A Procedure for Checking Equality of Regular Expressions

A. GINZBURG*

Carnegie Institute of Technology, Pittsburgh, Pa.

ABSTRACT. A simple "mechanical" procedure is described for checking equality of regular expressions. The procedure, based on the work of A. Salomaa, uses derivatives of regular expressions and transition graphs.

Given a regular expression R, a corresponding transition graph is constructed. It is used to generate a finite set of left-linear equations which characterize R. Two regular events R and S are equal if and only if each constant term in the set of left-linear equations formed for the pair $\begin{pmatrix} R \\ S \end{pmatrix}$ is $\begin{pmatrix} \phi \\ \phi \end{pmatrix}$ or $\begin{pmatrix} \wedge \\ \wedge \end{pmatrix}$.

The procedure does not involve any computations with or transformations of regular expressions and is especially appropriate for the use of a computer.

1. Let R denote a regular expression over the alphabet $\Sigma = \{0, 1\}$ (a two-letter alphabet is taken fos simplicity), and let x be a word in Σ^*. R_x will denote the derivative of R with respect to x, i.e., the set of words $w \in \Sigma^*$ such that $xw \in R$. For instance, for the empty word $\wedge$ one has $R_\wedge = R$. The basic properties of derivatives of regular expressions are derived in [2], where it is also proved that every R has only a finite number of unequal derivatives.

Let

$$R = R^{(1)}, R^{(2)}, \cdots, R^{(m)} \tag{1R}$$

be a set of derivatives of R such that every derivative of R is equal to at least one element in this set and assume that for every $R^{(i)}$ $(i = 1, 2, \cdots, m)$ one can single out $R^{(j_i)}$ and $R^{(k_i)}$ such that[1]

$$R_0^{(i)} = R^{(j_i)} \quad \text{and} \quad R_1^{(i)} = R^{(k_i)}. \tag{2R}$$

Then it is possible to construct the system of left-linear equations

$$R^{(i)} = 0R_0^{(i)} + 1R_1^{(i)} + \gamma^{(i)} = 0R^{(j_i)} + 1R^{(k_i)} + \gamma^{(i)} \quad (i = 1, 2, \cdots, m), \tag{3R}$$

where $\gamma^{(i)} = \wedge$ if $\wedge \in R^{(i)}$, and $\gamma^{(i)} = \phi$ (the empty set) otherwise. The system (3R) has a unique solution (up to equality of regular expressions) [2, 6].

2. Let S be another regular expression, and assume that the set

$$S = S^{(1)}, S^{(2)}, \cdots, S^{(n)} \tag{1S}$$

has the same properties as (1R); i.e., there can be found equalities (2S) similar to

* On leave from Technion, Israel Institute of Technology, Haifa, Israel.

[1] In this paper two regular expressions, R and S, are said to be equal (notation: $R = S$) if and only if the regular events described by these expressions are equal.

Journal of the Association for Computing Machinery, Vol. 14, No. 2, April 1967, pp. 355-362.

(2R). Using them, one can construct a system

$$S^{(i')} = 0S_0^{(i')} + 1S_1^{(i')} + \delta^{(i')} = 0S^{(ji')} + 1S^{(ki')} + \delta^{(i')}$$
$$(i' = 1, 2, \cdots, n, \quad \delta^{(i')} = \wedge \text{ or } \phi) \quad (3$$

similar to (3R).

Using (3R) and (3S), one can build the following "compound system" for R a S. (This construction appears essentially in [6].) Starting with the pair (the "colu vector") $\begin{pmatrix} R \\ S \end{pmatrix}$, i.e., $\begin{pmatrix} R_\wedge \\ S_\wedge \end{pmatrix}$, one writes

$$\begin{pmatrix} R^{(1)} \\ S^{(1)} \end{pmatrix} = \begin{pmatrix} R_\wedge \\ S_\wedge \end{pmatrix} = 0\begin{pmatrix} R_0 \\ S_0 \end{pmatrix} + 1\begin{pmatrix} R_1 \\ S_1 \end{pmatrix} + \begin{pmatrix} \gamma_\wedge \\ \delta_\wedge \end{pmatrix}.$$

Using (3R) and (3S) or, if these systems are not explicitly written, using (2R) a (2S), one replaces in the right-hand side of this equation the derivatives of R a of S by equal derivatives from (1R) and (1S), respectively.

