Derivatives of Regular Expressions and an Application

Haiming Chen[1] and Yu Shen[2]

[1]State Key Laboratory of Computer Science, ISCAS

[2]Department of Computer Science, University of Western Ontario

WTCS 2012, Auckland 2012.2.22

Derivatives $w^{-1}(E)$ Brzozowski, 1964

➔ (left) quotient of language $w^{-1}L(E)$

Berry and Sethi 's Result 1986

Derivatives of E → classes of *similar* derivatives

Why? *E* linear

Our work

A characterization of the structure of derivatives of linear E *implies Berry and Sethi 's Result*

Derivatives

Berry and Sethi's result

Structure of derivatives

Properties of repeating terms

An application

Conclusion

Regular expressions

$$E ::= \phi \mid \varepsilon \mid a \in \Sigma \mid E + E \mid EE \mid E^*$$

ACI-similar

$$E_1 \sim_{\text{aci}} E_2$$

Associativity $(E_1 + E_2) + E_3 = E_1 + (E_2 + E_3)$

Commutativity $E_1 + E_2 = E_2 + E_1$

Idempotence $E + E = E$

Marked expressions

$(a+b)^*ab(a+b)$ E $(a_1+b_1)^*a_2b_2(a_3+b_3)$ $\overline{E}$

The same notation used for dropping of subscripts:

$\overline{\overline{E}} = E$

Note

Marking is not unique

For example $(a_1+b_2)^*a_3b_4(a_5+b_6)$

(left) quotient set of a language L

$$w^{-1}(L)=\{u \mid wu\in L\}$$

$$L = wL(w^{-1}(L))$$

Derivatives (Brzozowski)

$$\begin{aligned}
a^{-1}(\emptyset) &= a^{-1}(\varepsilon) = \emptyset \\
a^{-1}(b) &= \begin{cases} \varepsilon, & \text{if } b = a \\ \emptyset, & \text{otherwise} \end{cases} \\
a^{-1}(F+G) &= a^{-1}(F) + a^{-1}(G) \\
a^{-1}(FG) &= \begin{cases} a^{-1}(F)G + a^{-1}(G), & \text{if } \varepsilon \in L(F) \\ a^{-1}(F)G, & \text{otherwise} \end{cases} \\
a^{-1}(F^*) &= a^{-1}(F)F^*
\end{aligned}$$

$$\varepsilon^{-1}(E) = E,\ (wa)^{-1}(E) = a^{-1}(w^{-1}(E))$$

$$L(w^{-1}(E)) = w^{-1}(L(E))$$

Derivatives

Berry and Sethi’s result

Structure of derivatives

Properties of repeating terms

An application

Conclusion

Regular expressions with distinct symbols (linear):

One symbol occurs only once

Next we consider this kind of expressions

Derivatives

Berry and Sethi’s result

Structure of derivatives

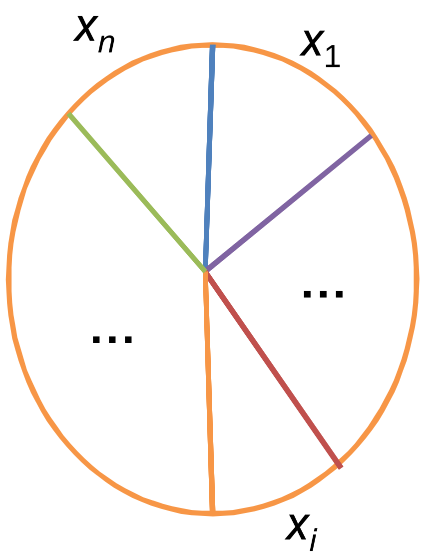
Properties of repeating terms

An application

Conclusion

Berry and Sethi proved that

Let all symbols in E be distinct. Given a fixed $x \in \Sigma_E$, $(wx)^{-1}(E)$ is either $\emptyset$ or unique modulo $\sim_{aci}$ for all words w.



Non-null derivatives of *E*

$|\Sigma_E| = n$

Theorem 1 Let all symbols in E be distinct. Given a fixed $x \in \Sigma_E$, for all words w, each non-null $(wx)^{-1}(E)$ must be of one of the following forms: F or $F + \ldots + F$, where F is a non-null regular expression called the repeating term of $(wx)^{-1}(E)$ which does not contain + at the top level.

