数理解析研究所講究録
960巻 1996年 71-76

Height Functions and Formal Languages

京都産業大学理学部　伊藤　正美（ Masami Ito ）

Let X be an alphabet and let X^* be the free monoid generated by X. We assume the cardinality of X is greater than 1. By X^+ we denote the free semigroup generated by X. Any element of X^* is called a *word* over X and any subset of X^* is called a *language* over X. Let $u \in X^*$. By $|u|$ we denote the *length* of u and by $Sub(u)$ we denote the language $\{v \in X^* \mid u = xvy$ for some $x, y \in X^*\}$. Moreover, let $L \subseteq X^*$. By $Sub(L)$ we denote the language $\cup_{u \in L} Sub(u)$. A language L is said to be *dense* if $Sub(L) = X^*$. In several references [2 - 5], dense languages and their properties have been studied. However, the above concept of the density seems to be rather rough from the point of view of the classification of languages. Introducing height functions, we proposed a new classification of dense and non-dense languages in [1,2].

Consider a total order $\leq$ on X^* satisfying the following condition (*):

If $|u| < |v|$, then $u < v$ where $u < v$ means that $u \leq v$ and $u \neq v$.

Let $M \subseteq X^*$. By $min(M)$ we denote the minimal element in M according to the total order $\leq$. Let $u \in X^*$. By $Height_{\leq}(u)$, we denote the word $min(X^* - Sub(u))$. Now we are ready to define a height function. Let $L \subseteq X^*$. Then $Height_{\leq}(L) = \{Height_{\leq}(u) \mid u \in L\}$. The following result is fundamental.

Proposition 1 *Let $L \subseteq X^*$. Then L is dense if and only if $Height_{\leq}(L)$ is infinite.*

First we show that any height function preserves the class of regular languages.

Lemma 1 *Let $|X| = 2$, let $a, b \in X$, $a \neq b$, and let $L \subseteq X^*$ be a regular language. Then there exists a positive integer N such that if $t > N$ and $ab^t \in Height_{\leq}(L)$ then $ab^{t-3} \in Height_{\leq}(L)$.*

Proof. Let $A = (D, X, d_0, \delta, W)$ be a finite automaton which accepts L where D is the set of states. Let $|D| = r$. It can be easily verified that there

exixt the following positive integers m, $3 \le m \le r^r$ and k, $k \ge 1$: $\delta(d, b^{k+m}) = \delta(d, b^k)$ for any $d \in D$.

Let $N = k + m$. Now assume that $ab^t \in Height_{\le}(L)$ and $t > N$. Since $ab^t \in Height_{\le}(L)$, $ab^t = Height_{\le}(b^{t+p}f_1f_2 \cdots f_q)$ and $p \ge 0$, $b^{t+p}f_1f_2 \cdots f_q \in L$ where $f_i \in \{a, ab, ab^2, ab^3, \ldots, ab^{t-5}, ab^{t-4}, ab^{t-3}, ab^{t-2}, ab^{t-1}\}$, $1 \le i \le q$. Now we define the mapping ϕ as follows: $\phi(a) = a$, $\phi(ab) = ab$, $\phi(ab^2) = ab^2$, $\phi(ab^3) = ab^3$, ... , $\phi(ab^{t-5}) = ab^{t-5}$, $\phi(ab^{t-4}) = ab^{t-4}$, $\phi(ab^{t-3}) = ab^{t-3-m}$, $\phi(ab^{t-2}) = ab^{t-2-m}$, and $\phi(ab^{t-1}) = ab^{t-1-m}$. Now consider $b^{t+p}\phi(f_1)\phi(f_2) \cdots \phi(f_q)$. By definition, $\delta(d_0, b^{t+p}f_1f_2 \cdots f_q) = \delta(d_0, b^{t+p}\phi(f_1)\,\phi(f_2) \cdots \phi(f_q)) \in W$, i.e. $b^{t+p}\phi(f_1)\phi(f_2) \cdots \phi(f_q) \in L$. Now let $u < ab^{t-3}$. Then obviously $aua < ab^t$. Since $|u| \le t - 2$, u can be represented as follows: (1) $u = b^i$, $0 \le i \le t - 2$, (2) $u = b^ja$, $0 \le j \le t - 3$ and (3) $u = a^ib^jav$, $0 \le i, j \le t - 4$ and $v = \cdots b^s \cdots \Rightarrow s \le t - 4$.

