Derivatives of Regular Expressions [*]

Haiming Chen
State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences
Beijing 100190, China
chm@ios.ac.cn
Telephone: +86-10-62661612
Fax: +86-10-62661627

2009-10

Abstract

The paper proposes a characterization of the structure of derivatives, and proves several properties of derivatives. The above work can be used to solve an issue in using Berry and Sethi's result, i. e., finding the unique representatives of derivatives.

keywords: Regular expressions, derivatives, finite automata.

1 Introduction

The construction of finite automata from regular expressions is an important issue and has been studied for a long time. An elegant construction of deterministic finite automata, based on the derivatives of regular expressions, was proposed by Brzozowski [4]. Among the well-known constructions of ϵ-free non-deterministic finite automata (NFA), the position automaton was proposed separately by Glushkov [7] and McNaughton and Yamada [9]. Berry and Sethi [2] showed that the position automaton has a natural connection with the notion of derivative [4], and related the above two approaches.

The paper continues the investigation of derivatives along the line of Berry and Sethi. It gives a characterization of the structure of derivatives of an expression E with distinct symbols, showing that each non-null derivative of E is composed of one or more identical expressions (called repeating terms), which implies Berry and Sethi's result [2]. The paper proves several facts, including computation of repeating terms, and several properties of repeating terms.

The above theoretical work solves an issue in using Berry and Sethi's result. Berry and Sethi showed that an arbitrary derivative in a certain class of derivatives of an expression E with distinct symbols corresponds to a state of the position automaton of E. This means that the derivatives corresponding to a state are not unique. In many cases, however, one needs a unique representative for that class of derivatives to correspond to a state. This, however, turns out to be not trivial as is discussed in Section 4. By the work on derivatives in the paper, the representatives are obtained immediately.

Section 2 introduces notations and notions required in the paper. Section 3 proposes a characterization of derivatives and several properties of derivatives. Section 4 contains concluding remarks.

[*] Work supported by the National Natural Science Foundation of China under Grants 60573013, 60721061.

2 Preliminaries

We assume the reader to be familiar with basic regular language and automata theory, e.g., from [11], so that we introduce here only some notations and notions used later in the paper.

2.1 Regular expressions and finite automata

Let Σ be an alphabet of symbols. The set of all words over Σ is denoted by Σ^*. The empty word is denoted ε. A regular expression over Σ is $\emptyset, \varepsilon$ or $a \in \Sigma$, or is obtained from these by applying the following rules finitely many times: the union $E_1 + E_2$, the concatenation E_1E_2, and the star E_1^* for two regular expressions E_1 and E_2. For a regular expression E, the language specified by E is denoted by $L(E)$. The number of symbol occurrences in E, or the alphabetic width of E, is denoted $\|E\|$. The symbols that occur in E, which is the smallest alphabet of E, is denoted by Σ_E.

Two regular expressions E_1 and E_2 which reduce to the same expression using associativity, commutativity, and idempotence of $+$ are called *similar* [4], which is denoted $E_1 \sim_{aci} E_2$.

We assume that the rules $E + \emptyset = \emptyset + E = E, E\emptyset = \emptyset E = \emptyset$, and $E\varepsilon = \varepsilon E = E$ ($\emptyset\varepsilon$-rules) hold in the paper.

For a regular expression E over Σ, we define the following functions:

$$\begin{aligned}
&first(E) = \{a \mid aw \in L(E), a \in \Sigma, w \in \Sigma^*\}\\
&last(E) = \{a \mid wa \in L(E), w \in \Sigma^*, a \in \Sigma\}\\
&follow(E, a) = \{b \mid uabv \in L(E), u, v \in \Sigma^*, b \in \Sigma\},\ \text{for } a \in \Sigma
\end{aligned}$$

One can easily write equivalent inductive definitions of the above functions on E, which is omitted here.

For a regular expression we can mark symbols with subscripts so that in the marked expression each marked symbol occurs only once. For example $(a_1 + b_1)^*a_2b_2(a_3 + b_3)$ is a marking of the expression $(a + b)^*ab(a + b)$. A marking of an expression E is denoted by $\overline{E}$. The same notation will also be used for dropping of subscripts from the marked symbols: $\overline{\overline{E}} = E$. We extend the notation for words and automata in the obvious way. It will be clear from the context whether $\bar{\cdot}$ adds or drops subscripts.

In this way the subscribed symbols are called *positions* of the expression. In the literature, positions are usually defined as the subscripts. This definition of positions, however, has drawbacks because it separate subscripts from symbols. When both subscripts and related symbols are required, presentations are awkward. For example, the definitions in [8] for the $first, last$, and $follow$ functions are actually not rigorous. On the other hand, rigorous definitions will be very tedious in this manner. Here we use symbols in $\Sigma_{\overline{E}}$ as the positions, which makes related definitions concise, and is more flexible (subscripts can be same, as in the above example).

A finite automaton is a quintuple $M = (Q, \Sigma, \delta, q_0, F)$, where Q is a finite set of states, Σ is the alphabet, $\delta \subseteq Q \times \Sigma \times Q$ is the transition mapping, q_0 is the start state, and $F \subseteq Q$ is the set of accepting states. Denote the language accepted by the automaton M by $L(M)$.

