正規表現と Büchi の定理

Sakaé Fuchino (渕野 昌)

Graduate School of System Informatics
Kobe University
(神戸大学大学院 システム情報学研究科)

https://fuchino.ddo.jp/

情報基礎特論 2015 年前期講義

(23. Juni 2022 (01:19 JST) version)

June 1, 2015

This presentation is typeset by pLaTeX with beamer class.

正規表現と有限オートマトン

- 空でない記号の有限集合 (アルファベット) S に対し，S^* で S の記号の有限列の全体をあらわす．

定理 1. S をアルファベットとするとき，$L \subseteq S^*$ に対して次は同値である:

(a) ある正規表現 r に対し，
$L = L^*(r) = \{\sigma \in S^* : \sigma$ は r にマッチする $\}$ となる．

(b) ある非決定性有限オートマトン A に対し，
$L = L^*(A) = \{\sigma \in S^* : A$ は σ を受理する $\}$ となる．

(c) ある (決定性) 有限オートマトン A' に対し，
$L = L^*(A') = \{\sigma \in S^* : A'$ は σ を受理する $\}$ となる．

- 上の定理の証明は，特に，(a) を満たすような σ が与えられたとき，それに対応する A を作り，A から A' を作るアルゴリズムを与えるものとなっている．

- この応用として，lex (flex), 正規表現のコンピュータへの実装，SKK などがある．

正規表現

- アルファベット S に対し，S に属さない記号 '$|$', '$*$', '(', ')' を，S に付け加えて得られる言語を $S+$ と呼ぶことにする．

- $s \in S+^*$ が **正規表現** ということを以下のように定義する:

  (1) 各 $s \in S^*$ は正規表現である． (2) $s \in S+^*$ が正規表現なら，$(s)*$ は正規表現である． (3) s, t が正規表現なら，st も正規表現である． (4) s, t が正規表現なら，$(s|t)$ も正規表現である．
  (5) 以上のみ．

- 文字列 $u \in S^*$ が **正規表現 s にマッチする** ということを以下のように再帰的に定義する:

  (1) $s \in S^*$ なら，u が s にマッチするのは $u = s$ のときのみである． (2) $s = (t)*$ なら，u が s にマッチするのは，t にマッチする文字列 $u_1, \ldots, u_n$ $(n = 0, 1, \ldots)$ があって，$u = u_1 \cdots u_n$ となるときである ($n = 0$ のときには，u は空列). (3) $s = s_0 s_1$ なら，

u が s にマッチするのは:

  (4) $s = (s_0|s_1)$ なら，**u が s にマッチするのは:**

  (5) 以上のみ．

有限オートマトン

- **(決定性／非決定性) 有限オートマトン** は (決定性／非決定性) チューリングマシンに以下のような制限を加えたものである.

  - テープは読みとり専用である.
  - 各ステップごとにヘッドは右に 1 コマ動く (止まったり左に動いたりしない).

- (非決定性) 有限オートマトン A が 記号列 s を受理するのは, s の書きこまれたテープの s の最後の記号の書かれたマスの右にヘッドが移動したときに, A 内部状態が受理状態になっている (受理状態になっているような遷移が存在する) ときである.

正規表現と有限オートマトン

- 空でない記号の有限集合 (アルファベット) S に対し，S^* で S の記号の有限列の全体をあらわす.

定理 1. S をアルファベットとするとき，$L \subseteq S^*$ に対して次は同値である:

(a) ある正規表現 r に対し，
$L = L^*(r) = \{\sigma \in S^* : \sigma$ は r にマッチする $\}$ となる.

(b) ある非決定的有限オートマトン A に対し，
$L = L^*(A) = \{\sigma \in S^* : A$ は σ を受理する $\}$ となる.

(c) ある (決定的) 有限オートマトン A' に対し，
$L = L^*(A')\{\sigma \in S^* : A'$ は σ を受理する $\}$ となる.

- S^* の要素を **語** (word) とよぶ. ▶ S^* の部分集合を **言語** とよぶ (この言葉の使い方は後出の論理での “言語” とは異る!).
- 言語 $L \subseteq S^*$ は上の定理の (a), (b), (c) (のどれか，またはすべて) を満たすとき，**正規言語** (正則言語 regular language) とよばれる.

