Information Processing Letters 111 (2011) 614–619

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Parikh's theorem: A simple and direct automaton construction

Javier Esparza [a], Pierre Ganty [b,*,1], Stefan Kiefer [c,2], Michael Luttenberger [a]

[a] *Institut für Informatik, Technische Universität München, 85748 Garching, Germany*
[b] *The IMDEA Software Institute, Madrid, Spain*
[c] *Oxford University Computing Laboratory, Oxford, UK*

ARTICLE INFO

Article history:
Received 22 June 2010
Received in revised form 24 March 2011
Accepted 24 March 2011
Communicated by A. Muscholl

Keywords:
Formal languages
Context-free language
Regular language
Parikh image

ABSTRACT

Parikh's theorem states that the Parikh image of a context-free language is semilinear or, equivalently, that every context-free language has the same Parikh image as some regular language. We present a very simple construction that, given a context-free grammar, produces a finite automaton recognizing such a regular language.
© 2011 Elsevier B.V. All rights reserved.

The *Parikh image* of a word w over an alphabet $\{a_1, \ldots, a_n\}$ is the vector $(v_1, \ldots, v_n) \in \mathbb{N}^n$ such that v_i is the number of occurrences of a_i in w. For example, the Parikh image of $a_1a_1a_2a_2$ over the alphabet $\{a_1, a_2, a_3\}$ is $(2, 2, 0)$. The Parikh image of a language is the set of Parikh images of its words. Parikh images are named after Rohit Parikh, who in 1966 proved a classical theorem of formal language theory which also carries his name. Parikh's theorem [1] states that the Parikh image of any context-free language is *semilinear*. Since semilinear sets coincide with the Parikh images of regular languages, the theorem is equivalent to the statement that every context-free language has the same Parikh image as some regular language. For instance, the language $\{a^nb^n \mid n \geqslant 0\}$ has the same Parikh image as $(ab)^*$. This statement is also often referred to as Parikh's theorem, see e.g. [10], and in fact it has been considered a more natural formulation [13].

Parikh's proof of the theorem, as many other subsequent proofs [8,13,12,9,10,2], is constructive: given a context-free grammar G, the proof produces (at least implicitly) an automaton or regular expression whose language has the same Parikh image as $L(G)$. However, the constructions are relatively complicated, not given in detail, or they yield crude upper bounds, namely automata of size $\mathcal{O}(n^n)$ for grammars in Chomsky normal form with n variables (see Section 4 for a detailed discussion). In this note we present an explicit and very simple construction that yields an automaton with $\mathcal{O}(4^n)$ states for grammars in Chomsky normal form, for a lower bound of $\Omega(2^n)$. An application of the automaton is briefly discussed in Section 3: the automaton can be used to algorithmically derive the semilinear set, and, using recent results on Parikh images of NFAs [15,11], it leads to the best known upper bounds on the size of the semilinear set for a given context-free grammar.

* Corresponding author.
E-mail addresses: esparza@model.in.tum.de (J. Esparza), pierre.ganty@imdea.org (P. Ganty), stefan.kiefer@comlab.ox.ac.uk (S. Kiefer), luttenbe@in.tum.de (M. Luttenberger).
1 This author is sponsored by the Comunidad de Madrid's Program prometidos-cm (S2009TIC-1465), by the people-cofund's program amarout (PCOFUND-2008-229599), and by the Spanish Ministry of Science and Innovation (TIN2010-20639).
2 This author is supported by a postdoctoral fellowship of the German Academic Exchange Service (DAAD).

0020-0190/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ipl.2011.03.019

1. The construction

We follow the notation of [3, Chapter 5]. Let $G = (V, T, P, S)$ be a context-free grammar with a set $V = \{A_1, \ldots, A_n\}$ of *variables* or *nonterminals*, a set T of *terminals*, a set $P \subseteq V \times (V \cup T)^*$ of *productions*, and an *axiom* $S \in V$. We construct a nondeterministic finite automaton (NFA) whose language has the same Parikh image as $L(G)$. The transitions of this automaton will be labelled with words of T^*, but note that by adding intermediate states (when the words have length greater than one) and removing ϵ-transitions (i.e., when the words have length zero), such an NFA can be easily brought in the more common form where transition labels are elements of T.