Let $\equiv \subseteq Q \times Q$ be an equivalence relation. For $q \in Q$, $[q]_\equiv$ denotes the equivalence class of q w.r.t. $\equiv$ and $Q/_\equiv$ denotes the quotient set $Q/_\equiv = \{[q]_\equiv \mid q \in Q\}$. We say that $\equiv$ is right invariant w.r.t. M iff (1) $\equiv \subseteq (Q - F)^2 \cup F^2$ and (2) for any $p, q \in Q, a \in \Sigma$, if $p \equiv q$, then $\delta(p, a)/_\equiv = \delta(q, a)/_\equiv$. If $\equiv$ is right invariant, the quotient automaton $M/_\equiv$ is $M/_\equiv = (Q/_\equiv, \Sigma, \delta_\equiv, [q_0]_\equiv, F/_\equiv)$, where $\delta_\equiv = \{([p]_\equiv, a, [q]_\equiv) \mid (p, a, q) \in \delta\}$. One can prove that $L(M/_\equiv) = L(M)$.

2.2 Derivatives

Given a language L and a finite word w, the derivative (or left quotient set) of L w. r. t. w is $w^{-1}(L) = \{u \mid wu \in L\}$. It has $L = wL(w^{-1}(L))$.

Derivatives of regular expressions were introduced by Brzozowski [4].

Definition 1 (Brzozowski [4]) *Given a regular expression E and a symbol a, the derivative $a^{-1}(E)$ of E with respect to a is defined inductively as follows:*

$$\begin{array}{rcl} a^{-1}(\emptyset) & = & a^{-1}(\varepsilon) = \emptyset \\ a^{-1}(b) & = & \left\{ \begin{array}{ll} \varepsilon, & \text{if } b = a \\ \emptyset, & \text{otherwise} \end{array} \right. \\ a^{-1}(F+G) & = & a^{-1}(F) + a^{-1}(G) \\ a^{-1}(FG) & = & \left\{ \begin{array}{ll} a^{-1}(F)G + a^{-1}(G), & \text{if } \varepsilon \in L(F) \\ a^{-1}(F)G, & \text{otherwise} \end{array} \right. \\ a^{-1}(F^*) & = & a^{-1}(F)F^* \end{array}$$

Derivative with respective to a word is computed by $\varepsilon^{-1}(E) = E$, $(wa)^{-1}(E) = a^{-1}(w^{-1}(E))$. It is known that $L(w^{-1}(E)) = w^{-1}(L(E))$. Brzozowski showed that an expression E has a finite number of dissimiliar derivatives [4], which were used as states to construct a deterministic finite automaton of E.

Partial derivatives were introduced by Antimirov [1].

Definition 2 (Antimirov [1]) *Given a regular expression E and a symbol a, the set of partial derivatives $\partial_a(E)$ of E with respect to a is defined as follows:*

$$\begin{array}{rcl} \partial_a(\emptyset) & = & \partial_a(\varepsilon) = \emptyset \\ \partial_a(b) & = & \left\{ \begin{array}{ll} \{\varepsilon\}, & \text{if } b = a \\ \emptyset, & \text{otherwise} \end{array} \right. \\ \partial_a(F+G) & = & \partial_a(F) \cup \partial_a(G) \\ \partial_a(FG) & = & \left\{ \begin{array}{ll} \partial_a(F)G \cup \partial_a(G), & \text{if } \varepsilon \in L(F) \\ \partial_a(F)G, & \text{otherwise} \end{array} \right. \\ \partial_a(F^*) & = & \partial_a(F)F^* \end{array}$$

Partial derivative with respect to a word is computed by $\partial_\varepsilon(E) = \{E\}$, $\partial_{wa}(E) = \bigcup_{p \in \partial_w(E)} \partial_a(p)$. The language denoted by $\partial_w(E)$ is $L(\partial_w(E)) = \bigcup_{p \in \partial_w(E)} L(p)$[1].

It is proved in [1] that the cardinality of the set $PD(E) = \cup_{w \in \Sigma^*} \partial_w(E)$ of all partial derivatives of a regular expression E is less than or equal to $\|E\| + 1$.

2.3 Position and equation automata

The position automaton was introduced independently by Glushkov [7] and McNaughton and Yamada [9].

Definition 3 *The position automaton of E is*

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

where

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

As shown by Glushkov [7], McNaughton and Yamada [9], $L(M_{\text{pos}}(E)) = L(E)$. $M_{\text{pos}}(E)$ can be computed in quadratic time [3, 6, 10].

The equation automaton [1] is constructed by partial derivatives.

[1] In the definition $RF = \{EF | E \in R\}$ for a set R of regular expressions and a regular expression F.

Definition 4 *The equation automaton of a regular expression E is*

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

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

It is proved [**?**] that $M_{\mathrm{pd}}(E)$ is a quotient of $M_{\mathrm{pos}}(E)$.

3 Regular expressions with distinct symbols

From Brzozowski [4] and Berry and Sethi [2] the following two facts are easily derived.