有限構造としての語

- S をアルファベットとして，$\sigma \in S^*$ とするとき．$\sigma = s_0 \cdots s_{\ell-1}$ として，有限構造 (文字列構造) $\mathfrak{A}_\sigma$ を

  $$\mathfrak{A}_\sigma = \langle \{0, \ldots, \ell - 1\}, <^{\mathfrak{A}_\sigma}, {R_a}^{\mathfrak{A}_\sigma} \rangle_{a \in S}$$

  のこととする．ただし $<^{\mathfrak{A}_\sigma}$ は数 $0, \ldots, \ell - 1$ 上の通常の大小関係，${R_a}^{\mathfrak{A}_\sigma}$ は $\{0, \ldots, \ell - 1\}$ 上の 1 変数関係で，

  $${R_a}^{\mathfrak{A}_\sigma}(k) \Leftrightarrow s_k = a$$

  となるものとする．

- この構造の (述語論理の意味での) 言語 $\{<, R_a\}_{a \in S}$ を $\mathcal{L}_S$ とよぶことにする．

- 述語論理の $\mathcal{L}_S$-文 φ に対し，$L^*(\varphi) = \{\sigma \in S^* : \mathfrak{A}_\sigma \models \varphi\}$ とする．

例 1. $S = \{a, b\}$ とする．φ を $\exists x \exists y\ ((x < y) \wedge R_a(x) \wedge R_b(y))$ とすると，$L^*(\varphi) = S^* a S^* b S^*$ である．

有限構造としての語

- S をアルファベットとして，$\sigma \in S^*$ とするとき．$\sigma = s_0 \cdots s_{\ell-1}$ として，有限構造 (文字列構造) $\mathfrak{A}_\sigma$ を
$$\mathfrak{A}_\sigma = \langle \{0, \ldots, \ell - 1\}, <^{\mathfrak{A}_\sigma}, {R_a}^{\mathfrak{A}_\sigma} \rangle_{a \in S}$$
のこととする． ただし $<^{\mathfrak{A}_\sigma}$ は数 $0, \ldots, \ell - 1$ 上の通常の大小関係，${R_a}^{\mathfrak{A}_\sigma}$ は $\{0, \ldots, \ell - 1\}$ 上の 1 変数関係で，
$${R_a}^{\mathfrak{A}_\sigma}(k) \Leftrightarrow s_k = a$$
となるものとする．

- この構造の (述語論理の意味での) 言語 $\{<, R_a\}_{a \in S}$ を $\mathcal{L}_S$ とよぶことにする．

- 述語論理の $\mathcal{L}_S$-文 φ に対し，$L^*(\varphi) = \{\sigma \in S^* : \mathfrak{A}_\sigma \models \varphi\}$ とする．

演習問題 1． $S = \{a, b\}$ とする． このとき，$L^*(\varphi) = (ab)^*$ となるような $\mathcal{L}_S$-文 φ を求めよ．

正規言語と述語論理

定理 2 (Büchi の定理 (1960) の一部). S をアルファベットとして φ を述語論理の $\mathcal{L}_S$-文とするとき，$L^*(\varphi)$ は正規言語である．

証明. φ の構成に関する帰納法による． □

命題 2. $S = \{a\}$ として，$(aa)^*$ は $L^*(\varphi)$ の形で表わせない．

証明. 各自然数 k に対し，$\bar{a}_k = \underbrace{aa \cdots a}_{k\,個}$ に対応する文字列構造を $\mathfrak{A}_k$ とあらわすことにする．

- $\bar{a}_k \in (aa)^* \Leftrightarrow k$ は偶数 である．
- Ehrenfeucht-Fraïssé の定理を応用すると，任意の自然数 n に対し十分に大きな k をとると すべての $k', k'' > k$ に対し，$\mathfrak{A}_{k'} \cong_{\Sigma^n} \mathfrak{A}_{k''}$ が成り立つ，ことが示せる．
- このことから，任意の $\mathcal{L}_S$-論理式 φ に対して $L^*(\varphi)$ が $(aa)^*$ と一致しないことがわかる． □

正規言語と述語論理

- 定理 2. と命題 3. から

  命題論理の文 φ により $L^*(\varphi)$ と表わせる言語の全体 $\subsetneq$ 正規言語の全体

  である．実は次が成り立つ:

> **定理 4. (McNaughton and Papert, 1977)** 言語 $L \subseteq S^*$ が述語論理の文 φ によって $L = L^*(\varphi)$ と表わせるのは，L が star-free な言語であるちょうどそのときである．