It is obvious that in (1) and (2) u is a subword of $b^{t+p}\phi(f_1)\,\phi(f_2) \cdots \phi(f_q)$. Consider the case (3). Since $aua < ab^t$, aua is a subword of $b^{t+p}f_1f_2 \cdots f_q$. Remark that $aua = \cdots b^s \cdots \Rightarrow s \le t - 4$. This means that aua is a subword of $b^{t+p}\phi(f_1)\phi(f_2) \cdots \phi(f_q)$ and hence u is so. On the other hand, ab^{t-3} is not a subword of $b^{t+p}\phi(f_1)\phi(f_2) \cdots \phi(f_q)$. Consequently, $ab^{t-3} = Height_{\le}(b^{t+p}\phi(f_1)\,\phi(f_2) \cdots \phi(f_q))$, i.e. $ab^{t-3} \in Height_{\le}(L)$. This completes the proof of the lemma.

Dually, we can prove the following.

Lemma 2 *Let $|X| = 2$, let $a, b \in X$, $a \ne b$, and let $L \subseteq X^*$ be a regular language. Then there exists a positive integer N such that if $s > N$ and $a^sb \in Height_{\le}(L)$ then $a^{s-3}b \in Height_{\le}(L)$.*

Theorem 1 *Let $L \subseteq X^*$ be a regular language. Then $Height_{\le}(L)$ is a regular language, too. Moreover, if L is dense, then $Height_{\le}(L)$ is dense.*

Proof. Let $|X| = 1$ and let $X = \{a\}$. Then $Height_{\le}(L) = aL$ and $Height_{\le}(L)$ is regular. Now consider the case $|X| \ge 3$. If L is not dense, then $Height_{\le}(L)$ is finite and hence regular. Assume that L is dense. Let $A = (V, X, v_0, \delta, W)$ be a finite automaton which accepts L where V is the set of states. Let $|V| = r$. Since L is dense, for any $w \in X^*$ there exist $\alpha, \beta \in X^*$ such that $|\alpha|, |\beta| \le r$ and $\alpha w\beta \in L$. Now let $u \in X^+$ with $|u| > r$. Then $u = avb$ for some $a, b \in X$ where $v \in X^*$. Let $a, b \ne c \in X$. Now let $u_1, u_2, \cdots, u_n$ be the strings in X^* such that $u_i < u$, $1 \le i \le n$, and construct the strings

$$w = c^{|u|}u_1c^{|u|}u_2c^{|u|} \cdots c^{|u|}u_nc^{|u|}.$$

From the above remark, there exist $\alpha, \beta \in X^*$ such that $|\alpha|, |\beta| \leq r$ and $\alpha w \beta = \alpha c^{|u|} u_1 c^{|u|} u_2 c^{|u|} \cdots c^{|u|} u_n c^{|u|} \beta \in L$. Then it can easily be seen that $u \notin Sub(\alpha w \beta)$, $u = Height_{\leq}(\alpha w \beta)$ and hence $u \in Height_{\leq}(L)$. This means that $Height_{\leq}(L) = X^+ - F$ where $F \subseteq X^*$ is a finite language. Obviously, $Height_{\leq}(L)$ is a dense regular language.

Now consider the case $|X| = 2$. Again, it is enough to assume that L is dense and also we consider the same finite automaton as above. Let $u \in X^*$ and $|u| > r$.