Proposition 1 *Let all symbols in E be distinct. Given $x \in \Sigma_E$, for all words w,*
1. If $E = E_1 + E_2$, then

$$(wx)^{-1}(E_1 + E_2) = \begin{cases} (wx)^{-1}(E_1) & \text{if } x \in \Sigma_{E_1}, w \in \Sigma_{E_1}^* \\ (wx)^{-1}(E_2) & \text{if } x \in \Sigma_{E_2}, w \in \Sigma_{E_2}^* \\ \emptyset & \text{otherwise} \end{cases} \quad (1)$$

2. If $E = E_1E_2$, then

$$(wx)^{-1}(E_1E_2) = \begin{cases} (wx)^{-1}(E_1)E_2 & \text{if } x \in \Sigma_{E_1}, w \in \Sigma_{E_1}^* \\ (vx)^{-1}(E_2) & \text{if } w = uv, \varepsilon \in L(u^{-1}(E_1)), x \in \Sigma_{E_2}, \\ & u \in \Sigma_{E_1}^*, v \in \Sigma_{E_2}^* \\ \emptyset & \text{otherwise} \end{cases} \quad (2)$$

Proof. 1. It is directly from Berry and Sethi [2].
2. From Berry and Sethi [2] it is already known

$$(wx)^{-1}(E_1E_2) = \begin{cases} (wx)^{-1}(E_1)E_2 & \text{if } x \in \Sigma_{E_1}, w \in \Sigma_{E_1}^* \text{ (a)} \\ \Sigma_{w=uv, \varepsilon \in L(u^{-1}(E_1))}(vx)^{-1}(E_2) & \text{otherwise (b)} \end{cases}$$

Let us consider (b) and set $wx = a_1a_2 \ldots a_t$. For a concrete sequence of $a_1 \ldots a_t$, a subterm $(a_{r+1} \ldots a_t)^{-1}(E_2)$ in (b) can exist only if $a_1, \ldots, a_r \in E_1$ and $a_{r+1}, \ldots, a_t \in E_2$. Since $a_n, 1 \leq n \leq t$ is either in E_1 or in E_2, there is at most one such subterm in (b). If such condition is not satisfied, then $(wx)^{-1}(E_1E_2) = \emptyset$. □

Proposition 2 *Given $x \in \Sigma_E$, for all words w, $(wx)^{-1}(E^*)$ is equivalent to a sum of subterms chosen from the set $\{(vx)^{-1}(E)E^* \mid wx = uvx\}$.*

Proof. It is directly from Brzozowski [4] or Berry and Sethi [2]. □

Berry and Sethi [2] proved that

Proposition 3 (Berry and Sethi [2]) *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.*

This is a very important property which was used to connect the class of non-null $(wx)^{-1}(\overline{E})$ to the state x of $M_{\mathrm{pos}}(E)$ for an expression E.

We further investigate the structure of non-null $(wx)^{-1}(E)$ here.

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.*

Proof. We prove it by induction on the structure of E. If $E = \emptyset$ or ε, then no symbol is in E, and no non-null derivative exists. Thus no repeating term exists. If $E = a, a \in \Sigma_E$, then the only symbol in a is a, and $a^{-1}(a) = \varepsilon$, $(wx)^{-1}(E) = \emptyset$ for $w \neq \varepsilon$ or $x \neq a$. Thus ε is the repeating term of $a^{-1}(a)$, in which no $+$ appears.

1. $E = E_1 + E_2$. By equation (1), a non-null $(wx)^{-1}(E)$ is either $(wx)^{-1}(E_1)$ or $(wx)^{-1}(E_2)$. Suppose the first, then $(wx)^{-1}(E_1)$ is non-null, and the repeating term of $(wx)^{-1}(E)$ is the same as $(wx)^{-1}(E_1)$. The inductive hypothesis applies to it, and no top-level $+$ will be added. The same is for the second.

2. $E = E_1E_2$. By equation (2), a non-null $(wx)^{-1}(E)$ is either $(wx)^{-1}(E_1)E_2$ or $(vx)^{-1}(E_2)$ for some v such that $w = uv$. If $(wx)^{-1}(E) = (wx)^{-1}(E_1)E_2$, by the inductive hypothesis, $(wx)^{-1}(E_1)$ is F or $F + \ldots F$ where F does not contain $+$ at the top level. Then FE_2 is the repeating term of $(wx)^{-1}(E)$, which does not contain top-level $+$. If $(wx)^{-1}(E) = (wx)^{-1}(E_2)$, the proof is the same as in the above case 1.

3. $E = E_1^*$. From Proposition 2 it is known that $(wx)^{-1}(E)$ is the sum of subterms of the form $(vx)^{-1}(E_1)E_1^*$ where $wx = uvx$. From the inductive hypothesis, each non-null $(vx)^{-1}(E_1)$ is F or $F + \ldots + F$ where F does not contain $+$ at the top level, so $(vx)^{-1}(E_1)E_1^*$ is FE_1^* or $FE_1^* + \ldots + FE_1^*$. If $(wx)^{-1}(E)$ is non-null, it is a sum of one or more FE_1^*, which does not contain $+$ at the top level. □

Therefore each $(wx)^{-1}(E)$ is either $\emptyset$ or a sum of one or more repeating terms of $(wx)^{-1}(E)$.