We need to introduce a few notions. Given $\alpha \in (V \cup T)^*$, we denote by $\Pi_V(\alpha)$ (resp. $\Pi_T(\alpha)$) the Parikh image of α where the components not in V (resp. T) have been projected away. Moreover, we denote by $\alpha_{/V}$ (resp. $\alpha_{/T}$) the projection of α onto V (resp. T). For instance, if $V = \{A_1, A_2\}$, $T = \{a, b, c\}$, and $\alpha = aA_2bA_1A_1$, then $\Pi_V(\alpha) = (2, 1)$, $\Pi_T(\alpha) = (1, 1, 0)$ and $\alpha_{/T} = ab$. Given $\alpha, \beta \in (V \cup T)^*$, let $\mathcal{P}(\alpha, \beta)$ be the set of productions of G that can transform α into β, i.e., $\mathcal{P}(\alpha, \beta) = \{(A \to \gamma) \in P \mid \exists \alpha_1, \alpha_2 \in (V \cup T)^*\colon\ \alpha = \alpha_1 A \alpha_2 \wedge \beta = \alpha_1 \gamma \alpha_2\}$. If $\mathcal{P}(\alpha, \beta) \neq \emptyset$ then we call (α, β) a *step*, denoted by $\alpha \Rightarrow \beta$.

The NFA whose language has the same Parikh image as $L(G)$ will be a member of the following family:

Definition 1.1. Let $G = (V, T, P, S)$ be a context-free grammar, let $n = |V|$, and let $k \geqslant 1$. The *k-Parikh automaton* of G is the NFA $M_G^k = (Q, T^*, \delta, q_0, \{q_f\})$ defined as follows:

- $Q = \{(x_1, \ldots, x_n) \in \mathbb{N}^n \mid \sum_{i=1}^n x_i \leqslant k\}$;
- $\delta = \{(\Pi_V(\alpha), \gamma_{/T}, \Pi_V(\beta)) \mid \exists (A \to \gamma) \in \mathcal{P}(\alpha, \beta)\colon \Pi_V(\alpha), \Pi_V(\beta) \in Q\}$;
- $q_0 = \Pi_V(S)$;
- $q_f = \Pi_V(\varepsilon) = (0, \ldots, 0)$.

It is easily seen that M_G^k has exactly $\binom{n+k}{n}$ states. Fig. 1 shows the 3-Parikh automaton of the context-free grammar with productions $A_1 \to A_1A_2 | a$, $A_2 \to bA_2aA_2 | cA_1$ and axiom A_1. The states are all pairs (x_1, x_2) such that $x_1 + x_2 \leqslant 3$. For instance, transition $(0, 2) \xrightarrow{ba} (0, 3)$ comes (among others) from the step $A_2A_2 \Rightarrow bA_2aA_2A_2$, and can be interpreted as follows: applying the production $A_2 \to bA_2aA_2$ to a word with zero occurrences of A_1 and two occurrences of A_2 leads to a word with one new occurrence of a and b, zero occurrences of A_1, and three occurrences of A_2.

We define the *degree* of G by $m := -1 + \max\{|\gamma_{/V}|\colon (A \to \gamma) \in P\}$; i.e., $m + 1$ is the maximal number of variables on the right-hand sides of the productions. For instance, the degree of the grammar in Fig. 1 is 1. Notice that if G is in Chomsky normal form then $m \leqslant 1$, and $m \leqslant 0$ iff G is regular.