For each pair $\begin{pmatrix} R^{(i)} \\ S^{(i')} \end{pmatrix}$ obtained in the right-hand side of the equation, one adds equation

$$\begin{pmatrix} R^{(i)} \\ S^{(i')} \end{pmatrix} = 0\begin{pmatrix} R_0^{(i)} \\ S_0^{(i')} \end{pmatrix} + 1\begin{pmatrix} R_1^{(i)} \\ S_1^{(i')} \end{pmatrix} + \begin{pmatrix} \gamma^{(i)} \\ \delta^{(i')} \end{pmatrix},$$

and the pairs of derivatives in its right-hand side are replaced once more by ments from (1R) and (1S), using (2R) and (2S). The procedure is continued u there are no new pairs. It follows from the existence of (1R) and (1S) that number u of distinct pairs will satisfy $u \le mn$. By enumerating the pairs, one tains the compound system

$$\begin{pmatrix} R_{(\alpha)} \\ S_{(\alpha)} \end{pmatrix} = 0\begin{pmatrix} R_{(\alpha_0)} \\ S_{(\alpha_0)} \end{pmatrix} + 1\begin{pmatrix} R_{(\alpha_1)} \\ S_{(\alpha_1)} \end{pmatrix} + \begin{pmatrix} \gamma_{(\alpha)} \\ \delta_{(\alpha)} \end{pmatrix},$$

where $\alpha = 1, 2, \cdots, u$, $1 \le \alpha_0 \le u$, $1 \le \alpha_1 \le u$, $R_{(1)} = R_\wedge$, $S_{(1)} =$ and $\gamma_{(\alpha)}$ and $\delta_{(\alpha)}$ are $\wedge$ of ϕ. If $\gamma_{(\alpha)} = \delta_{(\alpha)}$ for every α, one has in (4) two ident systems of equations for the $R_{(\alpha)}$ and $S_{(\alpha)}$; hence, $R_{(\alpha)} = S_{(\alpha)}$ $(\alpha = 1, 2, \cdots,$ particularly $R = R_{(1)} = S_{(1)} = S$.

Conversely, if $R = S$, then in the compound system (4), obtained by (3R) (3S) in the above way, one has necessarily $\gamma_{(\alpha)} = \delta_{(\alpha)}$. (This is explicitly show [6] with "right derivatives" instead of the "left" ones used here.)

Thus, the equality $R = S$ of two regular expressions can be established by sh ing that in the compound system (4) for R and S, $\gamma_{(\alpha)} = \delta_{(\alpha)}$ for all α. This car done by computing derivatives of R and S. Unfortunately, the derivation is o quite cumbersome and involves also the comparison of the results in order to fir finite set containing all unequal derivatives. Therefore, it seems to be of interes find a simple "mechanical" procedure for construction of (4). Such a procedu described below.

3. Given a regular expression R, there exist straightforward algorithms for c structing a transition graph (called also a transition system in [3]) representin

For example, let $R = [10 + (0 + 11)0^*1]^*1$. Consider the transition grap in Figure 1. The vertices (in the present case vertex 1 only) denoted by →

called *initial*, while those denoted by + (vertex 5) are called *final*. This transition graph represents the given R, because every path starting at an initial vertex and ending at a final one corresponds to a word in R, and, conversely, to every word in R there corresponds such a path in G. For example, the path 1-4-3-3-1-2-1-5 describes the word 1101101 $\in R$.

4. The same transition graph G can be used also to describe derivatives of R. To this end, denote by A_x the set of *all* vertices in G which can be reached from the initial vertices following a path corresponding to the word $x \in \Sigma^*$. It follows immediately from the definition of the derivative that R_x consists of all words and only of these words, which correspond to paths leading from the vertices in A_x to the final vertices in G. In short, R_x is represented by the same transition graph, but with A_x as initial vertices.