Example 1 *Let* $E = (a + b)(a^* + ba^* + b^*)^*$, *then*
$\overline{E} = (a_1 + b_2)(a_3^* + b_4a_5^* + b_6^*)^*$,
$a_1^{-1}(\overline{E}) = (a_3^* + b_4a_5^* + b_6^*)^* = \tau_1$,
$(a_1a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_1) = a_3^*\tau_1 = \tau_2$,
$(a_1a_3a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_2) = \tau_2 + \tau_2$,
$\ldots$
The repeating term for $(wa_1)^{-1}(\overline{E})$ *is* τ_1, *the repeating term for* $(wa_3)^{-1}(\overline{E})$ *is* τ_2.

Denote by $rt_x(E)$ the repeating term of $(wx)^{-1}(E)$. From Theorem 1 we have

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

Corollary 1 is a more precise version of Berry and Sethi's result (i. e., Propostion 3), that is, Theorem 1 implies Berry and Sethi's result, but not vice versa.

Below we consider the question: 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. The answer is positive. We show it by a construction, the first appearance.

Let all symbols in E be distinct. We associate symbols in Σ_E with an order. This is achieved by setting up a one-to-one function $ind : \Sigma_E \to \{1, \ldots, \|E\|\}$: $ind(x) = d$ if x is the dth occurrence of symbols from left to right in E. For $x, y \in \Sigma_E$, define $x < y$ iff $ind(x) < ind(y)$. For any words $w_1, w_2 \in \Sigma_E^*$, define the graded lexicographical order by $w_1 \prec w_2$ if either $|w_1| < |w_2|$, or $|w_1| = |w_2|$ and the condition is satisfied: let $w_1 = x_1 \ldots x_n, w_2 = x_1' \ldots x_n'$, there exists an integer $k, 1 \leq k \leq n$, such that $x_t = x_t'$ 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, denoted $F_x(E)$, if for any other non-null $(w_1x)^{-1}(E)$ it has $w \prec w_1$. From Berry and Sethi [2] a non-null $(wx)^{-1}(E)$ exists for all $x \in \Sigma_E$, which ensures the existence of $F_x(E)$.

Example 2 *For* $E = (a+b)(a^* + ba^* + b^*)^*$, $\overline{E} = (a_1 + b_2)(a_3^* + b_4a_5^* + b_6^*)^*$. *The first appearances of derivatives w.r.t. symbols in* $\overline{E}$, *in which the symbols are underlined, are computed as follows.*

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

From Example 2 we can see that no first appearance has duplicated repeating terms while other derivatives may have. Generally we have

Proposition 4 *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.*

Proof. We prove it by induction on the structure of E. The cases for $E = \varepsilon, \emptyset, x,\ x \in \Sigma_E$ are obvious. Suppose wx is such that $F_x(E)$ is $(wx)^{-1}(E)$.

1. $E = E_1 + E_2$. Consider equation (1). If $(wx)^{-1}(E) = (wx)^{-1}(E_1)$, we show that $F_x(E_1)$ is $(wx)^{-1}(E_1)$. Otherwise there is a word $w_1 \prec w$ such that $(w_1x)^{-1}(E_1) \neq \emptyset$. So $(w_1x)^{-1}(E) \neq \emptyset$, which is a contradiction. Therefore $(wx)^{-1}(E_1)$ is the first appearance and the inductive hypothesis applies to it. The same is for $(wx)^{-1}(E) = (wx)^{-1}(E_2)$.

2. $E = E_1E_2$. Consider equation (2). If $(wx)^{-1}(E) = (wx)^{-1}(E_1)E_2$, similarly as above we can prove that $(wx)^{-1}(E_1)$ is the first appearance, and the inductive hypothesis applies to it. If $(wx)^{-1}(E) = (v_1x)^{-1}(E_2)$ for some v_1 such that $wx = uv_1x$, we show that this subterm is $F_x(E_2)$. Suppose the converse. Then there is a word $v \prec v_1$ such that $(vx)^{-1}(E_2) \neq \emptyset$. So it is easy to see that $(uvx)^{-1}(E) \neq \emptyset$. But $uvx \prec wx$, which is a contradiction. Therefore $(v_1x)^{-1}(E_2)$ is the first appearance and the inductive hypothesis applies to it.

3. $E = E_1^*$. From Proposition 2 $(wx)^{-1}(E)$ is the sum of subterms of the form $(vx)^{-1}(E_1)E_1^*$ where $wx = uvx$. We show that when $(wx)^{-1}(E)$ is $F_x(E)$ the above becomes $(wx)^{-1}(E) = (wx)^{-1}(E_1)E_1^*$. Suppose $(wx)^{-1}(E)$ contains another non-null subterm $(vx)^{-1}(E_1)E_1^*, w = uv, w \neq v$. Then $(vx)^{-1}(E)$ is not $\emptyset$ since it contains $(vx)^{-1}(E_1)E_1^*$ as a summand. However $v \prec w$, which is a contradiction. Similarly we can prove that $(wx)^{-1}(E_1)$ is $F_x(E_1)$, so the inductive hypothesis applies to it. □

The choice of the order is not significant. Actually for different *ind*, the resulting $F_x(E)$ is the same.