- アルファベット S に対する $\mathcal{F} \subseteq \mathcal{P}(S^*)$ が **star free な言語の全体**であるとは，
  - ▷ $\emptyset, \{\not{\emptyset}\} \in \mathcal{F}$
  - ▷ すべての $s \in S$ に対し，$\{s\} \in \mathcal{F}$
  - ▷ $\mathcal{F}$ は (集合の) ブール演算に関して閉じている
  - ▷ $L, L' \in \mathcal{F}$ なら $L ^\frown L' = \{\sigma ^\frown \sigma' : \sigma \in L, \sigma' \in L'\} \in \mathcal{F}$

  を満たすような最小の族となっていることである．
- $\mathcal{F}$ をある S に対する star free な言語の全体として，$L \in \mathcal{F}$ となるとき，L は **star free な言語** である，という．

正規言語と monadic な 2 階の論理

- 述語論理を以下のようにして拡張する:
- 新しい (2 階の monadic な) 変数記号 X, Y,... を用意しておく.
- $\mathcal{L}$-論理式の生成規則に次の規則を加える:
  - t が $\mathcal{L}$-項で X が 2 階の monadic な変数記号なら, $X(t)$ は論理式である.
  - φ が論理式で X が monadic な変数記号なら, $\exists X\,\varphi$, も論理式である.
- $\mathcal{L}$-構造 $\mathfrak{A} = \langle A, \ldots\rangle$ で, monadic な変数記号は A の部分集合と解釈する.
- $t = t(x_0, \ldots, x_{k-1})$ を $\mathcal{L}$-項として X を monadic な変数記号とするとき, $U \subseteq A$, $a_0, \ldots, a_{k-1} \in A$ として, $\mathfrak{A} \models X(t)[U, a_0, \ldots, a_{k-1}]$ $\Leftrightarrow$ $t^{\mathfrak{A}}(a_0, \ldots, a_{k-1}) \in U$ とする.
- $\mathfrak{A} \models \exists X \varphi(X, \ldots) \Leftrightarrow$ ある $U \subseteq A$ に対し, $\mathfrak{A} \models \varphi(U, \ldots)$ とする.

Büchi の定理

定理 5. (Büchi, 1960) 言語 $L \subseteq S^*$ が正規言語となるのは，monadic な 2 階論理の $\mathcal{L}_S$-文 φ で，$L = L^*(\varphi) = \{\sigma \in S^* : \mathfrak{A}_\sigma \models \varphi\}$ となるものがある丁度そのときである．

証明. ▶ $L = L^*(\varphi)$ なら L は正規言語となることの証明は φ の構成に関する帰納法で示せる．

- ある決定性有限オートマトン A に対し，$L = L^*(A)$ となっているとして，$\mathfrak{s}_0, \ldots, \mathfrak{s}_\ell$ が A の内部状態のすべてで，$\mathfrak{s}_0$ が初期状態 $\mathfrak{s}_\ell$ が停止状態とする．
- monadic な 2 階算術の $\mathcal{L}_S$-文 φ を $\exists X_{\mathfrak{s}_0}\exists X_{\mathfrak{s}_1}\cdots\exists X_{\mathfrak{s}_\ell}\ (\cdots)$ とする．ただし，$(\cdots)$ はこの文を $\mathfrak{A}_\sigma$, $\sigma = s_0 \cdots s_{n-1}$ で解釈したとき，

  > $X_{\mathfrak{s}_0}, \ldots, X_{\mathfrak{s}_\ell}$ は $\{0, \ldots, n-1\}$ の分割で，A を σ の上で走らせたとき $X_{\mathfrak{s}_k}(i)$ なら，A のヘッドが i 番目の記号に移動したときの内部状態は k で，$X_{\mathfrak{s}_0}(0)$, $X_{\mathfrak{s}_\ell}(n-1)$ である

  を表現するものとする．
- このとき，$L = L^*(\varphi)$ である． □

参考文献等

- Jean-Eric Pin, Logic on words, Bulletin of the European Association of Theoretical Computer Science 54 (1994), 145–165.

- 有限モデルを用いた $P \neq NP$ 問題へのアプローチも可能に思われる.

- 理論コンピュータ科学の人類への貢献度に関する認識は文化圏で大きく異るように思われる．たとえば,

  https://en.wikipedia.org/wiki/20th_century#Science_and_mathematics

  https://ja.wikipedia.org/wiki/20%E4%B8%96%E7%B4%80#.E7.A7.91.E5.AD.A6.E3.83.BB.E6.8A.80.E8.A1.93