(i) $u = avb$ for some $a \in X, v \in X^*$. Let $b \in X - \{a\}$ and $u_1, u_2, \cdots, u_n$ be the the strings in X^* such that $u_i < u, 1 \leq i \leq n$. Construct the string

$$w = b^{|u|} u_1 b^{|u|} u_2 b^{|u|} \cdots b^{|u|} u_n b^{|u|}.$$

As has been stated, there exist $\alpha, \beta \in X^*$ such that $|\alpha|, |\beta| \leq r$ and $\alpha w \beta = \alpha b^{|u|} u_1 b^{|u|} u_2 b^{|u|} \cdots b^{|u|} u_n b^{|u|} \beta \in L$. In this case, we have $u \notin Sub(\alpha w \beta)$, $u = Height_{\leq}(\alpha w \beta)$ and hence $u \in Height_{\leq}(L)$.

(ii) $u = avb$ for some $a, b \in X, a \neq b, v \in X^*$ and $v \notin a^*b^*$. Let $u_1, u_2, \cdots, u_n$ be the the strings in X^* such that $u_i < u$, $1 \leq i \leq n$, and construct the strings

$$w = a^{|u|} b^{|u|} u_1 a^{|u|} b^{|u|} u_2 a^{|u|} b^{|u|} \cdots a^{|u|} b^{|u|} u_n a^{|u|} b^{|u|}.$$

There exist $\alpha, \beta \in X^*$ such that $|\alpha|, |\beta| \leq r$ and $\alpha w \beta = \alpha a^{|u|} b^{|u|} u_1 a^{|u|} b^{|u|} u_2 a^{|u|} b^{|u|} \cdots a^{|u|} b^{|u|} u_n a^{|u|} b^{|u|} \beta \in L$. In this case, we have also $u \notin Sub(\alpha w \beta)$, $u = Height_{\leq}(\alpha w \beta)$ and hence $u \in Height_{\leq}(L)$.

(iii) $u = a^t b^s$, $a \neq b$ and $t, s \geq 1$. If $t \geq 2$ and $s \geq 2$, then we proceed as above: let $u_1, u_2, \cdots, u_n$ be the the strings in X^* with $u_i < u, 1 \leq i \leq n$, and construct the string

$$w = (ab)^{|u|} w_1 (ab)^{|u|} w_2 (ab)^{|u|} \cdots (ab)^{|u|} w_n (ab)^{|u|}.$$

There exist $\alpha, \beta \in X^*$ such that $|\alpha|, |\beta| \leq r$ and $\alpha w \beta \in L$, i.e.

$$\alpha (ab)^{|u|} w_1 (ab)^{|u|} w_2 (ab)^{|u|} \cdots (ab)^{|u|} w_n (ab)^{|u|} \beta \in L.$$

We have $u \notin Sub(\alpha w \beta)$ and hence $u = Height_{\leq}(\alpha w \beta)$, i.e. $u \in Height_{\leq}(L)$.

(iv) The case that $\{ab^t \in Height_{\leq}(L) \mid t \geq 1\}$ is infinite. In this case, by Lemma 1 there exists a set of positive integers T consisting of at most 3 elements such that $\{ab^t \in Height_{\leq}(L) \mid t \geq 1\} = (\cup_{t \in T} \{ab^{t+3k} \mid k \geq 0\}) \cup G$ where G is a finite set.

(v) The case that $\{a^s b \in Height_{\leq}(L) \mid s \geq 1\}$ is infinite. In this case, by Lemma 2 there exists a set of positive integers S consisting of at most 3 elements such that $\{a^s b^t \in Height_{\leq}(L) \mid s \geq 1\} = (\cup_{s \in S} \{a^{s+3k} b \mid k \geq 0\}) \cup H$ where H is a finite set.

