Atoms of regular languages

Hellis Tamm
Tallinn University of Technology

Stellenbosch, Oct 15, 2018

Main publications

- basic theory:
  - J. Brzozowski, H. Tamm: Theory of átomata (DLT 2011, TCS 2014)
- complexity:
  - J. Brzozowski, H. Tamm: Quotient complexities of atoms of regular languages (DLT 2012, IJFCS 2013)
  - S. Iván: Complexity of atoms, combinatorially (IPL 2016)
  - J. Brzozowski, G. Davies: Maximally atomic languages (AFL 2014)
  - J. Brzozowski: Towards a theory of complexity of regular languages (JALC 2018)
- minimal NFA:
  - H. Tamm: New interpretation and generalization of the Kameda-Weiner method (ICALP 2016)
  - H. Tamm, B. van der Merwe: Lower bound methods for the size of nondeterministic finite automata revisited (LATA 2017)
- generalization:
  - H. Tamm, M. Veanes: Theoretical aspects of symbolic automata (SOFSEM 2018)

Quotients and atoms

Let L be a regular language over an alphabet Σ.

The left quotient of a language L by a word w is the language $w^{-1}L = \{x \in \Sigma^* \mid wx \in L\}$.

Let $K_0, \ldots, K_{n-1}$ be the quotients of L.

An atom of L is any non-empty language of the form

$$\widetilde{K_0} \cap \widetilde{K_1} \cap \cdots \cap \widetilde{K_{n-1}},$$

where $\widetilde{K_i}$ is either K_i or $\overline{K_i}$.

Quotients and atoms

Let L be a regular language over an alphabet Σ.

The left quotient of a language L by a word w is the language $w^{-1}L = \{x \in \Sigma^* \mid wx \in L\}$.

Let $K_0, \ldots, K_{n-1}$ be the quotients of L.

An atom of L is any non-empty language of the form

$$\widetilde{K_0} \cap \widetilde{K_1} \cap \cdots \cap \widetilde{K_{n-1}},$$

where $\widetilde{K_i}$ is either K_i or $\overline{K_i}$.

- Any quotient K_i of L (including L itself) is a union of atoms.
- Atoms define a partition of Σ^*.
- Atoms are the classes of the left congruence of L (Iván 2016): for $x, y \in \Sigma^*$, x is equivalent to y if for every $u \in \Sigma^*$, $ux \in L$ if and only if $uy \in L$.

The átomaton

Let $K_0 = L$ be the initial quotient of L.
Let $A = \{A_0, \ldots, A_{m-1}\}$ be the set of atoms of L.

- An atom is initial if it has K_0 (rather than $\overline{K_0}$) as a term.
- Let $I_A \subseteq A$ be the set of initial atoms.
- An atom is final if it contains ε.
- There is exactly one final atom A_{m-1}.
- The átomaton of L is the NFA $\mathcal{A} = (A, \Sigma, \alpha, I_A, \{A_{m-1}\})$, where $A_j \in \alpha(A_i, a)$ if $A_j \subseteq a^{-1}A_i$.

Some properties of the átomaton

- The language accepted by $\mathcal{A}$ is L.
- The (right) language of state A_i of $\mathcal{A}$ is the atom A_i.
- The reverse automaton $\mathcal{A}^R$ of $\mathcal{A}$ is a minimal DFA for L^R.
- The determinized automaton $\mathcal{A}^D$ of $\mathcal{A}$ is a minimal DFA of L.
- If $\mathcal{D}$ is a minimal DFA of L, then $\mathcal{A}$ is isomorphic to $\mathcal{D}^{RDR}$.

Atomic automata

An NFA $\mathcal{N}$ is atomic if for every state q of $\mathcal{N}$, the right language of q is a union of some atoms of $L(\mathcal{N})$.

Let L be a regular language. Some examples of atomic automata:

- átomaton of L
- minimal DFA of L
- canonical residual NFA of L
- universal automaton of L

Brzozowski's Theorem and DFA Minimization

Theorem (Brzozowski, 1962). For an NFA $\mathcal{N}$ without empty states, if $\mathcal{N}^R$ is deterministic, then $\mathcal{N}^D$ is minimal.

Brzozowski's (double-reversal) DFA minimization:

Given a DFA $\mathcal{D}$ of L, the minimal DFA is obtained by $\mathcal{D}^{RDRD}$.

Works also, if $\mathcal{D}$ is replaced by an NFA.