Proposition 5 *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 $(w_1x)^{-1}(E), (w_2x)^{-1}(E) \neq \emptyset$, and there is no w, such that $|w| < |w_1|$ and $(wx)^{-1}(E) \neq \emptyset$, then $(w_1x)^{-1}(E) = (w_2x)^{-1}(E)$.*

Proof. We prove it by induction on the structure of E. If $E = \emptyset$ or ε, no non-null derivative exists. If $E = a$ for a symbol a, the only non-null derivative is $a^{-1}(E)$, in which case $w_1 = w_2 = \varepsilon$ and $x = a$. So $(w_1x)^{-1}(E) = (w_2x)^{-1}(E)$.

1. $E = E_1 + E_2$. If $x \in \Sigma_{E_1}$, from equation (1), we have $(w_1x)^{-1}(E) = (w_1x)^{-1}(E_1)$ and $(w_2x)^{-1}(E) = (w_2x)^{-1}(E_1)$. We can see that there is no w, such that $|w| < |w_1|$ and $(wx)^{-1}(E_1) \neq \emptyset$. Otherwise $(wx)^{-1}(E) = (wx)^{-1}(E_1) \neq \emptyset$ which is a contradiction. So the inductive hypothesis applies to E_1. The proof is the same for $x \in \Sigma_{E_2}$.

2. $E = E_1E_2$. If $x \in \Sigma_{E_1}$, from equation (2), we have $(w_1x)^{-1}(E) = (w_1x)^{-1}(E_1)E_2$ and $(w_2x)^{-1}(E) = (w_2x)^{-1}(E_1)E_2$. Similar as in case 1 we can prove $(w_1x)^{-1}(E_1) = (w_2x)^{-1}(E_1)$. Thus $(w_1x)^{-1}(E) = (w_2x)^{-1}(E)$.

If $x \in \Sigma_{E_2}$, from equation (2), we have $(w_1x)^{-1}(E) = (v_1x)^{-1}(E_2)$ and $(w_2x)^{-1}(E) = (v_2x)^{-1}(E_2)$ for some v_1, v_2 such that $w_1 = u_1v_1, w_2 = u_2v_2, \varepsilon \in L(u_1^{-1}(E_1)), \varepsilon \in L(u_2^{-1}(E_1)), u_1, u_2 \in \Sigma_{E_1}^*, v_1, v_2 \in \Sigma_{E_2}^*$. We show $|v_1| = |v_2|$. Suppose the converse. Without losing generality suppose

$|v_1| < |v_2|$. Notice $|w_1| = |w_2|$, then $|u_1| > |u_2|$. Since $\varepsilon \in L(u_2^{-1}(E_1)), u_2 \in \Sigma_{E_1}^*, v_1 \in \Sigma_{E_2}^*$, by equation (2) $(u_2v_1x)^{-1}(E) = (u_2v_1x)^{-1}(E_2) \neq \emptyset$. But $|u_2v_1| < |w_1|$ which is a contradiction. So $|v_1| = |v_2|$ and the inductive hypothesis applies to E_2.

3. $E = E_1^*$. Similar as the proof of case 3 in the proof of Proposition 4, we can prove $(w_1x)^{-1}(E) = (w_1x)^{-1}(E_1)E_1^*$ and $(w_2x)^{-1}(E) = (w_2x)^{-1}(E_1)E_1^*$. Then by the inductive hypothesis $(w_1x)^{-1}(E_1) = (w_2x)^{-1}(E_1)$, thus $(w_1x)^{-1}(E) = (w_2x)^{-1}(E)$. □

In the above proposition it is easy to see that we have $(w_1x)^{-1}(E) = (w_2x)^{-1}(E) = F_x(E)$. Therefore $F_x(E)$ is the same for varying *ind*.

Then

Proposition 6 *Let all symbols in E be distinct. There exists a word $w \in \Sigma_E^*$ for each $x \in \Sigma_E$, such that $(wx)^{-1}(E) = rt_x(E)$.*

Proof. The first appearance $F_x(E)$ is one such $(wx)^{-1}(E)$ satisfying $F_x(E) = rt_x(E)$. □

Thus repeating terms are derivatives of E, and any non-null derivative of E is built from one of them. Next we present other properties for $rt_x(E)$.

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.

Proof. (1) From Berry and Sethi [2] it is known that a non-null $(wx)^{-1}(E)$ exists for each $x \in \Sigma_E$. Then from Theorem 1 $rt_x(E)$ exists and $rt_x(E) \neq \emptyset$.

(2) Suppose $rt_x(E)$ is not unique. That is, for some $x \in \Sigma_E$, there are two repeating terms F and F_1, such that $F \neq F_1$. From Theorem 1 and Proposition 6 it implies $F = F_1 + \ldots + F_1$ and $F_1 = F + \ldots + F$, which is a contradiction. Therefore $rt_x(E)$ is unique. □

If $E = \emptyset$ or ε, no symbol is in E, so $rt_x(E)$ is undefined. We let $rt_x(\emptyset) = rt_x(\varepsilon) = \emptyset$ for any $x \in \Sigma_E$ for the sake of completeness. Then

Proposition 8 *Let all symbols in E be distinct. For each $x \in \Sigma_E$, $rt_x(E)$ can be computed inductively:*

$$\begin{array}{rcl}
rt_x(\emptyset) & = & \emptyset \\
rt_x(\varepsilon) & = & \emptyset \\
rt_a(a) & = & \varepsilon \\
rt_x(F+G) & = & \begin{cases} rt_x(F) & \text{if } x \in \Sigma_F \\ rt_x(G) & \text{if } x \in \Sigma_G \end{cases} \\
rt_x(FG) & = & \begin{cases} rt_x(F)G & \text{if } x \in \Sigma_F \\ rt_x(G) & \text{if } x \in \Sigma_G \end{cases} \\
rt_x(F^*) & = & rt_x(F)F^*
\end{array}$$

Proof. $rt_x(\emptyset)$ and $rt_x(\varepsilon)$ is get directly.

