AN EFFICIENT $LALR(1)$ AND $LR(1)$ LOOKAHEAD SET ALGORITHM

PAULO S. L. M. BARRETO

ABSTRACT. The most popular method for computing $LALR(1)$ lookaheads is not the most efficient one, mainly (as we conjecture) due to certain deficiencies in the description of core algorithms. Similarly, full $LR(1)$ parser generators are not as common as their $LALR(1)$ counterparts, despite their usefulness regarding conventional semantic routine attachment and attribute grammar compilation. We address these two issues by describing an efficient algorithm to compute either $LALR(1)$ or $LR(1)$ lookahead sets without requiring any modification of the underlying transitive closure algorithm.

1. INTRODUCTION

Most compiler building tools employ $LL(1)$ or $LALR(1)$ parser generation algorithms. Modern methods for generating $LALR(1)$ lookahead sets are almost invariably based on computing the transitive closure of certain relations and associated set-valued functions, a problem for which efficient algorithms are known, notably the one by Eve and Kurki-Suonio [5].

Among these methods, the most popular is undoubtedly DeRemer and Pennello's [4], although the algorithm by Park, Choe and Chang [14] (henceforth called simply PCC) is arguably more efficient, and another by Ives [8] is claimed to be even more so. Recent research into the subject includes the method of Anzai [1].

A natural question to ask is why the more efficient algorithms are not as popular as DeRemer and Pennello's. We conjecture that the main reasons are:

- *Clarity of presentation.* DeRemer and Pennello provide a very clear and complete exposition of the problem, the theory underlying the proposed solution, and the implementation steps to be followed. On the other hand, the description of PCC is heavily theoretical, and Ives's paper is very brief and somewhat sketchy.
- *Lack of an efficient* Path/ctx *algorithm.* Both PCC and Ives define and use a new function between nonterminals (called Path by PCC and ctx by Ives[1]) that associates certain sets of terminals to pairs of nonterminals. The concept of Path/ctx is essential in their methods for lookahead computation. However, Ives does not provide any algorithm for computation of ctx, and the method employed by PCC for Path is only superficially described but even so requires a nontrivial change in Eve and Kurki-Suonio's closure algorithm; besides, no formal proof of correctness is given for the modified algorithm.

We also notice that, in many circumstances, it might be convenient to resort to the full power of $LR(1)$ parsers. A typical situation is semantic routine invocation, which usually takes place at nonterminal reductions. If a semantic routine must be activated somewhere

---

Key words and phrases. Compilers, Parser generator, Lookahead set algorithm.

[1]Actually the definition of ctx is slightly different from that of Path, as it includes a special symbol Λ in $\text{ctx}(A, B)$ to flag that $A \xrightarrow[\text{rm}]{*} B$. This is, however, unnecessary, as the unit derivation relationship can be indicated explicitly.

within a production (not at its end), the usually adopted strategy is to introduce into the grammar a new nonterminal that only generates the empty string and serves exclusively as placeholder for the semantic routine call. This technique may be implicitly applied to implement certain classes of attribute grammars like the $ECLR$ class described in [15]. Unfortunately, this procedure may destroy the $LALR(k)$ property for any k [2], a feature clearly undesirable in a compiler construction tool. The use of full $LR(1)$ items in parser automaton states may also be attractive to remove parser conflicts arising in the compilation of regular right-part grammars [12]. Syntax error repair methods like that of [11] also benefit from finer $LR(1)$ lookahead sets.

TBS: discuss error recovery, like the proposals of [9] **and** [3]**.**

Although the canonical $LR(1)$ construction [10] may be too expensive for practical applications due to the large number of states in the generated parser, an efficient algorithm was proposed in [13] and refined in [6] and [16] to build equivalent but state-minimized $LR(1)$ automata. These methods are based on iterative generation and merging of compatible $LR(1)$ states, and are still inefficient in the sense that the transitive closure of a state (the *completer* operation) must be computed repeatedly.

