let $a \in \{k, e\}$ *and* $b \in \{p, s, f, w\}$*. Then* $aba \equiv bab \equiv ab$*.*

In a similar fashion, we obtain many kinds of Kuratowski-style identities involving the operations k, e, c, p, s, f, w, and r.

Proposition

Let $a \in \{k, e\}$ *and* $b \in \{p, s, f, w\}$*. Then we have the following identities:*

- $abcacaca \equiv abca$
- $bcbcbcab \equiv bcab$
- $abcbcabcab \equiv abcab$

Additional identities (I)

We obtain many additional identities connecting the operations k, e, c, p, s, f, w, and r.

Proposition

We have the following identities:

- $rp \equiv sr$; $rs \equiv pr$
- $rf \equiv fr$; $rw \equiv wr$; $rc \equiv cr$; $rk \equiv kr$
- $ps \equiv sp \equiv pf \equiv fp \equiv sf \equiv fs \equiv f$
- $pw \equiv wp \equiv sw \equiv ws \equiv fw \equiv wf \equiv w$
- $rkw \equiv kw \equiv wk$
- $ek \equiv ke \equiv k$
- $fks \equiv pks$; $fkp \equiv skp$
- $rkf \equiv skf \equiv pkf \equiv fkf \equiv kf$

Additional identities (II)

Proposition

For all languages L, we have

- $pcs(L) = \Sigma^*$ *or* $\emptyset$.
- *The same result holds for* pcf, fcs, fcf, scp, scf, fcp, wcp, wcs, wcf, pcw, scw, fcw, *and* wcw.

Let's prove this for pcs:

If $s(L) = \Sigma^*$, then $cs(L) = \emptyset$ and $pcs(L) = \emptyset$.

Otherwise, $s(L)$ omits some word w.
Then $s(L) \cap \Sigma^* w = \emptyset$.
Then $\Sigma^* w \subseteq cs(L)$.
Then $\Sigma^* = p(\Sigma^* w) \subseteq pcs(L)$, hence $pcs(L) = \Sigma^*$.

Additional identities (III)

Proposition

For all languages L, we have

- $sckp(L) = \Sigma^*$ *or* $\emptyset$.
- *The same result holds for* $fckp$, $pcks$, $fcks$, $pckf$, $sckf$, $fckf$, $wckp$, $wcks$, $wckf$, $wckw$, $pckw$, $sckw$, $fckw$.

Proposition

For all languages L, we have

- $scskp(L) = \Sigma^*$ *or* $\emptyset$.
- *The same result holds for* $pcpks$.

Additional identities (IV)

Proposition

For all languages L and for all $b \in \{p, s, f, w\}$, we have

- $kcb(L) = cb(L) \cup \{\epsilon\}$
- $kckb(L) = ckb(L) \cup \{\epsilon\}$
- $kckck(L) = ckck(L) \cup \{\epsilon\}$
- $kbcbckb(L) = bcbckb(L) \cup \{\epsilon\}$.

Let's prove $kcp(L) \subseteq cp(L) \cup \{\epsilon\}$:

Assume $x \in kcp(L)$ and $x \neq \epsilon$.
We have $x = x_1 x_2 \cdots x_n$ for some $n \geq 1$, where each $x_i \in cp(L)$.
Then $x_1 x_2 \cdots x_n \notin p(L)$, because if it were, then $x_1 \in p(L)$.
Hence $x \in cp(L)$.

Main Result

Theorem (C-Domaratzki-Harju-Shallit 2011)

Let $S = \{k, e, c, p, f, s, w, r\}$. Then for every language L, the orbit $\mathcal{O}_S(L)$ contains at most 5676 distinct languages.

Sketch of the proof

We used breadth-first search to examine the set $S^* = \{k, e, c, p, f, s, w, r\}^*$ w.r.t. the radix order with $k < e < c < p < f < s < w < r$.

As each new word x is examined, we test it to see if any factor is of the form given by "certain identities".

If it is, then the corresponding language must be either Σ^*, $\emptyset$, $\{\epsilon\}$, or Σ^+; furthermore, each descendant language will be of this form. In this case the word x is discarded.

Otherwise, we use the remaining identities to try to reduce x to an equivalent word that we have previously encountered. If we succeed, then x is discarded.

Otherwise we append all the words in Sx to the end of the queue.

Sketch of the proof (cont'd)

If the process terminates, then $\mathcal{O}_S(L)$ is of finite cardinality.

