Inclusion of Unambiguous #REs is NP-Hard

Pekka Kilpeläinen
University of Kuopio
Department of Computer Science
Pekka.Kilpelainen@cs.uku.fi

May 27, 2004

Abstract

We show that testing inclusion between languages represented by regular expressions with numerical occurrence indicators (#REs) is NP-hard, even if the expressions satisfy the requirement of "unambiguity", which is required for XML Schema content model expressions.

1 Proof of the result

We have seen before [3] that testing for inclusion and overlap of languages represented by #REs is NP-hard. Testing for the overlap was seen hard also for expressions that satisfy the XML requirement of "unambiguity". On the other hand, the NP-hardness proof of #RE inclusion used ambiguous expressions. Here we show that unambiguity does not make the testing of inclusion essentially easier. The proof is based on a polynomial time Turing reduction [1, Chap. 5] from PARTITION, which is one of the best-known NP-complete problems [2, 1].

Theorem 1.1 *The #RE inclusion problem is NP-hard, also for unambiguous #REs.*

Proof. Let a set $A = \{a_1, \ldots, a_k\}$ and a positive integer weight $w(a)$ of each $a \in A$ form an instance of PARTITION. The problem is to decide whether A can be split in two equal-weight subsets A' and $A - A'$, that is, whether

$$\sum_{a \in A'} w(a) = \sum_{a \in A - A'} w(a) \qquad (1)$$

holds for some $A' \subseteq A$. Notice that (1) can hold only if the total weight of the set A is even. Therefore we can assume that $\sum_{a \in A} w(a) = 2n$ for some positive integer n, which means that (1) holds if and only if

$$\sum_{a \in A'} w(a) = n \tag{2}$$

for some $A' \subseteq A$.

For shortness, denote the weight $w(a_i)$ of an item $a_i \in A$ by w_i.

Now form the following two #REs over the alphabet $\Sigma = \{a_0, a_1, \ldots, a_k\}$:

$$\begin{aligned} E_1 &= a_0^{n+1..n+1}(a_1^{w_1..w_1}|\epsilon)(a_2^{w_2..w_2}|\epsilon)\cdots(a_k^{w_k..w_k}|\epsilon) \\ E_2 &= ((a_0|a_1|\cdots|a_k)^{n+1..2n})^{1..2} \end{aligned}$$

Notice that both expressions are trivially unambiguous since each symbol of Σ appears exactly once in both of them. Expression E_1 describes words of the form $a_0^{n+1}u$, where the length of the suffix u equals the total weight of some subset of A. Therefore $L(E_1) \subseteq \{v \in \Sigma^* \mid n+1 \leq |v| \leq 3n+1\}$. Obviously E_1 accepts a word of length $2n+1$ if and only if a partition that satisfies (2) exists. Expression E_2, on the other hand, *rejects* any words of length $2n+1$:

$$\begin{aligned} L(E_2) &= \bigcup_{i=n+1}^{2n} \Sigma^i \cup \bigcup_{i=2n+2}^{4n} \Sigma^i \\ &= \{v \in \Sigma^* \mid n+1 \leq |v| \leq 4n, |v| \neq 2n+1\} \end{aligned}$$

Now $L(E_1) \subseteq L(E_2)$ holds iff E_1 does not accept any word of length $2n+1$, which holds if and only if no partition which satisfies (1) exists. □

So, a polynomial-time algorithm for testing the inclusion of unambiguous #REs would imply $P = NP$, which is considered most unlikely.

References

[1] M.R. Garey and D.S. Johnson. *Computers and Intractability.* W.H. Freeman and Company, New York, 1979.

[2] R.M. Karp. Reducibility among combinatorial problems. In R.E. Miller and J.W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, New York, 1972.

[3] P. Kilpeläinen and R. Tuhkanen. Regular expressions with numerical occurrence indicators—preliminary results. In *Proc. of the Eighth Symp. on Programming Languages and Software Tools*, pages 163–173. University of Kuopio, 2003.