In the above example, R_1 is described by the same G with initial vertices $A_1 = \{2, 4, 5\}$. The final vertex 5 remains unchanged. Notice that $\wedge \in R_1$, because the vertex 5 is initial and final for R_1.

5. Thus, to every derivative R_x of R there can be put in correspondence a set A_x of vertices of G. The original initial vertices form the $A_\wedge$. The correspondence between the subsets of the set of vertices of G and the unequal derivatives of R is not one-to-one. To every derivative there corresponds at least one such subset, but there are subsets to which no derivative corresponds, and there can be also distinct subsets describing equal derivatives (see the examples below).

Every regular expression can be represented by a finite transition graph, and, thus the mentioned result from [2], that every R has only a finite number of unequal derivatives, follows directly.

6. A system of equations (3) can be derived using the subsets $A_\wedge, A_0, A_1, A_{00}, \cdots$ only, without actual computation of the derivatives. Indeed, consider Table I, which corresponds to G, in Figure 1.

The entries in the first column ("inputs") are words $x \in \Sigma^*$ ordered by length and for the same length by the numerical magnitude. In the second column ("vertices of G") the corresponding subsets of vertices A_x are marked. Thus, $A_\wedge = \{1\}$, $A_0 = \{3\}$, $A_1 = \{2, 4, 5\}$, $A_{00} = \{3\}$, and so on, as can be read directly from

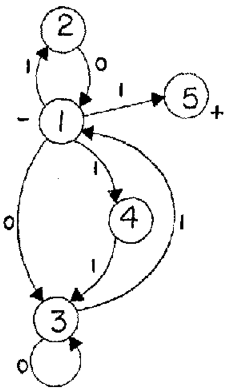

Fig. 1

Figure 1. In the third column ("equal to"), 0 appears in the row of 00, because $A_{00} = A_0$ (i.e., $R_{00} = R_0$). $A_{01} = A_\wedge$ implies $\wedge$ in the row of 01, etc. A row (and the corresponding derivative) with an entry in the column "equal to" will be called a *terminal*. Here all derivatives of "second order" are terminal; i.e., they are equal to derivatives of smaller orders and, clearly, so will be all "higher" derivatives. Thus the table need not be prolonged. As a rule, if the row of x is terminal, one does not enter in the table more inputs beginning with x.

In the last column ("includes $\wedge$") a "yes" appears, if and only if the corresponding A_x includes a final vertex (these vertices are labeled with a +).

For any x which is *not terminal*, the rows $x0$ and $x1$ are added to the table. The process is stopped when there are no new nonterminal words. (There is only a finite number of subsets in a finite set!)

The obtained table can be used to write the system (3R) for R, because the set of the nonterminal derivatives fulfills clearly the properties of (1R). One has

$$\begin{aligned} R &= R_\wedge = 0R_0 + 1R_1 \\ R_0 &= 0R_{00} + 1R_{01} = 0R_0 + 1R \\ R_1 &= 0R_{10} + 1R_{11} + \wedge = 0R + 1R_0 + \wedge. \end{aligned} \tag{5}$$

Notice that $\wedge$ appears in the equations for the derivatives with a "yes" in the last column of the table.

7. The above technique is now used to check an equality $R = S$.

Example 1. An equality from [5]:

$$\begin{aligned} R &\equiv [10 + (0 + 11)0^*1]^*1 \\ &= (10)^*1 + (10)^*(11 + 0)[0 + 1(10)^*(11 + 0)]^*1(10)^*1 \equiv S. \end{aligned}$$

R was discussed above. Now the same procedure will be applied to S.

A transition graph H for S is given in Figure 2.