Example 1 *Let* $E = (a + b)(a^* + ba^* + b^*)^*$, *then*

$$\overline{E} = (a_1 + b_2)(a_3^* + b_4 a_5^* + b_6^*)^*,$$
$$a_1^{-1}(\overline{E}) = (a_3^* + b_4 a_5^* + b_6^*)^* = \tau_1,$$
$$(a_1 a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_1) = a_3^* \tau_1 = \tau_2,$$
$$(a_1 a_3 a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_2) = \tau_2 + \tau_2,$$
$$\ldots$$

repeating term for
$(wa_1)^{-1}(E)$: τ_1
$(wa_3)^{-1}(E)$: τ_2

Denote by $rt_x(E)$ the repeating term of $(wx)^{-1}(E)$

Corollary 1 Let all symbols in E be distinct. If $(wx)^{-1}(E)$ is non-null, then $(wx)^{-1}(E) \sim_{aci} rt_x(E)$.

a more precise version of Berry and Sethi's result

Q: For each $x \in \Sigma_E$, whether there is a non-null $(wx)^{-1}(E)$ containing one $rt_x(E)$, that is, $rt_x(E)$ is a derivative of E.

A: positive see below

The first appearance $F_x(E)$

$ind : \Sigma_E \to \{1, \ldots, \|E\|\}$: $ind(x) = d$ if x is the dth occurrence of symbols from left to right in E

$x < y$ iff $ind(x) < ind(y)$ $x, y \in \Sigma_E$

$w_1 \prec w_2$ if either $|w_1| < |w_2|$, or $|w_1| = |w_2|$ and

let $w_1 = x_1 \ldots x_n, w_2 = x'_1 \ldots x'_n, 1 \leq k \leq n$,

$x_t = x'_t$ for $t = 1, \ldots, k-1$, and $x_k < x'_k$

A non-null $(wx)^{-1}(E)$ is called the *first appearance* of derivative of E w.r.t. x if for any other non-nul $(w_1x)^{-1}(E)$ it has $w \prec w_1$

Example 2 *For* $E = (a+b)(a^* + ba^* + b^*)^*$, $\overline{E} = (a_1 + b_2)(a_3^* + b_4 a_5^* + b_6^*)^*$

$$\underline{a_1}^{-1}(\overline{E}) = (a_3^* + b_4 a_5^* + b_6^*)^* = \tau_1, \qquad \underline{b_2}^{-1}(\overline{E}) = (a_3^* + b_4 a_5^* + b_6^*)^* = \tau_1,$$
$$(a_1\underline{a_3})^{-1}(\overline{E}) = a_3^{-1}(\tau_1) = a_3^*\tau_1 = \tau_2, \qquad (a_1\underline{b_4})^{-1}(\overline{E}) = b_4^{-1}(\tau_1) = a_5^*\tau_1 = \tau_3,$$
$$(a_1\underline{b_6})^{-1}(\overline{E}) = b_6^{-1}(\tau_1) = b_6^*\tau_1 = \tau_4, \qquad (b_2a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_1) = \tau_2,$$
$$(b_2b_4)^{-1}(\overline{E}) = b_4^{-1}(\tau_1) = \tau_3, \qquad (b_2b_6)^{-1}(\overline{E}) = b_6^{-1}(\tau_1) = \tau_4,$$
$$(a_1a_3a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_2) = \tau_2 + \tau_2, \qquad (a_1a_3b_4)^{-1}(\overline{E}) = b_4^{-1}(\tau_2) = \tau_3,$$
$$(a_1a_3b_6)^{-1}(\overline{E}) = b_6^{-1}(\tau_2) = \tau_4, \qquad (a_1b_4a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_3) = \tau_2,$$
$$(a_1b_4b_4)^{-1}(\overline{E}) = b_4^{-1}(\tau_3) = \tau_3, \qquad (a_1b_4\underline{a_5})^{-1}(\overline{E}) = a_5^{-1}(\tau_3) = \tau_3.$$

Proposition 1 Let all symbols in E be distinct. Given a fixed $x \in \Sigma_E$, the first appearance $F_x(E)$ consists of only one repeating term.

