Derivatives of Regular Expressions and an Application*

Haiming Chen[1] and Sheng Yu[2]

[1] State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences
Beijing 100190, China
chm@ios.ac.cn

[2] Department of Computer Science
University of Western Ontario, London, Ontario, N6A 5B7, Canada
syu@csd.uwo.ca

Abstract. In this paper, we propose a characterization of the structure of derivatives and prove several new properties of derivatives for regular expressions. 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. As an application, an improvement of Ilie and Yu's proof of the relation between the partial derivative and Glushkov automata is presented.

1 Introduction

The construction of finite automata from regular expressions is an important issue and has been studied for a long time. Note that finite automata have always been one of Calude's research interests [5]. 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 nondeterministic finite automata (NFA), the Glushkov automaton was proposed separately by Glushkov [8] and McNaughton and Yamada [10]. Berry and Sethi [2] showed that the Glushkov automaton has a natural connection with the notion of derivative [4], and related the above two different approaches.

The notion of derivative was generalized to partial derivatives by Antimirov [1], which yields the partial derivative automaton, introduced in [1]. Champarnaud and Ziadi [6] proved that the partial derivative automaton is a quotient of the Glushkov automaton. Therefore the partial derivative automaton is smaller than or equal to the Glushkov automaton. The latter has size at most quadratic and can be computed in quadratic time [3,7,11]. They also proposed a quadratic algorithm [6] for computing the partial derivative automata which improved very much the original Antimirov's algorithm. It appears that the partial derivative automaton is among the very small automata converted from a regular expression. Follow automata were introduced in [9]. For a given regular expression E,

* The work of the first author was supported by the National Natural Science Foundation of China under Grants 61070038, 60573013, and that of the second author was supported by the NSERC Discovery Grant 41630.

M.J. Dinneen et al. (Eds.): WTCS 2012 (Calude Festschrift), LNCS 7160, pp. 343–356, 2012.

the size of the follow automaton constructed from E is at most $\frac{3}{2}|E| + \frac{5}{2}$, which is very close to a lower bound, where $|E|$ is the size of E.

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 work provides new and deeper insight into the nature of derivatives.

Berry and Sethi showed that the derivatives in a certain class of derivatives of an expression E with distinct symbols correspond to the same state of the Glushkov automaton of E. This means that the derivatives that correspond 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 non-trivial as discussed in Section 4. By the work on derivatives in the paper, the representatives can be obtained immediately.

As an application, an improved Ilie and Yu's proof of the relation between the partial derivative and Glushkov automata [9] is presented in this paper.

Section 2 introduces notations and notions required in the paper. Section 3 proposes a characterization of derivatives and several properties of derivatives. Section 4 presents a proof of the relation between the partial derivative and Glushkov automata. Section 5 concludes the paper.

2 Preliminaries

We assume that the reader is familiar with basic regular language and automata theory, e.g., from [12], so that we introduce here only some notations and notions used 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 by ε. A regular expression over Σ is $\emptyset$, ε or a for any $a \in \Sigma$, or is the union $E_1 + E_2$, the concatenation $E_1 E_2$, or the star E_1^* for regular expressions E_1 and E_2. For a regular expression E, the language specified by E is denoted by $L(E)$. The size of E is denoted by $|E|$ and is the length of E when written in postfix (parentheses are not counted). The number of symbol occurrences in E, or the alphabetic width of E, is denoted by $\|E\|$. 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 *ACI-similar* or *similar* [4], which is denoted by $E_1 \sim_{aci} E_2$.

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 sets:

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

One can easily write equivalent inductive definitions of the above sets 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 $(a_1 + b_1)^* a_2 b_2 (a_3 + b_3)$ is a marking of the expression $(a + b)^* ab(a + b)$. A marking of an expression E is denoted by $\overline{E}$. If E is a marked expression, then $\overline{E}$ means dropping of subscripts from E. It will be clear from the context whether $\overline{\cdot}$ adds or drops subscripts. We extend the notation for words and automata in the obvious way.

In this way the subscripted symbols are called *positions* of the expression. In the literature, positions are sometimes defined as the subscripts. This definition of positions, however, has drawbacks because it separates subscripts from symbols. When both subscripts and related symbols are required, this presentation is rather awkward. Here we use symbols in $\Sigma_{\overline{E}}$ as the positions, which makes related definitions concise and more flexible (subscripts can be the same, as in the above example).

A finite automaton is a quintuple $M = (Q, \Sigma, \delta, q_0, F)$, where Q is the 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. 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 $p_1 \equiv q_1$ for $p_1 \in \delta(p, a), q_1 \in \delta(q, a)$. If $\equiv$ is right invariant, then we can define a quotient automaton $M/_{\equiv}$ in the usual way. 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 $w^{-1}(L) = \{u \,|\, wu \in L\}$.