Our contributions in this paper are the following. We address the $LALR(1)$ issues by describing a new algorithm for $\mathrm{Path/ctx}$ computation that does not require any changes in the underlying transitive closure algorithm; rather, we simply construct a different relation graph and show that the new construction is formally correct. Besides, using a lemma by Park, Choe and Chang [14, lemma 3.1], we show how to use the $\mathrm{Path/ctx}$ function to eliminate the need for the repeated *completer* computation, thus removing this inefficiency from the full $LR(1)$ generation algorithm.

2. NOTATION

Throughout this paper we adopt the following notation.

- Σ is the set of *terminal symbols* and $\mathcal{N}$ is the set of *non-terminal symbols*, with $\Sigma \cap \mathcal{N} = \varnothing$. The *vocabulary* is the set $\mathcal{V} = \Sigma \cup \mathcal{N}$.
- $\mathcal{P}$ is the set of *productions*, which have the form $A \to \alpha$ where $A \in \mathcal{N}$ and $\alpha \in \mathcal{V}^*$.
- $A \xrightarrow{\mathrm{rm}} \alpha$ denotes a right derivation of $\alpha \in \mathcal{V}^*$ from $A \in \mathcal{N}$ in one step, and $A \xrightarrow{*}_{\mathrm{rm}} \alpha$ denotes a right derivation in any number of steps. We also use the shorthand $A \xrightarrow{*}_{\mathrm{rm}} B \ldots$, meaning $\exists w \in \Sigma^* : A \xrightarrow{*}_{\mathrm{rm}} Bw$.
- $\forall \alpha \in \mathcal{V}^* : \mathrm{FIRST}(\alpha) \equiv \{t \in \Sigma \mid \alpha \to^* t\beta,\ \beta \in \mathcal{V}^*\}$.
- $\forall \rho, \sigma \subseteq \Sigma^* : \rho \oplus \sigma \equiv \mathrm{FIRST}(rs) : r \in \rho \wedge s \in \sigma$. That is, $\rho \oplus \sigma$ is simply $\mathrm{FIRST}(\rho)$ if $\varepsilon \notin \rho$, or $\mathrm{FIRST}(\rho) \cup \mathrm{FIRST}(\sigma)$ otherwise (the $\oplus$ operation is called *ε-free* FIRST).
- $\forall A, B \in \mathcal{N} : \mathrm{ctx}(A, B) \equiv \{t \in \Sigma \mid A \xrightarrow{\mathrm{rm}}^* Bt\alpha, \alpha \in \mathcal{V}^*\}$. Our notation is only slightly different from that of [8], where $\mathrm{ctx}(A, B)$ contains ε if $A \xrightarrow{*}_{\mathrm{rm}} B$; we prefer to indicate this fact explicitly when needed, and to extend the semantics of $\oplus$ so that:

$$\mathrm{ctx}(A, B) \oplus \sigma = \begin{cases} \mathrm{ctx}(A, B) \cup \sigma, & \text{if } A \xrightarrow{*}_{\mathrm{rm}} B; \\ \mathrm{ctx}(A, B), & \text{otherwise.} \end{cases}$$

3. Construction of the Minimal $LR(1)$ Automaton

The first efficient method to construct the full $LR(1)$ automaton is described in [13]. It is an iterative, fixed-point algorithm that partitions the generated states into *visited* and *unvisited* states, starting with a single unvisited state (corresponding to the grammar's start symbol) and proceeding while any unvisited states remain. At each step, an element is removed from the set of unvisited states, its successors are computed, new successor states are marked as unvisited, and the removed state is marked as visited.

The efficiency of this algorithm derives from the possibility of merging distinct states that satisfy certain compatibility criteria; in particular, one can show that one of the criteria described in [6] (called *strong compatibility*) produces the automaton with the minimum possible number of states for the given grammar.

A central aspect of the algorithm is thus the computation of successors of an $LR(1)$ state. For each item $[A \rightarrow \alpha \cdot X\gamma, u]$ of the state being processed, one must insert into the successor state not only the item $[A \rightarrow \alpha X \cdot \gamma, u]$ but also, if X is a nonterminal, all items of form $[N \rightarrow Y \cdot \beta, v]$, for each production $N \rightarrow Y\beta$ of each nonterminal N such that $X \xrightarrow[\text{rm}]{*} N \ldots$. The lookahead v of one item of this form is given, according to lemma 3.1 of [14], by $v = \text{ctx}(X, N) \oplus \text{FIRST}(\gamma u)$.