For $S = \{k, e, c, p, f, s, w, r\}$, the process terminated with 5672 nodes that could not be simplified using our identities. We did not count $\emptyset, \{\epsilon\}, \Sigma^+$, and Σ^*. The total is thus 5676.

(The longest word examined was *ckcpcpckpckpckpcpcpckckcr*, of length 25, and the same word with p replaced by s.)

If we use two arbitrary closure operations a and b with no relation between them, then the monoid generated by $\{a, b\}$ is infinite, since any two finite prefixes of $ababab\cdots$ are distinct.

Example

Define the exponentiation operation

$$t(L) = \{x^i : x \in L \text{ and } i \text{ is an integer} \geq 1\}.$$

Then t is a closure operation.

Hence the orbits $\mathcal{O}_{\{p\}}(L)$ and $\mathcal{O}_{\{t\}}(L)$ are finite, for all L.

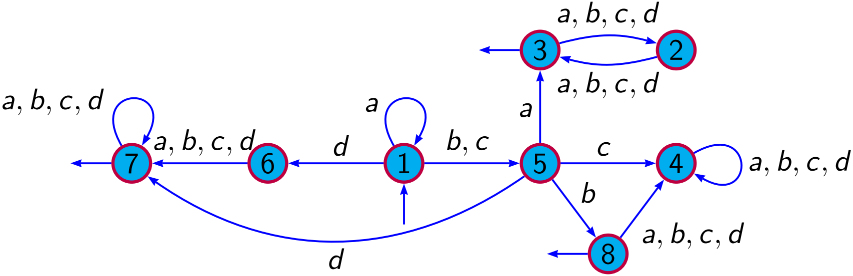
However, if $M = \{ab\}$, then the orbit $\mathcal{O}_{\{p,t\}}(M)$ is infinite, as

$$aba^i \in (pt)^i(M) \quad \text{and} \quad aba^i \notin (pt)^j(M) \text{ for } j < i.$$

.

Prefix and complement

In this case at most 14 distinct languages can be generated.
The bound of 14 can be achieved, e.g., by the regular language over $\Sigma = \{a, b, c, d\}$ accepted by the following DFA:



The following table gives the appropriate set of final states under the operations.

language	final states	language	final states
L	3,7,8	$pcpc(L)$	1,5,6,7
$c(L)$	1,2,4,5,6	$cpcp(L)$	2,3,6,7
$p(L)$	1,2,3,5,6,7,8	$cpcpc(L)$	2,3,4,8
$pc(L)$	1,2,3,4,5,6,8	$pcpcp(L)$	1,2,3,5,6,7
$cp(L)$	4	$pcpcpc(L)$	1,2,3,4,5,8
$cpc(L)$	7	$cpcpcp(L)$	4, 8
$pcp(L)$	1,4,5,8	$cpcpcpc(L)$	6, 7

Kleene star, prefix, suffix, factor

Here there are at most 13 distinct languages, given by the action of

$$\{\epsilon,\ k,\ p,\ s,\ f,\ kp,\ ks,\ kf,\ pk,\ sk,\ fk,\ pks,\ skp\}.$$

The bound of 13 is achieved, for example, by $L = \{abc\}$.

Factor, Kleene star, complement

Here breadth-first search gives 78 languages, so our bound is $78 + 4 = 82$. We can improve this bound by considering new kinds of arguments.

Lemma

For all languages L, we have either $f(L) = \Sigma^*$ or $fc(L) = \Sigma^*$.

Theorem (C-Domaratzki-Harju-Shallit 2011)

Let L be an arbitrary language. Then 50 is a tight upper bound for the size of $\mathcal{O}_{\{k,c,f\}}(L)$.

Sketch of the proof

The languages in $\mathcal{O}_{\{k,c,f\}}(L)$ that may differ from $\Sigma^*, \emptyset, \Sigma^+$, and $\{\epsilon\}$ are among the images of L and $c(L)$ under the 16 operations