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

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

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

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 dissimilar 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 of E with respect to a, $\partial_a(E)$, is defined as follows*[1]*:*

偏导数

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

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*

$$L(\partial_w(E)) = \bigcup_{p \in \partial_w(E)} L(p).$$

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 Glushkov and Partial Derivative Automata

The Glushkov or position automaton was introduced independently by Glushkov [8] and McNaughton and Yamada [10].

Definition 3. *The Glushkov automaton of E is*

$$M_{\mathrm{g}}(E) = (Q_{\mathrm{g}}, \Sigma, \delta_{\mathrm{g}}, q_E, F_{\mathrm{g}}),$$

where

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

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

As shown by Glushkov [8] and McNaughton and Yamada [10], $L(M_{\mathrm{g}}(E)) = L(E)$. $M_{\mathrm{g}}(E)$ can be computed in quadratic time [3,7,11].

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

Definition 4. *The partial derivative 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 $q \in PD(E), a \in \Sigma$.

Note that $PD(E)$ has been defined in the previous subsection. It is proved [6] that $M_{\mathrm{pd}}(E)$ is a quotient of $M_{\mathrm{g}}(E)$.

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 $a \in \Sigma_E$, for all words w,*

1. If $E = E_1 + E_2$, then

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

2. If $E = E_1E_2$, then

$$(wa)^{-1}(E_1E_2) = \begin{cases} (wa)^{-1}(E_1)E_2 \text{ if } a \in \Sigma_{E_1}, w \in \Sigma_{E_1}^* \\ (va)^{-1}(E_2) \qquad \text{if } w = uv, \varepsilon \in L(u^{-1}(E_1)), a \in \Sigma_{E_2}, \\ \qquad\qquad\qquad\quad u \in \Sigma_{E_1}^*, v \in \Sigma_{E_2}^* \\ \emptyset \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{2}$$

Proof. 1. It is directly from Berry and Sethi [2].

2. From Berry and Sethi [2] it is already known

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

Let us consider (b) and set $wa = a_1a_2 \ldots a_t$. 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 $(wa)^{-1}(E_1E_2) = \emptyset$.

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

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 $a \in \Sigma_E$, $(wa)^{-1}(E)$ 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 $(wa)^{-1}(\overline{E})$ to the state a of $M_{\mathrm{g}}(E)$ for an expression E.

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

Theorem 1. *Let all symbols in E be distinct. Given a fixed $a \in \Sigma_E$, for all words w, each non-null $(wa)^{-1}(E)$ 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 $(wa)^{-1}(E)$, 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 $E = b, b \in \Sigma_E$, then $a^{-1}(b) = \varepsilon$ for $a = b$, $(wa)^{-1}(E) = \emptyset$ for $w \neq \varepsilon$ or $a \neq b$. 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 $(wa)^{-1}(E)$ is either $(wa)^{-1}(E_1)$ or $(wa)^{-1}(E_2)$. Suppose the first, then $(wa)^{-1}(E_1)$ is non-null, and the inductive hypothesis applies to it. The repeating term of $(wa)^{-1}(E)$ is the same as $(wa)^{-1}(E_1)$, and no top-level $+$ will be added. The same is for the second.

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

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

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

In the following examples the expression is taken from [9].

Example 1. Let $E = (a + b)(a^* + ba^* + b^*)^*$, then
$\overline{E} = (a_1 + b_2)(a_3^* + b_4a_5^* + b_6^*)^*$,
$a_1^{-1}(\overline{E}) = (a_3^* + b_4a_5^* + b_6^*)^* = \tau_1$,
$(a_1a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_1) = a_3^*\tau_1 = \tau_2$,
$(a_1a_3a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_2) = \tau_2 + \tau_2$,
...
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 $rt_a(E)$ the repeating term of $(wa)^{-1}(E)$. From Theorem 1 we have

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

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

Below we consider the question: *For each $a \in \Sigma_E$, whether there is a non-null $(wa)^{-1}(E)$ containing one and only one $rt_a(E)$, that is, $rt_a(E)$ 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 $ind : \Sigma_E \to \{1, \ldots, \|E\|\}$: $ind(a) = d$ if a is the dth occurrence of symbols from left to right in E (Note that each symbol in E occurs only once). For $a, b \in \Sigma_E$, define $a < b$ iff $ind(a) < ind(b)$. 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 $w_1 = a_1 \ldots a_n, w_2 = a'_1 \ldots a'_n$, there exists an integer $k, 1 \leq k \leq n$, such that $a_t = a'_t$ for $t = 1, \ldots, k-1$, and $a_k < a'_k$. A non-null $(wa)^{-1}(E)$ is called the *first appearance* of derivative of E w.r.t. a, denoted by $F_a(E)$, if for any other non-null $(w_1 a)^{-1}(E)$ it has $w \prec w_1$. From Berry and Sethi [2] a non-null $(wa)^{-1}(E)$ exists for all $a \in \Sigma_E$, which ensures the existence of $F_a(E)$.