4. Efficient Computation of ctx

Our algorithm for the efficient computation of $\text{ctx}(A, B)$ is based on the following observations:

Lemma 1.

$$\forall (A \rightarrow C\alpha) \in \mathcal{P}, \forall B \in \mathcal{N} : \text{ctx}(C, B) \oplus \text{FIRST}(\alpha) \subseteq \text{ctx}(A, B).$$

Proof. Suppose initially that $C \xrightarrow[\text{rm}]{*} Bz$, where $z \in \Sigma^+$ so that $\text{FIRST}(z) \subseteq \text{ctx}(C, b)$. But $A \xrightarrow[\text{rm}]{} C\alpha \xrightarrow[\text{rm}]{*} Cw \xrightarrow[\text{rm}]{*} Bzw$, hence $\text{FIRST}(z) \subseteq \text{ctx}(A, B)$. Therefore, $\text{ctx}(C, B) = \bigcup_{C \xrightarrow[\text{rm}]{*} Bz} \text{FIRST}(z) \subseteq \text{ctx}(A, B)$.

Now suppose that $C \xrightarrow[\text{rm}]{*} B$. In this case, $A \xrightarrow[\text{rm}]{} C\alpha \xrightarrow[\text{rm}]{*} Cw \xrightarrow[\text{rm}]{*} Bw$, where $\alpha \xrightarrow[\text{rm}]{*} w$ e $w \in \Sigma^*$. Thus, $\text{FIRST}(\alpha) = \text{FIRST}(w) \subseteq \text{ctx}(A, B)$. Therefore, $\text{ctx}(C, B) \oplus \text{FIRST}(\alpha) \subseteq \text{ctx}(A, B)$. $\square$ $\square$

Lemma 2.

$$\text{ctx}(A, B) \quad = \bigcup_{\substack{(A \rightarrow C\alpha) \in \mathcal{P} \\ C \xrightarrow[\text{rm}]{*} B\ldots}} \text{ctx}(C, B) \oplus \text{FIRST}(\alpha).$$

Proof. It suffices to show that

$$\text{ctx}(A, B) \quad \subseteq \bigcup_{\substack{(A \rightarrow C\alpha) \in \mathcal{P} \\ C \xrightarrow[\text{rm}]{*} B\ldots}} \text{ctx}(C, B) \oplus \text{FIRST}(\alpha),$$

since the reverse inclusion is a corollary of the previous lemma. We have:

$$\text{ctx}(C, B) \oplus \text{FIRST}(\alpha) = \begin{cases} \text{ctx}(C, B) \cup \text{FIRST}(\alpha), & \text{if } C \xrightarrow[\text{rm}]{*} B \\ \text{ctx}(C, B), & \text{otherwise.} \end{cases}$$

Therefore,

$$
\bigcup_{\substack{(A \to C\alpha) \in \mathcal{P} \\ C \xrightarrow[\text{rm}]{*} B\ldots}} \mathrm{ctx}(C,B) \oplus \mathrm{FIRST}(\alpha) \quad =
$$

$$
\bigcup_{\substack{(A \to C\alpha) \in \mathcal{P} \\ C \xrightarrow[\text{rm}]{*} B\ldots \\ C \neq A}} \mathrm{ctx}(C,B) \quad \cup \quad \underbrace{\mathrm{ctx}(A,B)}_{\text{if}(A \to A\alpha) \in \mathcal{P}} \quad \cup \quad \bigcup_{\substack{(A \to C\alpha) \in \mathcal{P} \\ C \xrightarrow[\text{rm}]{*} B}} \mathrm{FIRST}(\alpha).
$$

$$\square \qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$$

For the main theorem, we need the following definitions:

Definition 1. $(A,B)\,\mathsf{inherits\_ctx}\,(C,B)$ *if, and only if,* $(A \to C\alpha) \in \mathcal{P}$, $C \neq A$, *and* B *is a nonterminal such that* $C \xrightarrow[\text{rm}]{*} B \ldots$.