The system of equations (3S) is here (see Table II):

$$\begin{aligned} S &= 0S_0 + 1S_1 \\ S_0 &= 0S_{00} + 1S_{01} = 0S_0 + 1S_{01} \\ S_1 &= 0S_{10} + 1S_{11} + \wedge = 0S + 1S_0 + \wedge \\ S_{01} &= 0S_{010} + 1S_{011} = 0S_0 + 1S_{011} \\ S_{011} &= 0S_{0110} + 1S_{0111} + \wedge = 0S_{01} + 1S_0 + \wedge. \end{aligned} \tag{6}$$

TABLE I

Inputs	Vertices of G 1	2	3	4	5+	Equal to	Includes $\wedge$
$\wedge$	V						
0			V				
1		V		V	V		yes
00			V			0	
01	V					$\wedge$	
10	V					$\wedge$	
11			V			0	

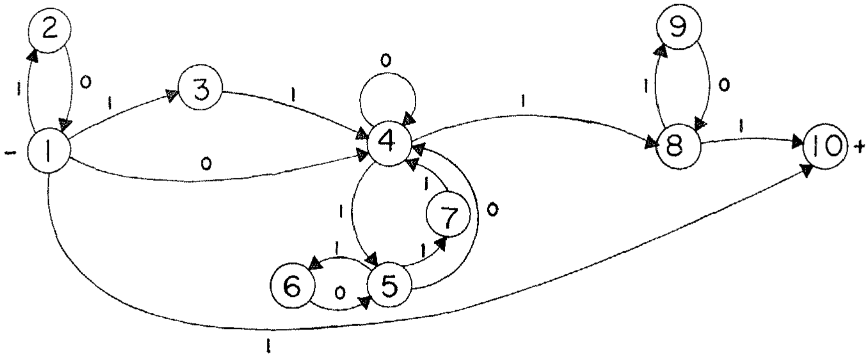

FIG. 2

TABLE II

Inputs	Vertices of H: 1	2	3	4	5	6	7	8	9	10+	*Equal to*	*Includes* Λ
Λ	V											
0				V								
1		V	V							V		yes
00				V							0	
01					V			V				
10	V										Λ	
11				V							0	
010				V							0	
011						V	V		V	V		yes
0110					V			V			01	
0111				V							0	

The compound system can be written using (5) and (6), or directly from the tables, which actually give the equalities (2R) and (2S). One obtains:

$$\begin{pmatrix} R_\Lambda \\ S_\Lambda \end{pmatrix} = 0\begin{pmatrix} R_0 \\ S_0 \end{pmatrix} + 1\begin{pmatrix} R_1 \\ S_1 \end{pmatrix}$$

$$\begin{pmatrix} R_0 \\ S_0 \end{pmatrix} = 0\begin{pmatrix} R_0 \\ S_0 \end{pmatrix} + 1\begin{pmatrix} R_\Lambda \\ S_{01} \end{pmatrix}$$

$$\begin{pmatrix} R_1 \\ S_1 \end{pmatrix} = 0\begin{pmatrix} R_\Lambda \\ S_\Lambda \end{pmatrix} + 1\begin{pmatrix} R_0 \\ S_0 \end{pmatrix} + \begin{pmatrix} \Lambda \\ \Lambda \end{pmatrix}$$

$$\begin{pmatrix} R_\Lambda \\ S_{01} \end{pmatrix} = 0\begin{pmatrix} R_0 \\ S_0 \end{pmatrix} + 1\begin{pmatrix} R_1 \\ S_{011} \end{pmatrix}$$

$$\begin{pmatrix} R_1 \\ S_{011} \end{pmatrix} = 0\begin{pmatrix} R_\Lambda \\ S_{01} \end{pmatrix} + 1\begin{pmatrix} R_0 \\ S_0 \end{pmatrix} + \begin{pmatrix} \Lambda \\ \Lambda \end{pmatrix}.$$

There are no new pairs, and for all appearing pairs $\gamma_{(\alpha)} = \delta_{(\alpha)}$; hence $R = S$. Notice that it follows that $S_{01} = S$ (because $R_{01} = R$), but this fact was not clear from the table for S. This is an example of two equal derivatives with distinct subsets A.

8. *Procedure for Checking an Equality* $R = S$.