Example 2. For $E = (a+b)(a^* + ba^* + b^*)^*$, $\overline{E} = (a_1 + b_2)(a_3^* + b_4 a_5^* + b_6^*)^*$. The first appearances of derivatives w.r.t. symbols in $\overline{E}$, in which the symbols are underlined, are computed as follows.

$$
\begin{array}{ll}
\underline{a_1}^{-1}(\overline{E}) = (a_3^* + b_4 a_5^* + b_6^*)^* = \tau_1, & \underline{b_2}^{-1}(\overline{E}) = (a_3^* + b_4 a_5^* + b_6^*)^* = \tau_1, \\
(a_1\underline{a_3})^{-1}(\overline{E}) = a_3^{-1}(\tau_1) = a_3^*\tau_1 = \tau_2, & (a_1\underline{b_4})^{-1}(\overline{E}) = b_4^{-1}(\tau_1) = a_5^*\tau_1 = \tau_3, \\
(a_1\underline{b_6})^{-1}(\overline{E}) = b_6^{-1}(\tau_1) = b_6^*\tau_1 = \tau_4, & (b_2 a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_1) = \tau_2, \\
(b_2 b_4)^{-1}(\overline{E}) = b_4^{-1}(\tau_1) = \tau_3, & (b_2 b_6)^{-1}(\overline{E}) = b_6^{-1}(\tau_1) = \tau_4, \\
(a_1 a_3 a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_2) = \tau_2 + \tau_2, & (a_1 a_3 b_4)^{-1}(\overline{E}) = b_4^{-1}(\tau_2) = \tau_3, \\
(a_1 a_3 b_6)^{-1}(\overline{E}) = b_6^{-1}(\tau_2) = \tau_4, & (a_1 b_4 a_3)^{-1}(\overline{E}) = a_3^{-1}(\tau_3) = \tau_2, \\
(a_1 b_4 b_4)^{-1}(\overline{E}) = b_4^{-1}(\tau_3) = \tau_3, & (a_1 b_4 \underline{a_5})^{-1}(\overline{E}) = a_5^{-1}(\tau_3) = \tau_3.
\end{array}
$$

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 $a \in \Sigma_E$, the first appearance $F_a(E)$ consists of only one repeating term.*

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

1. $E = E_1 + E_2$. Consider equation (1). If $(wa)^{-1}(E) = (wa)^{-1}(E_1)$, we show that $F_a(E_1)$ is $(wa)^{-1}(E_1)$. If this is not true, there is a word $w_1 \prec w$ such that $(w_1 a)^{-1}(E_1) \neq \emptyset$. So $(w_1 a)^{-1}(E) \neq \emptyset$, which is a contradiction. Therefore $(wa)^{-1}(E_1)$ is the first appearance and the inductive hypothesis applies to it. The same is for $(wa)^{-1}(E) = (wa)^{-1}(E_2)$.

2. $E = E_1E_2$. Consider equation (2). If $(wa)^{-1}(E) = (wa)^{-1}(E_1)E_2$, similarly as above we can prove that $(wa)^{-1}(E_1)$ is the first appearance, and the inductive hypothesis applies to it. If $(wa)^{-1}(E) = (v_1a)^{-1}(E_2)$ for some v_1 such that $wa = uv_1a$, we show that this subterm is $F_a(E_2)$. Suppose the converse. Then there is a word $v \prec v_1$ such that $(va)^{-1}(E_2) \neq \emptyset$. So it is easy to see that $(uva)^{-1}(E) \neq \emptyset$. But $uva \prec wa$, which is a contradiction. Therefore $(v_1a)^{-1}(E_2)$ is the first appearance and the inductive hypothesis applies to it.

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

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

Proposition 5. *Let all symbols in E be distinct. Given any words $w_1, w_2 \in \Sigma_E^*$ and $a \in \Sigma_E$, if $|w_1| = |w_2|$ and $(w_1a)^{-1}(E), (w_2a)^{-1}(E) \neq \emptyset$, and there is no $w \in \Sigma_E^*$, such that $|w| < |w_1|$ and $(wa)^{-1}(E) \neq \emptyset$, then $(w_1a)^{-1}(E) = (w_2a)^{-1}(E)$.*

Proof. We prove it by induction on the structure of E. If $E = \emptyset$ or ε, no non-null derivative exists. If $E = b$ for a symbol b, the only non-null derivative is ε, in which case $w_1 = w_2 = \varepsilon$ and $a = b$. So $(w_1a)^{-1}(E) = (w_2a)^{-1}(E)$.