Generalization of Brzozowski's Theorem

Theorem (Brzozowski and Tamm, 2011, 2014). For any NFA $\mathcal{N}$, $\mathcal{N}^D$ is minimal if and only if $\mathcal{N}^R$ is atomic.

Applications:

- A polynomial double-reversal DFA minimization algorithm (Vázquez de Parga, García, and López, 2013):

  Let $\mathcal{D}$ be a DFA with no unreachable states.
  The minimal DFA is obtained by $\mathcal{D}^{RARD}$, where A is an *atomization* algorithm (produces an atomic NFA).
- García, López, and Vázquez de Parga (2015) also showed a relationship between two main approaches for DFA minimization: partitioning of the states of a DFA, and the double-reversal method.

Quotient complexity of atoms

Quotient complexity = state complexity.

Let L have n quotients, $n \geqslant 1$.

Theorem (Brzozowski and Tamm, 2012, 2013).
For $n \geqslant 1$, the quotient complexity of the atoms with 0 or n complemented quotients is less than or equal to $2^n - 1$.
For $n \geqslant 2$ and r satisfying $1 \leqslant r \leqslant n-1$, the quotient complexity of any atom of L with r complemented quotients is less than or equal to

$$f(n,r) = 1 + \sum_{k=1}^{r} \sum_{h=k+1}^{k+n-r} \binom{n}{h} \binom{h}{k}.$$

Moreover, these bounds are tight.

Another proof for these results was suggested by Iván (2014).

Quotient complexities of atoms in language classes

- right, left and two-sided regular ideal languages (Brzozowski and Davies, 2015)
- prefix-closed, prefix-free, and proper prefix-convex regular languages (Brzozowski and Sinnamon, 2017)
- suffix-free languages (Brzozowski and Szykuła, 2017)
- bifix-free languages (Ferens and Szykuła, 2017)
- non-returning languages (Brzozowski and Davies, 2017)

Asymptotic behaviour of the quotient complexity of atoms was studied by Diekert and Walter (2015).

Maximally Atomic Languages

Brzozowski and Davies (2014) defined a new class of regular languages:

A language is maximally atomic if it has the maximal number of atoms, and if every atom has the maximal complexity.

Theorem (Brzozowski and Davies, 2014). Let L be a regular language with complexity $n \geqslant 3$, and let T be the transition semigroup of the minimal DFA of L. Then L is maximally atomic if and only if the subgroup of permutations in T is set-transitive and T contains a transformation of rank $n-1$.

Another proof for this result was presented by Iván (2014).

Finding a minimal NFA: Kameda-Weiner matrix

Reinterpretation of the Kameda-Weiner method of finding a minimal NFA of a language, in terms of atoms of the language (HT, 2016).

- Kameda and Weiner (1970) used minimal DFAs for a language L and its reverse L^R, to form a matrix, and based on the grids in this matrix, a minimal NFA was found.
- Trimmed minimal DFA $\mathcal{D}^T$ of L with a state set Q.
- By Brzozowski's theorem, $\mathcal{D}^{RDT}$ is trim minimal DFA of L^R with a state set $S \subseteq 2^Q \setminus \emptyset$.
- Form a matrix with rows corresponding to states q_i of $\mathcal{D}$, and columns, to states $S_j \in S$ of $\mathcal{D}^{RDT}$.
- The (i, j) entry is 1 if $q_i \in S_j$, and 0 otherwise.

Quotient-atom matrix

- We use $\mathcal{D}^{RDRT}$, the trim átomaton of L, instead of $\mathcal{D}^{RDT}$, since the state sets of these automata are the same.
- The states of the minimal DFA correspond to quotients, and the states of the átomaton correspond to atoms of L.
- Interpret rows of the matrix as quotients, and columns as atoms of L (exc. the empty quotient and the atom $\overline{K_0} \cap \cdots \cap \overline{K_{n-1}}$, if they exist).
- We call this matrix the quotient-atom matrix of L.
- Then the (i, j) entry is 1 if and only if $A_j \subseteq K_i$.

Grids and cover of the quotient-atom matrix

- A grid g of the matrix is the direct product $g = P \times R$ of a set P of quotients with a set R of atoms, such that every atom in R is a subset of every quotient in P.
- If $g = P \times R$ and $g' = P' \times R'$ are two grids, then $g \subseteq g'$ if and only if $P \subseteq P'$ and $R \subseteq R'$.
- A grid is maximal if it is not contained in any other grid.
- A cover is a set $G = \{g_0, \ldots, g_{k-1}\}$ of grids, such that every pair (K_i, A_j) with $A_j \subseteq K_i$ belongs to some grid g_i in G.