Summing up the results, we have the following:

Let $X = \{a, b\}$ and let $L \subseteq X^*$ be a dense regular language. Then $Height_{\leq}(L)$ can be represented as follows:

$$\begin{aligned} Height_{\leq}(L) \;=\; & [X^+ - (F \cup a^*b^* \cup b^*a^*)] \cup (\cup_{t \in T} \{ab^{t+3k} \mid k \geq 0\}) \\ & \cup (\cup_{t' \in T'} \{ba^{t'+3k} \mid k \geq 0\}) \cup (\cup_{s \in S} \{a^{s+3k}b \mid k \geq 0\}) \\ & \cup (\cup_{s' \in S'} \{b^{s'+3k}a \mid k \geq 0\}) \cup J \end{aligned}$$

where F and J are finite sets and T, T', S, S' are sets of positive integers consisting of at most of 3 elements.

Obviously, $Height_{\leq}(L)$ is a dense regular language. This completes the proof of the theorem.

However, the same type of result does not hold true for the case of context-sensitive languages.

Theorem 2 *Let $\leq$ be a lexicographic order on X^* satisfying the condition (*). Then there exists a context-sensitive language $L \subseteq X^*$ such that $Height_{\leq}(L)$ is not context-sensitive.*

We will discuss whether the same type of result as in Theorem 1 holds true for the case of linear languages.

Let $X = \{a, b, ...\}$. Moreover, let $1 = f(1) < f(2) < ... < f(k) < f(k+1) < ...$ be a sequence of positive integers and let $\leq$ be a total order on X^* satisfying the following conditions (**):

(1) $|u| < |v| \Rightarrow u < v$,

(2) $b^{f(k)} < a^{f(k)}$ for $k, k = 1, 2, 3, ...$,

(3) $a^i < u$ for any $u, u \neq a^i$, $|u| = i$ and $i \neq f(k), k = 1, 2, 3, ...$.

We consider the language $L = \{w\phi(\rho(w)) \mid w \in X^+\}$ where $\rho(w) = a_n a_{n-1} ... a_1$ for $w = a_1 a_2 ... a_n$, $a_i \in X$ and ϕ is the isomorphism of X^* onto X^* such that $\phi(a) = b$, $\phi(b) = a$ and $\phi(c) = c$ for any $c \in X - \{a, b\}$. Then L is a linear language whose grammar is $G = (V, X, S, P)$ where $P = \{S \to xS\phi(x), S \to x\phi(x) \mid x \in X\}$.

Lemma 3 *If $i = f(k)$ for some $k, k = 1, 2, 3, ...$, then $a^i \notin Height_{\leq}(L)$.*

Proof. Let $i = f(k)$ for some k, $k = 1, 2, 3, ...$. Assume $a^i \in Height_{\leq}(L)$. Then there exists $w \in X^+$ such that $a^i = Height_{\leq}(w\phi(\rho(w)))$. Since $b^{f(k)} < a^{f(k)}$,

$w\phi(\rho(w))$ contains $b^{f(k)}$ as a subword. Then we have the following three cases.

Case 1. w contains $b^{f(k)}$ as a subword. In this case, $\phi(\rho(w))$ contains $a^{f(k)}$ as a subword and hence $w\phi(\rho(w))$ contains $a^{f(k)}$ as a subword. This contradicts the assumption that $a^i = Height_{\leq}(w\phi(\rho(w))$.

Case 2. $\phi(\rho(w))$ contains $b^{f(k)}$ as a subword. In this case, w contains $a^{f(k)}$ as a subword and hence $w\phi(\rho(w))$ contains $a^{f(k)}$ as a subword. This contradicts the assumption that $a^i = Height_{\leq}(w\phi(\rho(w))$.

Case 3. $b^i = b^{p+q}$, $p, q \geq 1$, b^p is a suffix of w and b^q is a prefix of $\phi(\rho(w))$. However, this yields a contradiction because in this case a^p must be a prefix of $\phi(\rho(w))$.

In either case, we have a contradiction. Hence $a^i \notin Height_{\leq}(L)$.