1. $E = E_1 + E_2$. If $a \in \Sigma_{E_1}$, from equation (1), we have $(w_1a)^{-1}(E) = (w_1a)^{-1}(E_1)$ and $(w_2a)^{-1}(E) = (w_2a)^{-1}(E_1)$. We can see that there is no w, such that $|w| < |w_1|$ and $(wa)^{-1}(E_1) \neq \emptyset$. Otherwise $(wa)^{-1}(E) = (wa)^{-1}(E_1) \neq \emptyset$ which is a contradiction. So the inductive hypothesis applies to E_1. The proof is the same for $a \in \Sigma_{E_2}$.

2. $E = E_1E_2$. If $a \in \Sigma_{E_1}$, from equation (2), we have $(w_1a)^{-1}(E) = (w_1a)^{-1}(E_1)E_2$ and $(w_2a)^{-1}(E) = (w_2a)^{-1}(E_1)E_2$. Similar as in case 1 we can prove $(w_1a)^{-1}(E_1) = (w_2a)^{-1}(E_1)$. Thus $(w_1a)^{-1}(E) = (w_2a)^{-1}(E)$.

If $a \in \Sigma_{E_2}$, from equation (2), we have $(w_1a)^{-1}(E) = (v_1a)^{-1}(E_2)$ and $(w_2a)^{-1}(E) = (v_2a)^{-1}(E_2)$ 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 $\varepsilon \in L(u_2^{-1}(E_1))$, $u_2 \in \Sigma_{E_1}^*$, and $v_1 \in \Sigma_{E_2}^*$, by equation (2) $(u_2v_1a)^{-1}(E) = (v_1a)^{-1}(E_2) \neq \emptyset$. But $|u_2v_1| < |w_1|$ which is a contradiction. So $|v_1| = |v_2|$. Similarly we can prove there is no $v \in \Sigma_{E_2}^*$, such that $|v| < |v_1|$ and $(va)^{-1}(E_2) \neq \emptyset$. So 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 $(w_1a)^{-1}(E) = (w_1a)^{-1}(E_1)E_1^*$ and $(w_2a)^{-1}(E) = (w_2a)^{-1}(E_1)E_1^*$ and there is no $w \in \Sigma_{E_1}^*$ such that $|w| < |w_1|$ and $(wa)^{-1}(E_1) \neq \emptyset$. Then by the inductive hypothesis $(w_1a)^{-1}(E_1) = (w_2a)^{-1}(E_1)$, thus $(w_1a)^{-1}(E) = (w_2a)^{-1}(E)$.

In the above proposition we easily see that $(w_1a)^{-1}(E) = (w_2a)^{-1}(E) = F_a(E)$. Therefore $F_a(E)$ 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 $a \in \Sigma_E$, such that $(wa)^{-1}(E) = rt_a(E)$.*

Proof. The first appearance $F_a(E)$ is one such $(wa)^{-1}(E)$ satisfying $F_a(E) = rt_a(E)$.

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 $rt_a(E)$.

Proposition 7. *Let all symbols in E be distinct. For each $a \in \Sigma_E$,*
(1) $rt_a(E)$ exists,
(2) $rt_a(E)$ is unique.

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

(2) Suppose $rt_a(E)$ is not unique. That is, for some $a \in \Sigma_E$, 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 $rt_a(E)$ is unique.

If $E = \emptyset$ or ε, no symbol is in E, so $rt_a(E)$ is undefined. We let $rt_a(\emptyset) = rt_a(\varepsilon) = \emptyset$ for any $a \in \Sigma$ for the sake of completeness.

The following lemma will be used in the proof of Proposition 8.

Lemma 1. *Let all symbols in E be distinct. If $(wa)^{-1}(E) \sim_{aci} E$ for some $w \in \Sigma_E^*$, then $rt_a(E) = E$.*

Proof. We prove by induction on the structure of E. If $E = \emptyset$, then $(wa)^{-1}(E) = \emptyset$. By assumption, $rt_a(E) = \emptyset$. So $rt_a(E) = E$. If $E = \varepsilon$, then $(wa)^{-1}(E) = \emptyset$, $(wa)^{-1}(E) \not\sim_{aci} E$. If $E = a$, then $a^{-1}(E) = \varepsilon$, $(wb)^{-1}(E) = \emptyset$ for $w \neq \varepsilon$ or $b \neq a$. So $(wa)^{-1}(E) \not\sim_{aci} E$.