Definition 2. $\mathrm{ctx}_0(A,B) = \bigcup\{\mathrm{FIRST}(\alpha) \mid (A \to C\alpha) \in \mathcal{P},\, C \xrightarrow[\text{rm}]{*} B\}$.

Theorem 1. *In a grammar without useless productions or inaccessible symbols,* $\mathrm{ctx}(A,B) = \mathrm{ctx}_0(A,B) \cup \bigcup\{\mathrm{ctx}(C,B) \mid (A,B)\,\mathsf{inherits\_ctx}\,(C,B)\}$.

Proof. We first prove the reverse inclusion:

$$\forall (A \to C\alpha) \in \mathcal{P} \text{ s. t. } C \xrightarrow[\text{rm}]{*} B, \forall w \in \Sigma^* \text{ s. t. } \alpha \xrightarrow[\text{rm}]{*} w :$$

$$A \to C\alpha \xrightarrow[\text{rm}]{*} Cw \xrightarrow[\text{rm}]{*} Bw$$

$$\Rightarrow \quad \mathrm{FIRST}(w) \subseteq \mathrm{ctx}(A,B)$$

$$\Rightarrow \quad \mathrm{FIRST}(\alpha) = \bigcup\{\mathrm{FIRST}(w) \mid \alpha \xrightarrow[\text{rm}]{*} w\} \subseteq \mathrm{ctx}(A,B)$$

$$\Rightarrow \quad \mathrm{ctx}_0(A,B) = \bigcup\{\mathrm{FIRST}(\alpha) \mid (A \to C\alpha) \in \mathcal{P},\, C \xrightarrow[\text{rm}]{*} B\} \subseteq \mathrm{ctx}(A,B).$$

Furthermore, we have:

$$\forall (A \to C\alpha) \in \mathcal{P}, \forall t \in \mathrm{ctx}(C,B), \forall w \in \Sigma^* \text{ s. t. } \alpha \xrightarrow[\text{rm}]{*} w,$$

$$\exists v \in \Sigma^* :$$

$$C \xrightarrow[\text{rm}]{*} Btv$$

$$\Rightarrow \quad A \to C\alpha \xrightarrow[\text{rm}]{*} Cw \xrightarrow[\text{rm}]{*} Btvw$$

$$\Rightarrow \quad t \in \mathrm{ctx}(A,B)$$

$$\Rightarrow \quad \mathrm{ctx}(C,B) \subseteq \mathrm{ctx}(A,B)$$

$$\Rightarrow \quad \bigcup\{\mathrm{ctx}(C,B) \mid (A \to C\alpha) \in \mathcal{P},\, C \xrightarrow[\text{rm}]{*} B \ldots\} \subseteq \mathrm{ctx}(A,B)$$

$$\Rightarrow \quad \bigcup\{\mathrm{ctx}(C,B) \mid (A,B)\,\mathsf{inherits\_ctx}\,(C,B)\} \subseteq \mathrm{ctx}(A,B).$$

The forward inclusion is a consequence of the following reasoning. If $\mathrm{ctx}(A,B) = \varnothing$ the inclusion is trivial. Suppose then that $\mathrm{ctx}(A,B) \neq \varnothing$, and let $t \in \mathrm{ctx}(A,B)$. By the definition of ctx, $\exists w \in \Sigma^* : A \xrightarrow[\text{rm}]{*} Btw$. The first production applied in this derivation must begin with a nonterminal, given the presence of B as head of the sentential form. Let

$A \to C\alpha$ be this first applied production, so that $\exists z \in \Sigma^* : A \to C\alpha \xrightarrow[\text{rm}]{*} Cz \xrightarrow[\text{rm}]{*} Btw$. There are two possibilities:

(1) $z = tw \Rightarrow C \xrightarrow[\text{rm}]{*} B \wedge \alpha \xrightarrow[\text{rm}]{*} tw$. Thus, $\exists (A \to C\alpha) \in (P) : C \xrightarrow[\text{rm}]{*} B \wedge t \in \text{FIRST}(\alpha)$, hence $t \in \text{ctx}_0(A, B)$.