I. Construct transition graphs for R and S.

II. Construct the corresponding tables.

III. Write the set of the distinct pairs, which will appear in the compound syster (use the columns "equal to" of the tables).

IV. $R = S$ if and only if *both elements in each pair* simultaneously do or do nc include $\wedge$. (Use the columns "includes $\wedge$" for checking this property.)

9. *Example* 2.

$$R \equiv [(1^*0)^*01^*]^* = \wedge + 0(0+1)^* + (0+1)^*00(0+1)^* \equiv S$$

This equality and the transition graph for R appear in [4]. For R, see Figure 3 an Table III.

Notice that in the case when there are arrows with $\wedge$ in the transition grapl $i \in A_x$ implies that every vertex which can be reached from i by a chain of $\wedge$ arrov is also an element of A_x. In the last case, for example, $A_\wedge$ includes additional to also 2 and 3, and $4 \in A_x \Rightarrow 1, 2, 3 \in A_x$.

For S, see Figure 4 and Table IV. There will appear the following pairs (or omits R and S):

First $\begin{pmatrix}\wedge\\ \wedge\end{pmatrix}$; it implies $\begin{pmatrix}0\\ 0\end{pmatrix}$ and $\begin{pmatrix}1\\ 1\end{pmatrix}$.

The pair $\begin{pmatrix}0\\ 0\end{pmatrix}$ implies $\begin{pmatrix}0\\ 0\end{pmatrix}$ and $\begin{pmatrix}0\\ 01\end{pmatrix}$.

The pair $\begin{pmatrix}1\\ 1\end{pmatrix}$ implies $\begin{pmatrix}10\\ 10\end{pmatrix}$ and $\begin{pmatrix}1\\ 1\end{pmatrix}$.

The two added pairs $\begin{pmatrix}0\\ 01\end{pmatrix}$ and $\begin{pmatrix}10\\ 10\end{pmatrix}$ do not imply new ones; i.e., the set of a appearing pairs is

$$\begin{pmatrix}\wedge\\ \wedge\end{pmatrix}, \begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}1\\ 1\end{pmatrix}, \begin{pmatrix}0\\ 01\end{pmatrix}, \begin{pmatrix}10\\ 10\end{pmatrix}.$$

As both elements in the pairs $\begin{pmatrix}\wedge\\ \wedge\end{pmatrix}, \begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}0\\ 01\end{pmatrix}$ include $\wedge$ and both elements in tl pairs $\begin{pmatrix}1\\ 1\end{pmatrix}, \begin{pmatrix}10\\ 10\end{pmatrix}$ do not include $\wedge$, the checks IV are fulfilled and consequentl $R = S$.

TABLE III

Inputs	1 +	2	3	4	*Equal to*	*Includes* $\wedge$
$\wedge$	$\vee$	$\vee$	$\vee$			yes
0	$\vee$	$\vee$	$\vee$	$\vee$		yes
1			$\vee$			
00	$\vee$	$\vee$	$\vee$	$\vee$	0	yes
01	$\vee$	$\vee$	$\vee$	$\vee$	0	yes
10		$\vee$	$\vee$			
11			$\vee$		1	
100	$\vee$	$\vee$	$\vee$	$\vee$	0	yes
101			$\vee$		1	

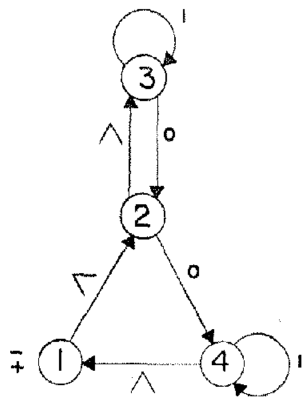

FIG. 3

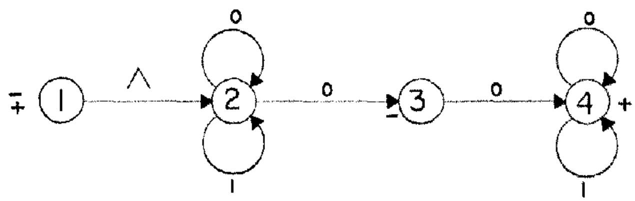

FIG. 4

TABLE IV

Inputs	1 +	2	3	4 +	*Equal to*	*Includes* ∧
∧	V	V	V			yes
0		V	V	V		yes
1		V				
00		V	V	V	0	yes
01		V		V		yes
10		V	V			
11		V			1	
010		V	V	V	0	yes
011		V		V	01	yes
100		V	V	V	0	yes
101		V			1	