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

2. $E = FG$. By the rules ($\emptyset\varepsilon$-rules), $F, G \neq \emptyset$ or ε, then since $(wa)^{-1}(E) \sim_{aci} E \neq \emptyset$, by equation (2) $wa^{-1}(E)$ is either $(wa)^{-1}(F)G$ or $(va)^{-1}(G)$ for some v such that $w = uv$. If $wa^{-1}(E) = (wa)^{-1}(F)G$, then $(wa)^{-1}(F)G \sim_{aci} FG$. So $(wa)^{-1}(F) \sim_{aci} F$. By the inductive hypothesis, we have $rt_a(F) = F$. By equation (2) $wa^{-1}(E) = (wa)^{-1}(F)G$ implies $a \in \Sigma_F$. Hence, from the proof of Theorem 1 we know $rt_a(E) = rt_a(F)G = FG$. If $wa^{-1}(E) = (va)^{-1}(G)$, then $(va)^{-1}(G) \sim_{aci} FG$. Since $(va)^{-1}(G)$ does not contain symbols in F, we have $F = \varepsilon$, which is a contradiction.

3. $E = F^*$. If $E = \emptyset$, then $rt_a(E) = E$. Otherwise $E \neq \emptyset$, then $(wa)^{-1}(E) \neq \emptyset$. From the proof of Theorem 1 we have $rt_a(E) = rt_a(F)F^*$. Thus $(wa)^{-1}(E)$ is a sum of one or more $rt_a(F)F^*$. Since $(wa)^{-1}(E) \sim_{aci} F^*$, we have $rt_a(F) = \varepsilon$. Hence $rt_a(E) = rt_a(F)F^* = F^* = E$.

This means if $w^{-1}(E) \sim_{aci} E$ then E does not contain any top-level $+$, or, equivalently, if E contains any top-level $+$, then $w^{-1}(E) \not\sim_{aci} E$ for any $w \in \Sigma_E^*$.

Remark. If the rules ($\emptyset\varepsilon$-rules) do not hold, the above lemma can also be proved without difficulty.

Proposition 8. *Let all symbols in E be distinct. If there are non-null $(w_1a_1)^{-1}$ (E) and $(w_2a_2)^{-1}(E)$, such that $(w_1a_1)^{-1}(E) \sim_{aci} (w_2a_2)^{-1}(E)$, then $rt_{a_1}(E) = rt_{a_2}(E)$, and vice versa.*

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

1. $E = F+G$. From equation (1), the non-null $(w_1a_1)^{-1}(E)$ is either $(w_1a_1)^{-1}$ (F) or $(w_1a_1)^{-1}(G)$. Likewise, the non-null $(w_2a_2)^{-1}(E)$ is either $(w_2a_2)^{-1}(F)$ or $(w_2a_2)^{-1}(G)$.

If

$$(w_1a_1)^{-1}(E) = (w_1a_1)^{-1}(F), (w_2a_2)^{-1}(E) = (w_2a_2)^{-1}(F) \quad (a),$$

then $(w_1a_1)^{-1}(F) \sim_{aci} (w_2a_2)^{-1}(F)$. By the inductive hypothesis, we have $rt_{a_1}(F) = rt_{a_2}(F)$. In addition, (a) implies $a_1, a_2 \in \Sigma_F$. Then from the proof of Theorem 1 we know $rt_{a_1}(E) = rt_{a_1}(F)$, and $rt_{a_2}(E) = rt_{a_2}(F)$. Hence $rt_{a_1}(E) = rt_{a_2}(E)$.

If

$$(w_1a_1)^{-1}(E) = (w_1a_1)^{-1}(F), (w_2a_2)^{-1}(E) = (w_2a_2)^{-1}(G) \quad (b),$$

then $(w_1a_1)^{-1}(F) \sim_{aci} (w_2a_2)^{-1}(G)$. Since symbols in F and G are distinct, we have $(w_1a_1)^{-1}(F) = (w_2a_2)^{-1}(G) = \varepsilon$. Then from Theorem 1 we have $rt_{a_1}(F) = rt_{a_2}(G) = \varepsilon$. In addition, (b) implies $a_1 \in \Sigma_F$ and $a_2 \in \Sigma_G$. Hence similarly from the proof of Theorem 1 we know $rt_{a_1}(E) = rt_{a_1}(F)$ and $rt_{a_2}(E) = rt_{a_2}(G)$, So $rt_{a_1}(E) = rt_{a_2}(E)$.

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

2. $E = FG$. From equation (2), the non-null $(w_1a_1)^{-1}(E)$ is either $(w_1a_1)^{-1}$ $(F)G$ or $(v_1a_1)^{-1}(G)$ for some v_1 such that $w_1 = u_1v_1$. Likewise, the non-null $(w_2a_2)^{-1}(E)$ is either $(w_2a_2)^{-1}(F)G$ or $(v_2a_2)^{-1}(G)$.