(2) $z \neq tw \Rightarrow \exists v \in \Sigma^* : w = vz$ (because z is a suffix of tw). Thus, $Cz \xrightarrow[\text{rm}]{*} Btvz$, that is, $\exists (A \to C\alpha) \in \mathcal{P} : C \xrightarrow[\text{rm}]{*} B \ldots \wedge t \in \text{ctx}(C, B)$, hence $t \in \bigcup \{\text{ctx}(C, B) \mid (A, B) \, \mathsf{inherits\_ctx} \, (C, B)\}$.

Remark: there must exist (except possibly if $A = B$) some production $A \to C\alpha$, $C \neq A$; otherwise, $A \xrightarrow[\text{rm}]{*} \!\!\!\!\!\! / \;\; B \ldots$. In this case, even if the first production applied in the derivation $A \xrightarrow[\text{rm}]{*} Btw$ is of form $A \to A\beta$, it is necessary that $A \to A\beta \xrightarrow[\text{rm}]{*} Au \to C\alpha u \xrightarrow[\text{rm}]{*} Cyu$, which reduces to the above case with $z = yu$. If $A = B \wedge \nexists C : (A \to C\alpha) \in \mathcal{P}, C \neq A$, then all productions of the derivation $A \xrightarrow[\text{rm}]{*} Atw$ are of form $A \to A\beta$. In this case, $\forall t \in \text{ctx}(A, A), \exists (A \to A\beta) \in \mathcal{P} : t \in \text{FIRST}(\beta)$, and thus $t \in \text{ctx}_0(A, B)$ since $A \xrightarrow[\text{rm}]{*} A$ trivially. $\qquad \square \qquad\qquad\qquad\qquad \square$

This theorem makes it possible to apply, *without any modification*, any transitive closure algorithm (but particularly that of [5]) to the relation $\mathsf{inherits\_ctx}$, so as to compute the values of $\text{ctx}(A, B)$. These ideas are captured in the following algorithm to compute lookahead sets $u \subseteq \Sigma$ for $LR(1)$ items of form $[A \to \alpha \cdot \beta, u]$, where $A \in \mathcal{N}$ and $(A \to \alpha\beta) \in \mathcal{P}$. Alternatively, one can use the computed $\text{ctx}(A, B)$ values in an $LALR(1)$ parser generator to achieve the efficiency gains reported in [14] and [8] over the results in [4].

Algorithm to compute lookahead sets:

(1) For every nonterminal C, compute all nonterminals A satisfying $C \xrightarrow[\text{rm}]{*} A \ldots$ and all nonterminals B satisfying $C \xrightarrow[\text{rm}]{*} B$.

(2) Build a relation $\mathsf{inherits\_ctx}$ between pairs of nonterminals: for every production $(A \to C\alpha) \in \mathcal{P}$ such that $C \neq A$, and every nonterminal B such that $C \xrightarrow[\text{rm}]{*} B \ldots$, define $(A, B) \, \mathsf{inherits\_ctx} \, (C, B)$.

(3) For all pairs (A, B) present in the above relation, initialize $\text{ctx}(A, B)$ to the union of $\text{FIRST}(\alpha)$ for all α such that $(A \to C\alpha) \in \mathcal{P}$ and $C \xrightarrow[\text{rm}]{*} B$.

(4) Compute $\text{ctx}(A, B)$ by taking the transitive closure of $\mathsf{inherits\_ctx}$.

(5) For each item $[A \to \alpha \cdot X\gamma, u]$ in the state being processed where X is a nonterminal, merge into the successor state all items of form $[N \to Y \cdot \beta, v]$, for each production $N \to Y\beta$ of each nonterminal N such that $X \xrightarrow[\text{rm}]{*} N \ldots$, where $v = \text{ctx}(X, N) \oplus \text{FIRST}(\gamma u)$.