The other cases can be get from the proof of Theorem 1. □

Example 3 *For $E = (a+b)(a^* + ba^* + b^*)^*$, $\overline{E} = (a_1 + b_2)(a_3^* + b_4a_5^* + b_6^*)^*$.*

$rt_{a_1}(\overline{E}) = rt_{a_1}(a_1+b_2)(a_3^* + b_4a_5^* + b_6^*)^* = rt_{a_1}(a_1)(a_3^* + b_4a_5^* + b_6^*)^* = \varepsilon(a_3^* + b_4a_5^* + b_6^*)^* = (a_3^* + b_4a_5^* + b_6^*)^* = \tau_1$,

$rt_{b_2}(\overline{E}) = \varepsilon(a_3^* + b_4a_5^* + b_6^*)^* = \tau_1$,

$rt_{a_3}(\overline{E}) = rt_{a_3}(a_3^*+b_4a_5^*+b_6^*)^* = rt_{a_3}(a_3^*+b_4a_5^*+b_6^*)\tau_1 = rt_{a_3}(a_3^*)\tau_1 = rt_{a_3}(a_3)a_3^*\tau_1 = a_3^*\tau_1 = \tau_2$,

$rt_{b_4}(\overline{E}) = rt_{b_4}(a_3^* + b_4a_5^* + b_6^*)^* = rt_{b_4}(b_4a_5^*)\tau_1 = a_5^*\tau_1 = \tau_3$,

$rt_{a_5}(\overline{E}) = rt_{a_5}(a_3^* + b_4a_5^* + b_6^*)^* = rt_{a_5}(b_4a_5^*)\tau_1 = a_5^*\tau_1 = \tau_3$, and

$rt_{b_6}(\overline{E}) = rt_{b_6}(a_3^* + b_4a_5^* + b_6^*)^* = rt_{b_6}(b_6^*)\tau_1 = b_6^*\tau_1 = \tau_4$.

Now we have two approaches to compute $rt_x(E)$, one is by computing $F_x(E)$, the other is by Proposition 8. Of course the result is the same, but usually computation by Proposition 8 is simpler.

The following lemma will be used in the proof of Proposition 9.

Lemma 1 *Let all symbols in E be distinct. If $(wx)^{-1}(E) \sim_{aci} E$, then $rt_x(E) = E$.*

Proof. We prove by induction on the structure of E. If $E = \emptyset$, then $(wx)^{-1}(E) = \emptyset$. By Proposition 8, $rt_x(E) = \emptyset$. So $rt_x(E) = E$. If $E = \varepsilon$, then $(wx)^{-1}(E) = \emptyset$, $(wx^{-1}(E) \not\sim_{aci} E$. If $E = a$, then $a^{-1}(E) = \varepsilon$, $(wx)^{-1}(E) = \emptyset$ for $w \neq \varepsilon$ or $x \neq a$. So $(wx)^{-1}(E) \not\sim_{aci} E$.

Induction. 1. $E = F+G$. If $F = \emptyset$, then $(wx)^{-1}(E) = (wx)^{-1}(G)$. From $(wx)^{-1}(E) \sim_{aci} E$, we have $(wx)^{-1}(G) \sim_{aci} G$. By the inductive hypothesis $rt_x(G) = G$, thus $rt_x(E) = rt_x(G) = G = E$. Similarly, if $G = \emptyset$, we have $rt_x(E) = E$.

If $F, G \neq \emptyset$, then $(wx)^{-1}(E) \sim_{aci} E \neq \emptyset$. By equation (1), $(wx)^{-1}(E)$ is either $(wx)^{-1}(F)$ or $(wx)^{-1}(G)$. If $(wx)^{-1}(E) = (wx)^{-1}(F)$, then $(wx)^{-1}(F) \sim_{aci} F + G$. Since $(wx)^{-1}(F)$ does not contain symbols in G, we have $G = \emptyset$, which is a contradiction. Similarly, if $(wx)^{-1}(E) = (wx)^{-1}(G)$, we also have a contradiction.