If

$$(w_1a_1)^{-1}(E) = (w_1a_1)^{-1}(F)G, (w_2a_2)^{-1}(E) = (w_2a_2)^{-1}(F)G \quad (a),$$

then $(w_1a_1)^{-1}(F)G \sim_{aci} (w_2a_2)^{-1}(F)G$, which then implies $(w_1a_1)^{-1}(F) \sim_{aci}$ $(w_2a_2)^{-1}(F)$. By the inductive hypothesis, we have $rt_{a_1}(F) = rt_{a_2}(F)$. In addition, (a) implies $a_1, a_2 \in \Sigma_F$. Then from the proof of Theorem 1 we know $rt_{a_1}(E) = rt_{a_1}(F)G$, and $rt_{a_2}(E) = rt_{a_2}(F)G$. Hence $rt_{a_1}(E) = rt_{a_2}(E)$.

If

$$(w_1a_1)^{-1}(E) = (w_1a_1)^{-1}(F)G, (w_2a_2)^{-1}(E) = (v_2a_2)^{-1}(G) \quad (b),$$

then $(w_1a_1)^{-1}(F)G \sim_{aci} (v_2a_2)^{-1}(G)$. Since $(v_2a_2)^{-1}(G)$ does not contain symbols in F, we have $(w_1a_1)^{-1}(F) = \varepsilon$, and $G \sim_{aci} (v_2a_2)^{-1}(G)$. Since $(w_1a_1)^{-1}(F) = \varepsilon$, from Theorem 1 we have $rt_{a_1}(F) = \varepsilon$. By Lemma 1 $G \sim_{aci} (v_2a_2)^{-1}(G)$ implies $rt_{a_2}(G) = G$. In addition, (b) implies $a_1 \in \Sigma_F$ and $a_2 \in \Sigma_G$. Hence $rt_{a_1}(E) = rt_{a_1}(F)G = G = rt_{a_2}(G) = rt_{a_2}(E)$.

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

3. $E = F^*$. Since $(w_1a_1)^{-1}(E) \sim_{aci} (w_2a_2)^{-1}(E)$, by Corollary 1 we have $rt_{a_1}(E) \sim_{aci} rt_{a_2}(E)$. From the proof of Theorem 1 we know

$$rt_{a_1}(E) = rt_{a_1}(F)F^*, rt_{a_2}(E) = rt_{a_2}(F)F^* \quad .$$

So $rt_{a_1}(F) \sim_{aci} rt_{a_2}(F)$, which implies there are $(u_1a_1)^{-1}(F), (u_2a_2)^{-1}(F) \neq \emptyset$, such that $(u_1a_1)^{-1}(F) \sim_{aci} (u_2a_2)^{-1}(F)$. Then from the inductive hypothesis, we have $rt_{a_1}(F) = rt_{a_2}(F)$. Hence $rt_{a_1}(E) = rt_{a_1}(F)F^* = rt_{a_2}(E)$.

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

Corollary 2. *Let all symbols in E be distinct. If $rt_{a_1}(E) \sim_{aci} rt_{a_2}(E)$, then $rt_{a_1}(E) = rt_{a_2}(E)$.*

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

4 An Application

The above results solve an issue in using Berry and Sethi's result. Berry and Sethi showed that an arbitrary derivative in the class $\{(wx)^{-1}(\overline{E}) \mid w \in \Sigma^*_{\overline{E}}\}$ corresponds to the state x of the Glushkov automaton of E. This means that the derivatives that correspond 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 non-trivial as is discussed later in this section. By the theoretical work on derivatives in the paper, the representatives can be obtained immediately. As an application this section gives an improvement of Ilie and Yu's proof [9].

4.1 Background

Champarnaud and Ziadi's proof of the fact that the partial derivative automaton is a quotient of the Glushkov automaton resorts to their newly defined notion of c-derivative [6]. It is thus an interesting and attractive issue whether a proof can directly use only the notions of derivative and partial derivative. Ilie and Yu [9] presented such a proof, which is much simplified compared with Champarnaud and Ziadi's proof. The central issue to use Ilie and Yu's approach is to find a unique representative for a class of derivatives mentioned above. As we show shortly, the proof in [9] actually fails to find the correct representatives. See next subsection for details. The difficulty of finding the representatives may also be partly reflected by the fact that the first proof (Champarnaud and Ziadi [6]) has to use an indirect approach.

Since a correct proof directly using only derivatives and partial derivatives may provide insight and helpful techniques for related researches, for example research of algorithms for partial derivative automata, it is valuable to give such a proof.

In the following, based on our work on derivatives, and in the spirit of Ilie and Yu, an improved proof which directly uses only the notions of derivative and partial derivative is presented.