We have implemented a full $LR(1)$ parser generator in Java using our lookahead set algorithm. The generated parsers are table-driven and use a suitable interpreter also written in Java. The parser of the generator itself was initially handwritten and then bootstrapped from its own grammar. The generated bytecode is about the same size as the handwritten recursive-descent parser, but the source code was reduced in half due to the use of tables. The bootstrap generation time is about 0.5 s on a Pentium III 1 GHz.

5. Conclusion

We have presented an efficient algorithm for $LALR(1)$ and $LR(1)$ lookahead computation, conceptually simpler than other existing algorithms and not requiring any alteration in conventional transitive closure algorithms. We have exhibited a formal proof of validity of our method and observed its effectiveness in practice.

Although the basic $LR(1)$ algorithm may be modified to eliminate unit reductions [7], it is still unclear whether our new algorithm can be used in such a modified parser construction. It is also an open problem how to extend our algorithm to construct $LR(k)$ automata for $k > 1$. We conjecture that this is possible but nontrivial, in a similar fashion to the extension of the state merging algorithm itself [16].

References

[1] H. Anzai. Almost boolean algebraic computation of $LALR(1)$ look-ahead set. *Journal of Information Processing*, 14(1):1–15, 1991.

[2] J. C. Beatty. On the relationship between the $LL(1)$ and $LR(1)$ grammars. *Journal of the ACM*, 29(4):1007–1022, 1982.

[3] R. Corchuelo, J. A. Pérez, A. Ruiz, and M. Toro. Repairing syntax errors in LR parsers. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 24(6):698–710, 2002.

[4] F. DeRemer and T. Pennello. Efficient computation of $LALR(1)$ look-ahead sets. *ACM Transactions on Programming Languages and Systems*, 4(4):615–649, 1982.

[5] J. Eve and R. Kurki-Suonio. On computing the transitive closure of a relation. *Acta Informatica*, 8:303–314, 1977.

[6] S. Heilbrunner. A parsing automata approach to LR theory. *Theoretical Computer Science*, 15:117–157, 1981.

[7] S. Heilbrunner. Truly prefix-correct chain-free $LR(1)$ parsers. *Acta Informatica*, 22:499–536, 1985.

[8] F. Ives. Unifying view of recent $LALR(1)$ lookahead set algorithms. *SIGPLAN Notices*, 21(7):131–135, 1986.

[9] I.-S. Kim and K.-M. Choe. Error repair with validation in LR-based parsing. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 23(4):451–471, 2001.

[10] D. E. Knuth. On the translation of languages from left to right. *Information and Control*, 8:607–639, 1965.

[11] Bruce J. McKenzie, Corey Yeatman, and Lorraine de Vere. Error repair in shift-reduce parsers. *ACM Transactions on Programming Languages and Systems*, 17(4):672–689, 1995.

[12] S. Morimoto and M. Sassa. Yet another generation of $LALR$ parsers for regular right part grammars. *Acta Informatica*, 37(9):671–697, 2001.

[13] D. Pager. A practical general method for constructing $LR(k)$ parsers. *Acta Informatica*, 7:249–268, 1977.

[14] J. C. H. Park, K. M. Choe, and C. H. Chang. A new analysis of $LALR$ formalisms. *ACM Transactions on Programming Languages and Systems*, 7(1):159–175, 1985.

[15] M. Sassa, H. Ishizuka, and I. Nakata. Rie, a compiler generator based on a one-pass-type attribute grammar. *Software - Practice and Experience*, 25(3):229–250, 1995.

[16] L. Schmitz. On the correct elimination of chain productions from LR parsers. *International Journal of Computer Mathematics*, 15:99–116, 1984.

This report contains unpublished results obtained by the author in 1988–1989.

Universidade de São Paulo, Escola Politécnica, Av. Prof. Luciano Gualberto, tr. 3, n. 158, s. C1-46, BR 05508-900, São Paulo(SP), Brazil.

E-mail address: pbarreto@larc.usp.br