2. $E = FG$. If $F = \emptyset$ or $G = \emptyset$, then $E = \emptyset$, $rt_x(E) = E$. Otherwise $F, G \neq \emptyset$, then since $(wx)^{-1}(E) \sim_{aci} E \neq \emptyset$, by equation (2) $wx^{-1}(E)$ is either $(wx)^{-1}(F)G$ or $(vx)^{-1}(G)$ for some v such that $w = uv$. If $wx^{-1}(E) = (wx)^{-1}(F)G$, then $(wx)^{-1}(F)G \sim_{aci} FG$. So $(wx)^{-1}(F) \sim_{aci} F$. By the inductive hypothesis, we have $rt_x(F) = F$. By equation (2) $wx^{-1}(E) = (wx)^{-1}(F)G$ implies $x \in \Sigma_F$. Hence, from Proposition 8, $rt_x(E) = rt_x(F)G = FG$. If $wx^{-1}(E) = (vx)^{-1}(G)$, then $(vx)^{-1}(G) \sim_{aci} FG$. Since $(vx)^{-1}(G)$ does not contain symbols in F, we have $F = \varepsilon$. Then $(vx)^{-1}(G) \sim_{aci} G$. By the inductive hypothesis, we have $rt_x(G) = G$. By equation (2) $wx^{-1}(E) = (vx)^{-1}(G)$ implies $x \in \Sigma_G$. Hence, from Proposition 8, $rt_x(E) = rt_x(G) = G = E$.

3. $E = F^*$. If $E = \emptyset$, then $rt_x(E) = E$. Otherwise $E \neq \emptyset$, then $(wx)^{-1}(E) \neq \emptyset$. From Proposition 8 we have $rt_x(E) = rt_x(F)F^*$. Thus $(wx)^{-1}(E)$ is a sum of one or more $rt_x(F)F^*$. Since $(wx)^{-1}(E) \sim_{aci} F^*$, we have $rt_x(F) = \varepsilon$. Hence $rt_x(E) = rt_x(F)F^* = F^* = E$. □

Proposition 9 *Let all symbols in E be distinct. If there are non-null $(w_1x_1)^{-1}(E)$ and $(w_2x_2)^{-1}(E)$, such that $(w_1x_1)^{-1}(E) \sim_{aci} (w_2x_2)^{-1}(E)$, then $rt_{x_1}(E) = rt_{x_2}(E)$, and vice versa.*

Proof. ($\Rightarrow$) We prove it by induction on the structure of E. The cases for $E = \varepsilon, \emptyset, x$, $x \in \Sigma_E$ are obvious.

1. $E = F + G$. From equation (1), the non-null $(w_1x_1)^{-1}(E)$ is either $(w_1x_1)^{-1}(F)$ or $(w_1x_1)^{-1}(G)$. Likewise, the non-null $(w_2x_2)^{-1}(E)$ is either $(w_2x_2)^{-1}(F)$ or $(w_2x_2)^{-1}(G)$.

If $(w_1x_1)^{-1}(E) = (w_1x_1)^{-1}(F), (w_2x_2)^{-1}(E) = (w_2x_2)^{-1}(F)$ (a),
then $(w_1x_1)^{-1}(F) \sim_{aci} (w_2x_2)^{-1}(F)$. By the inductive hypothesis, we have $rt_{x_1}(F) = rt_{x_2}(F)$. In addition, (a) implies $x_1, x_2 \in \Sigma_F$. Then from Proposition 8, $rt_{x_1}(E) = rt_{x_1}(F)$, and $rt_{x_2}(E) = rt_{x_2}(F)$. Hence $rt_{x_1}(E) = rt_{x_2}(E)$.

If $w_1x_1)^{-1}(E) = (w_1x_1)^{-1}(F), (w_2x_2)^{-1}(E) = (w_2x_2)^{-1}(G)$ (b),
then $(w_1x_1)^{-1}(F) \sim_{aci} (w_2x_2)^{-1}(G)$. Since symbols in F and G are distinct, we have $(w_1x_1)^{-1}(F) = (w_2x_2)^{-1}(G) = \varepsilon$. Then from Theorem 1 we have $rt_{x_1}(F) = rt_{x_2}(G) = \varepsilon$. In addition, (b) implies $x_1 \in \Sigma_F$ and $x_2 \in \Sigma_G$. Hence from Proposition 8, $rt_{x_1}(E) = rt_{x_1}(F) = rt_{x_2}(G) = rt_{x_2}(E)$.

Proofs for the remaining two cases are similar to the above cases.

2. $E = FG$. From equation (2), the non-null $(w_1x_1)^{-1}(E)$ is either $(w_1x_1)^{-1}(F)G$ or $(v_1x_1)^{-1}(G)$ for some v_1 such that $w_1 = u_1v_1$. Likewise, the non-null $(w_2x_2)^{-1}(E)$ is either $(w_2x_2)^{-1}(F)G$ or $(v_2x_2)^{-1}(G)$.

If $(w_1x_1)^{-1}(E) = (w_1x_1)^{-1}(F)G, (w_2x_2)^{-1}(E) = (w_2x_2)^{-1}(F)G$ (a),
then $(w_1x_1)^{-1}(F)G \sim_{aci} (w_2x_2)^{-1}(F)G$, which implies $(w_1x_1)^{-1}(F) \sim_{aci} (w_2x_2)^{-1}(F)$. By the inductive hypothesis, we have $rt_{x_1}(F) = rt_{x_2}(F)$. In addition, (a) implies $x_1, x_2 \in \Sigma_F$. Then from Proposition 8, $rt_{x_1}(E) = rt_{x_1}(F)G$, and $rt_{x_2}(E) = rt_{x_2}(F)G$. Hence $rt_{x_1}(E) = rt_{x_2}(E)$.