4.2 Ilie and Yu's Proof

It is claimed in the proof [9] that, by using the rules ($\emptyset\varepsilon$-rules), for a fixed $x \in \Sigma_{\overline{E}}$ and for all words w, $(wx)^{-1}(\overline{E})$ is either $\emptyset$ or unique. However this result is incorrect, which can be seen from the following example.

Example 3. In Example 1, $(a_1a_3)^{-1}(\overline{E})$ and $(a_1a_3a_3)^{-1}(\overline{E})$ are distinct.

The whole proof is based on this uniqueness assumption.

4.3 An Improved Proof

From Theorem 1 and the definitions of derivatives and partial derivatives it is easy to see that for an expression E if $\partial_{wx}(\overline{E}) \neq \emptyset$ then $\partial_{wx}(\overline{E}) = \{rt_x(\overline{E})\}$.

For a letter $x \in \Sigma_{\overline{E}}$, recall that Berry and Sethi's continuation, denoted $C_x(\overline{E})$, is any expression $(wx)^{-1}(\overline{E}) \neq \emptyset$. We use $rt_x(\overline{E})$ instead of arbitrary $(wx)^{-1}(\overline{E}) \neq \emptyset$ to represent $C_x(\overline{E})$, i. e., $C_x(\overline{E}) = rt_x(\overline{E})$. Now the continuation $C_x(\overline{E})$ is unique. Denote also $C_{q_E}(\overline{E}) = \overline{E}$ (q_E is the start state of the Glushkov automaton of E). Berry and Sethi's continuation automaton of E is

$$M_{\mathrm{cont}}(E) = (Q, \Sigma, \delta, q, F),$$

where $Q = \{(x, C_x(\overline{E})) \,|\, x \in \Sigma_{\overline{E}} \cup \{q_E\}\}$, $\delta((x, C_x(\overline{E})), a) = \{(y, C_y(\overline{E})) \mid \overline{y} = a \text{ and } y \in first(C_x(\overline{E}))\}$ for $x \in \Sigma_{\overline{E}} \cup \{q_E\}$ and $a \in \Sigma$, $q = (q_E, \overline{E})$, $F = \{(x, C_x(\overline{E})) \mid \varepsilon \in L(C_x(\overline{E}))\}$.

Define $M_1 \simeq M_2$ if two automata M_1 and M_2 are isomorphic. It is proved [2] that $M_{\mathrm{cont}}(E) \simeq M_{\mathrm{g}}(E)$.

By definition, $M_{\mathrm{pd}}(\overline{E})$ takes elements in $\partial_{wx}(\overline{E})$ and $\{\overline{E}\}$ as states for $x \in \Sigma_{\overline{E}}$, with transitions labeled by letters in $\Sigma_{\overline{E}}$. Then $\overline{M_{\mathrm{pd}}(\overline{E})}$ is the automaton obtained from $M_{\mathrm{pd}}(\overline{E})$ by unmarking labels of transitions. From the correspondence between $\partial_{wx}(\overline{E})$ and $C_x(\overline{E})$ it is easy to see that the difference between

$\overline{M_{\text{pd}}(\overline{E})}$ and $M_{\text{cont}}(E)$, hence between $\overline{M_{\text{pd}}(\overline{E})}$ and $M_{\text{g}}(E)$, is whenever $C_x(\overline{E})$ and $C_y(\overline{E})$ are identical, they correspond to the same state in $\overline{M_{\text{pd}}(\overline{E})}$ and correspond to different states in $M_{\text{cont}}(E)$. This leads to the following proposition.

Define the equivalence $=_{c'} \subseteq Q^2$ by $(x_1, C_{x_1}(\overline{E})) =_{c'} (x_2, C_{x_2}(\overline{E}))$ iff $C_{x_1}(\overline{E}) = C_{x_2}(\overline{E})$. The equivalence is right invariant w.r.t. $M_{\text{cont}}(E)$. Define the equivalence $=_c \subseteq (\Sigma_{\overline{E}} \cup \{q_E\})^2$ by $x_1 =_c x_2$ iff $C_{x_1}(\overline{E}) = C_{x_2}(\overline{E})$. The equivalence is right invariant w.r.t. $M_{\text{g}}(E)$.

Proposition 9. $\overline{M_{\text{pd}}(\overline{E})} \simeq M_{\text{cont}}(E)/_{=_{c'}} \simeq M_{\text{g}}(E)/_{=_c}$.