Lemma 4 *If $i \neq f(k)$ for any k, $k = 1, 2, 3, \ldots$, then $a^i \in Height_{\leq}(L)$.*

Proof. Let $\{w_1, w_2, \ldots, w_r\} = \{w \in X^* \mid w < a^i\}$. By (3) in (**), $|w_t| \leq i - 1$ for any t, $t = 1, 2, 3, \ldots, r$. We can assume that $w_1 = a^{i-1}$ and $w_r = b^{i-1}$. Consider $u = w_1baw_2baw_3ba\ldots baw_{r-1}baw_r(aba)(bab)\phi(\rho(w_r))ba\phi(\rho(w_{r-1}))ba\ldots ba\phi(\rho(w_3))ba\phi(\rho(w_2))ba\phi(\rho(w_1)) \in L$. Remark that $|w_s|_a \leq i-2$ for any s, $s = 2, 3, \ldots, r$ where $|w_s|_a$ is the number of the occurrences of a in w_s. Suppose u contains a^i as a subword. Then a^i is a subword of $a\phi(\rho(w_k))$ for some k, $k = 2, 3, \ldots, r-1$. Since $|\phi(\rho(w_k))| = |w_k| \leq i - 1$, $\phi(\rho(w_k)) = a^{i-1}$. Thus $w_k = b^{i-1}$ and hence $k = r$, a contradiction. Therefore, $a^i = Height_{\leq}(u) \in Height_{\leq}(L)$.

Lemma 5 *Let $L \subseteq X^*$ be the above mentioned linear language. Then $a^+ \cap Height_{\leq}(L) = a^+ - \{a^{f(k)} \mid k = 1, 2, 3, \ldots\}$.*

Proof. Obvious from the previous lemmas.

Theorem 3 *Let $|X| \geq 2$. Then there exists a linear language $L \subseteq X^*$ such that $Height_{\leq}(L)$ is not even recursively enumerable under some total order $\leq$ on X^*.*

Proof. Let N be the set of all positive integers and let $1 = f(1) < f(2) < \ldots < f(k) < f(k+1) < \ldots$ be a sequence of positive integers such that $A = N - \{f(k) \mid k = 1, 2, \ldots\}$ is not recursively enumerable. Moreover, let $\leq$ be the total order on X^* satisfying the conditions (**) and let $L = \{w\ \phi(\rho(w)) \mid w \in X^+\}$. Consider $Height_{\leq}(L)$ under the order $\leq$. Suppose $Height_{\leq}(L)$ is recursively enumerable. Since the intersection of a regular set and a recursively enumerable set is recursively enumerable, $a^+ \cap Height_{\leq}(L)$ is recursively enu-

merable. By Lemma 5, $a^+ \cap Height_{\leq}(L) = \{a^i \mid i \in A\}$. However, this set is not recursively enumerable, a contradiction. Therefore, $Height_{\leq}(L)$ is not recursively enumerable.

Corollary *Let* $|X| \geq 2$. *Then there exists a context-free language* $L \subseteq X^*$ *such that* $Height_{\leq}(L)$ *is not even recursively enumerable under some total order* $\leq$ *on* X^*.

References

[1] J. Dassow, M. Ito and G. Paun, On the subword density of languages, The Southeast Asian Bulletin of Mathematics 18 (1994), 49 - 62.

[2] M. Ito, Dense and disjunctive properties of languages, Lecture Notes in Computer Science 710 (1993) (Springer-Verlag, New York), 31 - 49.

[3] M. Ito, Height functions and linear languages, Developments in Language Theory II (edit by J. Dassow) (World Scientific Publ. Co Pte Ltd, Singapore), to appear.

[4] M. Ito and C.M. Reis, Left dense covers of semigroups, Words, Languages and Combinatorics (1992) (World Scientific, Singapore), 202 - 218.

[5] M. Ito and G. Tanaka, Dense property of initial literal shuffles, International Journal of Computer Mathematics 34 (1990), 161 - 170.

[6] H.J. Shyr, Free Monoid and Languages, Lecture Notes, National Chung-Hsing University, Taichung, Taiwan, 1991.

Faculty of Science
Kyoto Sangyo University
Kyoto 603, Japan