10. The use of the tables in the above procedure can be replaced by the following relational technique.

A transition graph G can be described by a set of relations over its vertex set in the obvious way: to every input $\sigma \in \Sigma$ and to $\wedge$ there corresponds a relation T_σ, such that $aT_\sigma b$ if and only if there is in G a σ-arrow from the vertex a to the vertex b.

Denote by $\bar{T}_\wedge$ the transitive closure of $T_\wedge$ and by $\bar{\bar{T}}_\wedge$ the union $\bar{T}_\wedge \cup I$, where I is the identity relation. Then for any $x = \sigma_1\sigma_2 \cdots \sigma_k \in \Sigma^*$ one has

$$A_x = (A_\wedge)(T_{\sigma_1}\bar{\bar{T}}_\wedge T_{\sigma_2}\bar{\bar{T}}_\wedge \cdots T_{\sigma_k}\bar{\bar{T}}_\wedge).$$

(The operation in the brackets is the usual composition of relations, and $(A)T = \{b \mid \exists a \in A, aTb\}$.)

For example, for G in Figure 3,

$$T_0 = \begin{pmatrix} 2 & 3 \\ 4 & 2 \end{pmatrix}, \qquad T_1 = \begin{pmatrix} 3 & 4 \\ 3 & 4 \end{pmatrix}, \qquad T_\wedge = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 3 & 1 \end{pmatrix},$$

$$\bar{T}_\wedge = \begin{pmatrix} 1 & 2 & 4 & 1 & 4 & 4 \\ 2 & 3 & 1 & 3 & 2 & 3 \end{pmatrix}, \qquad \bar{\bar{T}}_\wedge = \begin{pmatrix} 1 & 2 & 3 & 4 & 1 & 2 & 4 & 1 & 4 & 4 \\ 1 & 2 & 3 & 4 & 2 & 3 & 1 & 3 & 2 & 3 \end{pmatrix}.$$

$$A_\wedge = \{1, 2, 3\} \qquad (A_\wedge = \{1\}\bar{\bar{T}}_\wedge),$$

$$A_{10} = (A_\wedge)(T_1 \;\; \bar{\bar{T}}_\wedge \;\; T_0 \;\; \bar{\bar{T}}_\wedge) = \{2, 3\}.$$

This computational approach is especially appropriate for the use of a computer.

ACKNOWLEDGMENT. The author thanks Professor David C. Cooper and Mr. Zohar Manna for their interest and stimulating discussions.

REFERENCES

1. AANDERAA, S. On the algebra of regular expressions. *Appl. Math.*, Harvard U., Cambridge, Mass., Jan. 1965, pp. 1–18 (ditto).
2. BRZOZOWSKI, J. A. Derivatives of regular expressions. *J. ACM 11* (Jan. 1964), 481–494.
3. HARRISON, M. A. *Introduction to Switching and Automata Theory*. McGraw-Hill, New York, 1965.
4. MCNAUGHTON, R. Techniques for manipulating regular expressions. Machines Structures Group Memo No. 10, MIT Project MAC, Cambridge, Mass., Nov. 1965.
5. MCNAUGHTON, R., AND YAMADA, H. Regular expressions and state graphs for automata. *Trans. IRE EC-9* (1960), 39–47.
6. SALOMAA, A. Two complete axiom systems for the algebra of regular events. *J. ACM 13* (1966), 158–169.

RECEIVED JUNE, 1966; REVISED NOVEMBER, 1966