If $(w_1x_1)^{-1}(E) = (w_1x_1)^{-1}(F)G, (w_2x_2)^{-1}(E) = (v_2x_2)^{-1}(G)$ (b), then $(w_1x_1)^{-1}(F)G \sim_{aci} (v_2x_2)^{-1}(G)$. Since $(v_2x_2)^{-1}(G)$ does not contain symbols in F, we have $(w_1x_1)^{-1}(F) = \varepsilon$, and $G \sim_{aci} (v_2x_2)^{-1}(G)$. Since $(w_1x_1)^{-1}(F) = \varepsilon$, from Theorem 1 we have $rt_{x_1}(F) = \varepsilon$. By Lemma 1 $G \sim_{aci} (v_2x_2)^{-1}(G)$ implies $rt_{x_2}(G) = G$. In addition, (b) implies $x_1 \in \Sigma_F$ and $x_2 \in \Sigma_G$. Hence $rt_{x_1}(E) = rt_{x_1}(F)G = G = rt_{x_2}(G) = rt_{x_2}(E)$.

Proofs for the remaining two cases are similar to the above cases.

3. $E = F^*$. Since $(w_1x_1)^{-1}(E) \sim_{aci} (w_2x_2)^{-1}(E)$, by Corollary 1 we have $rt_{x_1}(E) \sim_{aci} rt_{x_2}(E)$. By Proposition 8 $rt_{x_1}(E) = rt_{x_1}(F)F^*, rt_{x_2}(E) = rt_{x_2}(F)F^*$. So $rt_{x_1}(F) \sim_{aci} rt_{x_2}(F)$, which implies there are $(u_1x_1)^{-1}(F), (u_2x_2)^{-1}(F) \neq \emptyset$, such that $(u_1x_1)^{-1}(F) \sim_{aci} (u_2x_2)^{-1}(F)$. Then from the inductive hypothesis, we have $rt_{x_1}(F) = rt_{x_2}(F)$. Hence $rt_{x_1}(E) = rt_{x_1}(F)F^* = rt_{x_2}(E)$.

($\Leftarrow$) This is obvious from Corollary 1. □

Corollary 2 *Let all symbols in E be distinct. If $rt_{x_1}(E) \sim_{aci} rt_{x_2}(E)$, then $rt_{x_1}(E) = rt_{x_2}(E)$.*

Remark 1. From the previous discussions, it is clear that $rt_x(E)$'s are 'atomic' building blocks, in the following meanings. (1) Each non-null $(wx)^{-1}(E)$ is uniquely decomposed into a sum of $rt_x(E)$, that is, $(wx)^{-1}(E) = \Sigma\, rt_x(E)$. (2) $rt_x(E)$ and $rt_y(E)$ are either identical, or not equivalent modulo $\sim_{aci}$, if $x \neq y$.

4 Concluding remarks

The paper proposed a characterization of the structure of derivatives and proved several properties of derivatives for an expression with distinct symbols. Base on this, it gave a representative of derivatives and presented a simpler proof of the fact that the equation automaton is a quotient of the position automaton.

We believe that the characterization of derivatives given in the paper is a useful technique for relevant researches.

The results can have many applications. One example is the correction of Ilie and Yu's simplified proof of the relation between the equation and position automata [8]. Champarnaud and Ziadi [5] proved that the equation automaton [1] is a quotient of the position automaton. Ilie and Yu [8] presented a simplified proof, which relies only on the work of Berry and Sethi [2]. The central issue to use Ilie and Yu's approach is to find a unique representative for $(wx)^{-1}(\overline{E})$. However, the proof given by Ilie and Yu actually fails to find the correct representatives. Thus the proof is incorrect. This may partly reflects the difficulty of finding the representatives. On the other hand, by the results presented in the paper, the representatives are obtained immediately.

References

[1] V. Antimirov. Partial derivatives of regular expressions and finite automaton constructions. Theoretical Computer Science 155 (1996) 291–319.

[2] G. Berry, R. Sethi, From regular expressions to deterministic automata, Theoretical Computer Science 48 (1986) 117–126.

[3] A. Bruggemann-Klein, Regular expressions into finite automata, Theoretical Computer Science 120 (1993) 197–213.

[4] J. A. Brzozowski, Derivatives of regular expressions, J. ACM 11(4):481–494, 1964.

[5] J.-M. Champarnaud, D. Ziadi, Canonical derivatives, partial derivatives and finite automaton constructions, Theoretical Computer Science 289 (2002) 137–163.

[6] C.-H. Chang and R. Page, From regular expressions to DFAs using compressed NFA's, Theoretical Computer Science, 178(1-2):1–36, 1997

[7] V. M. Glushkov, The abstract theory of automata, Russian Math. Surveys 16 (1961) 1–53.

[8] L. Ilie and S. Yu. Follow automata. Information and Computation, 186(1):146–162, 2003.

[9] R. McNaughton, H. Yamada, Regular expressions and state graphs for automata, IEEE Trans. on Electronic Computers 9 (1) (1960) 39–47.

[10] J.-L. Ponty, D. Ziadi and J.-M. Champarnaud, A new Quadratic Algorithm to convert a Regular Expression into an Automaton, WIA'96, LNCS 1260, 1997, 109–119.

[11] S. Yu, Regular Languages, in: G. Rozenberg, A. Salomaa, eds., Handbook of Formal Languages, Vol. I, Springer-Verlag, Berlin, 1997, 41–110.