Define $\equiv_c \subseteq (\Sigma_{\overline{E}} \cup \{q_E\})^2$ by $x_1 \equiv_c x_2$ iff $\overline{C_{x_1}(\overline{E})} = \overline{C_{x_2}(\overline{E})}$. The equivalence is right invariant w.r.t. $M_{\text{g}}(E)$. It is easy to see that $=_c \subseteq \equiv_c$. That is, $M_{\text{g}}(E)/_{\equiv_c}$ is a quotient of $M_{\text{g}}(E)/_{=_c}$. Let us compute the quotient. Suppose $M_{\text{g}}(E) = (Q, \Sigma, \delta, q, F)$, then $M_{\text{g}}(E)/_{=_c} = (Q/_{=_c}, \Sigma, \delta_{=_c}, [q]_{=_c}, F/_{=_c})$, $M_{\text{g}}(E)/_{\equiv_c} = (Q/_{\equiv_c}, \Sigma, \delta_{\equiv_c}, [q]_{\equiv_c}, F/_{\equiv_c})$. Define the equivalence $\equiv \subseteq (Q/_{=_c})^2$ by $[x_1]_{=_c} \equiv [x_2]_{=_c}$ iff $\overline{C_{x_1}(\overline{E})} = \overline{C_{x_2}(\overline{E})}$. The equivalence is right invariant w.r.t. $M_{\text{g}}(E)/_{=_c}$. Then $M_{\text{g}}(E)/_{\equiv_c} \simeq M_{\text{g}}(E)/_{=_c}/_{\equiv}$. From Proposition 9, $M_{\text{g}}(E)/_{=_c}$ and $\overline{M_{\text{pd}}(\overline{E})}$ are isomorphic. The difference between $\overline{M_{\text{pd}}(\overline{E})}$ and $M_{\text{pd}}(E)$ is that, for $C_x(\overline{E})$ and $C_y(\overline{E})$, $C_x(\overline{E}) \neq C_y(\overline{E})$, whenever $\overline{C_{x_1}(\overline{E})} = \overline{C_{x_2}(\overline{E})}$, they correspond to different states in $\overline{M_{\text{pd}}(\overline{E})}$ and correspond to the same state in $M_{\text{pd}}(E)$. Therefore it is easy to further see that $M_{\text{g}}(E)/_{=_c}/_{\equiv}$ and $M_{\text{pd}}(E)$ are isomorphic. Therefore we have

Theorem 2. $M_{\text{pd}}(E) \simeq M_{\text{g}}(E)/_{\equiv_c}$.

Remark 2. The above proof is possible mainly due to (1) $C_x(\overline{E})$ is unique, which is enabled by selecting a representative, $rt_x(\overline{E})$, for it, and (2) $C_x(\overline{E})$ is still a derivative of $\overline{E}$.

Remark 3. After setting $C_x(\overline{E}) = rt_x(\overline{E})$, the remaining part of the proof for Theorem 2 is in the spirit of Ilie and Yu [9], but reformulated in a more rigorous form and corrects several flaws contained in the original proof.

5 Conclusion

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 an improved proof of Ilie and Yu [9] of the fact that the partial derivative automaton is a quotient of the Glushkov automaton.

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

References

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

2. Berry, G., Sethi, R.: From regular expressions to deterministic automata. Theoretical Computer Science 48, 117–126 (1986)
3. Bruggemann-Klein, A.: Regular expressions into finite automata. Theoretical Computer Science 120, 197–213 (1993)
4. Brzozowski, J.A.: Derivatives of regular expressions. J. ACM 11(4), 481–494 (1964)
5. Calude, C., Calude, E., Khoussainov, B.: Deterministic Automata: Simulation, Universality and Minimality. Developments in Language Theory, 519–537 (1997)
6. Champarnaud, J.-M., Ziadi, D.: Canonical derivatives, partial derivatives and finite automaton constructions. Theoretical Computer Science 289, 137–163 (2002)
7. Chang, C.-H., Page, R.: From regular expressions to DFAs using compressed NFA's. Theoretical Computer Science 178(1-2), 1–36 (1997)
8. Glushkov, V.M.: The abstract theory of automata. Russian Math. Surveys 16, 1–53 (1961)
9. Ilie, L., Yu, S.: Follow automata. Information and Computation 186(1), 146–162 (2003)
10. McNaughton, R., Yamada, H.: Regular expressions and state graphs for automata. IRE Trans. on Electronic Computers 9(1), 39–47 (1960)
11. Ponty, J.-L., Ziadi, D., Champarnaud, J.-M.: A New Quadratic Algorithm to Convert a Regular Expression into an Automaton. In: Raymond, D.R., Yu, S., Wood, D. (eds.) WIA 1996. LNCS, vol. 1260, pp. 109–119. Springer, Heidelberg (1997)
12. Yu, S.: Regular Languages. In: Rozenberg, G., Salomaa, A. (eds.) Handbook of Formal Languages, vol. I, pp. 41–110. Springer, Berlin (1997)