The choice of the order is not significant.

Proposition 2 Let all symbols in E be distinct. Given any words $w_1, w_2 \in \Sigma_E^*$ and $x \in \Sigma_E$, if $| w_1 | = | w_2 |$ and $(w_1 x)^{-1}(E), (w_2 x)^{-1}(E) \neq \phi$, and there is no w, such that $|w| < | w_1 |$ and), $(w x)^{-1}(E) \neq \phi$, then $(w_1 x)^{-1}(E) = (w_2 x)^{-1}(E)$.

Proposition 3 Let all symbols in E be distinct. There exists a word $w \in \Sigma_E^*$ for each $x \in \Sigma_E$, such that $(w x)^{-1}(E) = rt_x(E)$.

Thus repeating terms are derivatives of E, and any non-null derivative of E is built from one of them.

Derivatives

Berry and Sethi’s result

Structure of derivatives

Properties of repeating terms

An application

Conclusion

Proposition 7 *Let all symbols in E be distinct. For each $x \in \Sigma_E$,*

(1) $rt_x(E)$ exists, and $rt_x(E) \neq \emptyset$.

(2) $rt_x(E)$ is unique.

There cannot be two $rt_x(E)$ for $(w\,x)^{-1}(E)$

Example 3 *For* $E = (a+b)(a^* + ba^* + b^*)^*$, $\overline{E} = (a_1 + b_2)(a_3^* + b_4 a_5^* + b_6^*)^*$.

$$
\begin{aligned}
rt_{a_1}(\overline{E}) &= rt_{a_1}(a_1 + b_2)(a_3^* + b_4 a_5^* + b_6^*)^* = rt_{a_1}(a_1)(a_3^* + b_4 a_5^* + b_6^*)^* \\
&= \varepsilon(a_3^* + b_4 a_5^* + b_6^*)^* = (a_3^* + b_4 a_5^* + b_6^*)^* = \tau_1, \\
rt_{b_2}(\overline{E}) &= \varepsilon(a_3^* + b_4 a_5^* + b_6^*)^* = \tau_1, \\
rt_{a_3}(\overline{E}) &= rt_{a_3}(a_3^* + b_4 a_5^* + b_6^*)^* = rt_{a_3}(a_3^* + b_4 a_5^* + b_6^*)\tau_1 = rt_{a_3}(a_3^*)\tau_1 \\
&= rt_{a_3}(a_3)a_3^*\tau_1 = a_3^*\tau_1 = \tau_2, \\
rt_{b_4}(\overline{E}) &= rt_{b_4}(a_3^* + b_4 a_5^* + b_6^*)^* = rt_{b_4}(b_4 a_5^*)\tau_1 = a_5^*\tau_1 = \tau_3, \\
rt_{a_5}(\overline{E}) &= rt_{a_5}(a_3^* + b_4 a_5^* + b_6^*)^* = rt_{a_5}(b_4 a_5^*)\tau_1 = a_5^*\tau_1 = \tau_3, \text{ and} \\
rt_{b_6}(\overline{E}) &= rt_{b_6}(a_3^* + b_4 a_5^* + b_6^*)^* = rt_{b_6}(b_6^*)\tau_1 = b_6^*\tau_1 = \tau_4.
\end{aligned}
$$

Proposition 8 Let all symbols in E be distinct. If there are non-null $(w_1 x_1)^{-1}(E)$ and $(w_2 x_2)^{-1}(E)$, such that $(w_1 x_1)^{-1}(E) \sim_{aci} (w_2 x_2)^{-1}(E)$, then $rt_{x1}(E) = rt_{x2}(E)$, and vice versa.

Corollary 2 Let all symbols in E be distinct. If $rt_{x1}(E) \sim_{aci} rt_{x2}(E)$, then $rt_{x1}(E)=rt_{x2}(E)$.

Remark $rt_x(E)$'s are `atomic' building blocks

(1) Each non-null $(w\,x)^{-1}(E)$ is uniquely decomposed into a sum of $rt_x(E)$, that is, $(w\,x)^{-1}(E) = \sum rt_x(E)$.

(2) $rt_x(E)$ and $rt_y(E)$ are either identical, or not equivalent modulo $\sim_{aci}$, if $x \neq y$.