NFA minimization by the Kameda-Weiner method

Let f_G be the function that assigns to every non-empty quotient K_i, the set of grids $g = P \times R$ from a cover G, such that $K_i \in P$.

The constructed NFA is $\mathcal{N}_G = (G, \Sigma, \eta_G, I_G, F_G)$, where G is a cover consisting of (maximal) grids, $I_G = f_G(K_0)$ is the set of grids involving the initial quotient K_0, $g \in F_G$ if and only if $g \in f_G(K_i)$ implies that K_i is a final quotient, and $\eta_G(g, a) = \bigcap_{K_i \in P} f_G(a^{-1}K_i)$ for a grid $g = P \times R$ and $a \in \Sigma$.

It may be the case that $\mathcal{N}_G$ does not accept the language L.

A cover G is called legal if $L(\mathcal{N}_G) = L$.

To find a minimal NFA of a language L, the method tests the covers of the matrix in the order of increasing size to see if they are legal.

The first legal NFA is a minimal one.

Reinterpretation of the Kameda-Weiner method

Let R be a set of atoms and let $U(R) = \bigcup_{A_j \in R} A_j$.

Theorem

Let $G = \{g_0, \ldots, g_{k-1}\}$ be a cover consisting of maximal grids $g_i = P_i \times R_i$, and let $\mathcal{N}_G = (G, \Sigma, \eta_G, I_G, F_G)$ be the corresponding NFA, obtained by the Kameda-Weiner method. It holds that

- *$g_i \in I_G$ if and only if $U(R_i) \subseteq L$,*
- *$g_i \in F_G$ if and only if $\varepsilon \in U(R_i)$,*
- *$g_j \in \eta_G(g_i, a)$ if and only if $U(R_j) \subseteq a^{-1}U(R_i)$ holds, for any $g_i, g_j \in G$ and $a \in \Sigma$.*

We note that essentially the same approach to the Kameda-Weiner method which uses projections of grids, consisting of subsets of the state set of the DFA $\mathcal{D}^{RDT}$ (corresponding to sets of atoms), was presented by Champarnaud and Coulon (IJFCS, 2005).

Lower bound methods for the size of NFA

- We consider the following lower bound methods for the size of NFA:
  - fooling set technique
  - extended fooling set technique
  - biclique edge cover technique
- Lower bounds obtained by these methods are not necessarily tight; a minimal NFA may have more states than the obtained bound.
- Some classes of languages for which tight bounds can be achieved, are known.
- The class of regular languages for which the fooling set provides a tight bound, is known as the class of *biseparable languages*.
- The exact classes of languages for which the extended fooling set technique and the biclique edge cover technique provide tight bounds, are not known.

Fooling set techniques

- Fooling set technique (Glaister and Shallit, 1996):
  Let $L \subseteq \Sigma^*$ be a regular language, and suppose there exists a set of pairs $S = \{(x_i, y_i) \mid 1 \leqslant i \leqslant p\}$ such that
  (a) $x_i y_i \in L$, for $1 \leqslant i \leqslant p$, and
  (b) $x_i y_j \notin L$, for $1 \leqslant i, j \leqslant p$, $i \neq j$.
  Then any NFA accepting L has at least p states.
- Extended fooling set technique (Birget, 1992):
  (b') $x_i y_j \notin L$ or $x_j y_i \notin L$, for $1 \leqslant i, j \leqslant p$, $i \neq j$.

Extended fooling set technique may provide a better lower bound.

Lower bounds obtained by these techniques are not necessarily tight.

Biclique edge cover technique

Let $G = (X, Y, E)$ be a bipartite graph, with sets of vertices X and Y, and set of edges $E \subseteq X \times Y$.

A set $C = \{H_1, H_2, \ldots\}$ of bipartite subgraphs of G is an edge cover of G if every edge $e \in E$ is an edge of some H_i.

An edge cover C of G is a biclique edge cover if every H_i is a biclique, that is, if $H_i = (X_i, Y_i, E_i)$ with $E_i = X_i \times Y_i$.

The bipartite dimension of G, $d(G)$, is the size of the smallest biclique edge cover of G if it exists and is infinite otherwise.