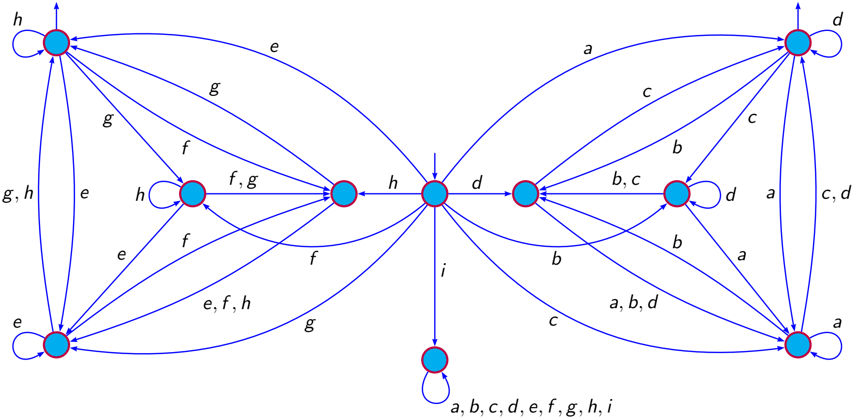
$$\begin{gathered} f\,, kf\,, kckf\,, kcf\,, fk, kcfk, fck, kfck, kckfck, kcfck, \\ fkck, kcfkck, fckck, kfckck, kckfckck, kcfckck, \end{gathered} \tag{1}$$

the complements of these images, together with the 14 languages in $\mathcal{O}_{\{k,c\}}(L)$.

We show that there are at most 32 distinct languages among the $64 = 16 \cdot 4$ languages given by the images of L and $c(L)$ under the 16 operations (1) and their complements.

Adding the 14 languages in $\mathcal{O}_{\{k,c\}}(L)$, and $\Sigma^*, \emptyset, \Sigma^+$, and $\{\epsilon\}$, we obtain that $50 = 32 + 14 + 4$ is an upper bound for the size of the orbit of $\{k, c, f\}$.

Sketch of the proof (cont'd)



This DFA accepts a language L with orbit size 50 under $\{k, c, f\}^*$.

Summary of results

r	**2**	w	**2**	f	**2**
s	**2**	p	**2**	c	**2**
k	**2**	w, r	**4**	f, r	**4**
f, w	**3**	s, w	**3**	s, f	**3**
p, w	**3**	p, f	**3**	c, r	**4**
c, w	**6**$*$	c, f	**6**$*$	c, s	**14**
c, p	**14**	k, r	**4**	k, w	**4**
k, f	**5**	k, s	**5**	k, p	**5**
k, c	**14**	f, w, r	**6**	s, f, w	**4**
p, f, w	**4**	p, s, f	**4**	c, w, r	**10**$*$
c, f, r	**10**$*$	c, f, w	**8**$*$	c, s, w	**16**$*$
c, s, f	**16**$*$	c, p, w	**16**$*$	c, p, f	**16**$*$
k, w, r	**7**	k, f, r	**9**	k, f, w	**6**
k, s, w	**7**	k, s, f	**9**	k, p, w	**7**
k, p, f	**9**	k, c, r	**28**	k, c, w	**38**$*$
k, c, f	**50**$*$	k, c, s	1070	k, c, p	1070

Summary of results (Cont'd)

p, s, f, r	**8**	p, s, f, w	**5**	c, f, w, r	**12**$*$
c, s, f, w	**16**$*$	c, p, f, w	**16**$*$	c, p, s, f	**16**$*$
k, f, w, r	**11**	k, s, f, w	**10**	k, p, f, w	**10**
k, p, s, f	**13**	k, c, w, r	72$*$	k, c, f, r	**84**$*$
k, c, f, w	66$*$	k, c, s, w	1114	k, c, s, f	1450
k, c, p, w	1114	k, c, p, f	1450	p, s, f, w, r	**10**
c, p, s, f, r	**30**$*$	c, p, s, f, w	**16**$*$	k, p, s, f, r	**25**
k, p, s, f, w	**14**	k, c, f, w, r	120$*$	k, c, s, f, w	1474
k, c, p, f, w	1474	k, c, p, s, f	2818	c, p, s, f, w, r	**30**$*$
k, p, s, f, w, r	**27**	k, c, p, s, f, r	5628	k, c, p, s, f, w	2842
k, c, p, s, f, w, r	5676				

Further work

We plan to continue to refine our estimates of the previous tables, and pursue the status of other sets of operations.

For example, if t is the exponentiation operation, then, using the identities $kt \equiv tk \equiv k$, and the inclusion $t \subseteq k$, we get the additional Kuratowski-style identities

- $kctckck \equiv kck$,
- $kckctck \equiv kck$,
- $kctctck \equiv kck$,
- $tctctck \equiv tck$,
- $kctctct \equiv kct$.

This allows us to prove that $\mathcal{O}_{\{k,c,t\}}(L)$ is finite and of cardinality at most 126.