Derivatives

Berry and Sethi's result

Structure of derivatives

Properties of repeating terms

An application

Conclusion

Solves an issue in using Berry and Sethi's result: find a unique representative for $(w\,x)^{-1}(E)$

Glushkov automaton

$$M_{\text{pos}}(E) = (Q_{\text{pos}}, \Sigma, \delta_{\text{pos}}, q_E, F_{\text{pos}}),$$

where

1. $Q_{\text{pos}} = \Sigma_{\overline{E}} \cup \{q_E\}$, q_E *is a new state not in* $\Sigma_{\overline{E}}$
2. $\delta_{\text{pos}}(q_E, a) = \{x \mid x \in \mathit{first}(\overline{E}), \overline{x} = a\}$ *for* $a \in \Sigma$
3. $\delta_{\text{pos}}(x, a) = \{y \mid y \in \mathit{follow}(\overline{E}, x), \overline{y} = a\}$ *for* $x \in \Sigma_{\overline{E}}$ *and* $a \in \Sigma$
4. $F_{\text{pos}} = \begin{cases} \mathit{last}(\overline{E}) \cup \{q_E\}, & \text{if } \varepsilon \in L(E), \\ \mathit{last}(\overline{E}), & \text{otherwise} \end{cases}$

Berry and Sethi showed the class of derivatives $\{(wx)^{-1}(E)\}$ corresponds to a state x of $M_{\text{pos}}(E)$, $x \in \Sigma_{\overline{E}}$

In many cases, however, one needs a unique representative for the class of $\{(w\,x)^{-1}(E)\}$ to correspond to a state x

By the work, the representatives are obtained immediately

An improvement of Ilie and Yu‘s proof presented in (Ilie & Yu 2003)

A proof about the quotient relation between Glushkov and partial derivative automata

Partial derivatives

$$\partial_a(\emptyset) = \partial_a(\varepsilon) = \emptyset$$

$$\partial_a(b) = \begin{cases} \{\varepsilon\}, \text{if } b = a \\ \emptyset, \quad \text{otherwise} \end{cases}$$

$$\partial_a(F + G) = \partial_a(F) \cup \partial_a(G)$$

$$\partial_a(FG) = \begin{cases} \partial_a(F)G \cup \partial_a(G), \text{ if } \varepsilon \in L(F) \\ \partial_a(F)G, \qquad\qquad \text{otherwise} \end{cases}$$

$$\partial_a(F^*) = \partial_a(F)F^*$$

$$\partial_\varepsilon(E) = \{E\},\ \partial_{wa}(E) = \bigcup_{p \in \partial_w(E)} \partial_a(p)$$

$$PD(E) = \cup_{w \in \Sigma^*} \partial_w(E)$$

Partial derivative automaton

$$M_{\mathrm{pd}}(E) = (PD(E), \Sigma, \delta_{\mathrm{pd}}, E, \{q \in PD(E) \mid \varepsilon \in L(q)\}),$$

where $\delta_{\mathrm{pd}}(q, a) = \partial_a(q)$, *for any* $q \in PD(E), a \in \Sigma$.

$M_{\mathrm{pd}}(E)$ is a quotient of $M_{\mathrm{g}}(E)$

Ilie and Yu's proof

- The central issue is to find a unique representative for a class of derivatives
- The proof fails to find the correct representatives

It is claimed in the proof that, by using the rules (ϕε-rules), for a fixed $x \in \Sigma_{\overline{E}}$ and for all words *w*, $(wx)^{-1}(\overline{E})$ is either ϕ or unique.
incorrect

$$E + \emptyset = \emptyset + E = E$$
$$E\emptyset = \emptyset E = \emptyset$$
$$E\varepsilon = \varepsilon E = E$$

Rules (ϕ-rules)

Example. In Example 1, $(a_1a_3)^{-1}(\overline{E})$ and $(a_1a_3a_3)^{-1}(\overline{E})$ are distinct

An improved proof

Use $rt_x(\overline{E})$ as the unique representative.

See our paper

Derivatives

Berry and Sethi’s result

Structure of derivatives

Properties of repeating terms

An application

Conclusion

A characterization of the structure of derivatives

Several properties

An application

A useful technique

Thanks!