In the rest of the note we prove:

Theorem 1.1. *If G is a context-free grammar with n variables and degree m, then $L(G)$ and $L(M_G^{nm+1})$ have the same Parikh image.*

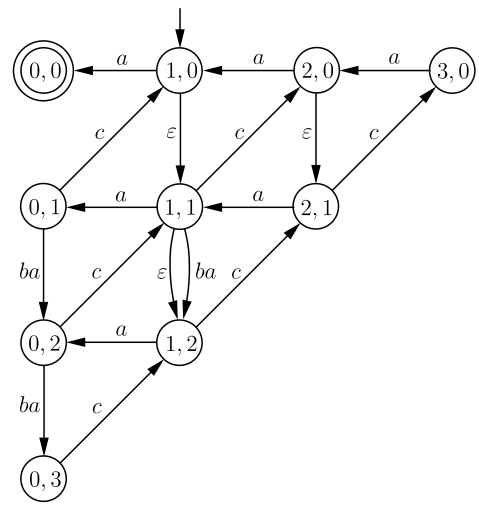

Fig. 1. The 3-Parikh automaton of $A_1 \to A_1A_2 | a$, $A_2 \to bA_2aA_2 | cA_1$ with $S = A_1$.

For the grammar of Fig. 1 we have $n = 2$ and $m = 1$, and Theorem 1.1 yields $L(G) = L(M_G^3)$. So the language of the automaton of the figure has the same Parikh image as the language of the grammar.

Using standard properties of binomial coefficients, for M_G^{nm+1} and $m \geqslant 1$ we get an upper bound of $2 \cdot (m+1)^n \cdot e^n$ states. For $m \leqslant 1$ (e.g. for grammars in Chomsky normal form), the automaton M_G^{n+1} has $\binom{2n+1}{n} \leqslant 2^{2n+1} \in \mathcal{O}(4^n)$ states. On the other hand, for every $n \geqslant 1$ the grammar G_n in Chomsky normal with productions $\{A_k \to A_{k-1}\ A_{k-1} \mid 2 \leqslant k \leqslant n\} \cup \{A_1 \to a\}$ and axiom $S = A_n$ satisfies $L(G_n) = \{a^{2^{n-1}}\}$, and therefore the smallest Parikh-equivalent NFA has $2^{n-1} + 1$ states. This shows that our construction is close to optimal.

2. The proof

Given $L_1, L_2 \subseteq T^*$, we write $L_1 =_\Pi L_2$ (resp. $L_1 \subseteq_\Pi L_2$) to denote that the Parikh image of L_1 is equal to (resp. included in) the Parikh image of L_2. Also, given $w, w' \in T^*$, we abbreviate $\{w\} =_\Pi \{w'\}$ to $w =_\Pi w'$.

We fix a context-free grammar $G = (V, T, P, S)$ with n variables and degree m. In terms of the notation we have just introduced, we have to prove $L(G) =_\Pi L(M_G^{nm+1})$. One inclusion is easy:

Proposition 2.1. *For every $k \geqslant 1$ we have $L(M_G^k) \subseteq_\Pi L(G)$.*

Proof. Let $k \geqslant 1$ arbitrary, and let $q_0 \xrightarrow{\sigma} q$ be a run of M_G^k on the word $\sigma \in T^*$. We first claim that there exists a step sequence $S \Rightarrow^* \alpha$ satisfying $\Pi_V(\alpha) = q$ and $\Pi_T(\alpha) = \Pi_T(\sigma)$. The proof is by induction on the length ℓ of $q_0 \xrightarrow{\sigma} q$. If $\ell = 0$, then $\sigma = \varepsilon$, and we choose $\alpha = S$, which satisfies $\Pi_V(S) = q_0$ and $\Pi_T(S) = (0, \ldots, 0) = \Pi_T(\varepsilon)$. If $\ell > 0$, then let $\sigma = \sigma'\gamma$ and $q_0 \xrightarrow{\sigma'} q' \xrightarrow{\gamma} q$. By induction hypothesis there is a step sequence $S \Rightarrow^* \alpha'$ satisfying $\Pi_V(\alpha') = q'$ and $\Pi_T(\alpha') = \Pi_T(\sigma')$. Moreover, since $q' \xrightarrow{\gamma} q$ is a transition of M_G^k, there is a production $A \to \gamma'$ and a step $\alpha_1 A \alpha_2 \Rightarrow \alpha_1 \gamma \alpha_2$ such that

Download English Version:

https://daneshyari.com/en/article/427309

Download Persian Version:

https://daneshyari.com/article/427309

Daneshyari.com