The biclique edge cover technique (Gruber and Holzer, 2006):

Theorem

Let $L \subseteq \Sigma^*$ be a regular language, let $X, Y \subseteq \Sigma^*$.
Suppose there exists a bipartite graph $G = (X, Y, E_L)$, where
for $x \in X$ and $y \in Y$, $(x, y) \in E_L$ if and only if $xy \in L$.
Then any NFA accepting L has at least $d(G)$ states.

Dependency graph of a language

Nerode right congruence is well known:
for $x, y \in \Sigma^*$, $x \equiv_L y$ if for every $v \in \Sigma^*$, $xv \in L$ if and only if $yv \in L$.

The left congruence is defined:
for $x, y \in \Sigma^*$, $x_L{\equiv}y$ if for every $u \in \Sigma^*$, $ux \in L$ if and only if $uy \in L$.

Gruber and Holzer (2006) defined the dependency graph of a language L as the bipartite graph $G_L = (X, Y, E_L)$, where $X = \Sigma^*/\equiv_L$ and $Y = \Sigma^*/_L{\equiv}$, and $([x]_L, {}_L[y]) \in E_L$ if and only if $xy \in L$.

They suggested that the maximal fooling sets and extended fooling sets, as well as the smallest biclique edge cover for L, can be found by inspecting the dependency graph G_L.

Quotient-atom graph of a language

Dependency graph of L was defined as $G_L = (X, Y, E_L)$, where $X = \Sigma^*/\equiv_L$ and $Y = \Sigma^*/_L\equiv$, and $([x]_L, {}_L[y]) \in E_L$ if and only if $xy \in L$.

Classes of $\equiv_L$ correspond to the quotients of L.

Classes of ${}_L\equiv$ are the atoms of L (Iván 2016).

We can define G_L in terms of quotients and atoms of L:

Let $K = \{K_1, \ldots, K_n\}$ be the set of quotients of L, and let $A = \{A_1, \ldots, A_m\}$ be the set of atoms of L.

Proposition

For any $x, y \in \Sigma^*$, $xy \in L$ if and only if $A_j \subseteq K_i$, where $y \in A_j$ and $K_i = x^{-1}L$.

We can express $G_L = (K, A, E_L)$, with $(K_i, A_j) \in E_L$ if and only if $A_j \subseteq K_i$.

With this view, we call G_L the quotient-atom graph of L.

Lower bound methods in terms of quotients and atoms

By Gruber and Holzer (2006), maximal fooling sets and extended fooling sets, as well as the smallest biclique edge cover for L, can be found by inspecting the dependency graph G_L, that is, the quotient-atom graph of L.

Consequently, we can express the above mentioned lower bound methods in terms of quotients and atoms.

The biclique edge cover technique can be presented by the following theorem:

Theorem

Let $L \subseteq \Sigma^*$ be a regular language, and let the quotient-atom graph of L be $G_L = (K, A, E_L)$, with $(K_i, A_j) \in E_L$ if and only if $A_j \subseteq K_i$. Then any NFA accepting L has at least $d(G_L)$ states.

Fooling set methods in terms of quotients and atoms

The fooling set technique and the extended fooling set technique can be expressed as the first and the second case, respectively, of the following theorem:

Theorem

Let $L \subseteq \Sigma^*$ be a regular language, and suppose there exists a set of quotient-atom pairs $S = \{(K_i, A_i) \mid 1 \leqslant i \leqslant p\}$ such that either

1. *(a) $A_i \subseteq K_i$ for $1 \leqslant i \leqslant p$,*
   (b) $A_i \not\subseteq K_j$ for $1 \leqslant i, j \leqslant p$ and $i \neq j$,
   or
2. *(a) $A_i \subseteq K_i$ for $1 \leqslant i \leqslant p$,*
   (b) $A_i \not\subseteq K_j$ or $A_j \not\subseteq K_i$ for $1 \leqslant i, j \leqslant p$ and $i \neq j$,

holds. Then any NFA accepting L has at least p states.

Conclusions

- We have introduced a natural set of languages – the atoms – that are defined by every regular language.
- We defined a unique NFA for every regular language, the átomaton, and related it to other known concepts.
- We characterized the class of NFAs for which the subset construction yields a minimal DFA.
- We introduced a new complexity measure for regular languages: the quotient complexity of atoms.
- We showed that atoms of regular languages have an important role in finding a minimal NFA.
- We presented the lower bound methods for the size of NFA in terms of quotients and atoms of the language.

Thanks