数理解析研究所講究録
第 697 巻 1989 年 57-66

Some Algebraic Properties of Comma-Free Codes

C.Y. Hsieh, S.C. Hsu and H.J. Shyr

1. Introduction

Let X be a finite alphabet and let X^* be the free monoid generated by X. Any element of X^* is called a *word* and any subset of X^* is called a *language*. We let $X^+ = X^* - \{1\}$ where 1 is the empty word. A code is a language $L \subseteq X^+$ such that $x_1x_2 \cdots x_n = y_1y_2 \cdots y_m$, $x_i, y_j \in L$ implies $n = m$ and $x_i = y_i$ for $i = 1, 2, \cdots, n$. In recent years many different types of codes are studied, which include prefix codes, suffix codes, bifix codes, infix codes, outfix codes, uniform codes, etc. S.W. Colomb and others studied a particular kind of codes called comma-free codes. John A. Llewellyn quoted that a comma-free code is a directory of code words such that for any sequence of symbols, synchronization can be achieved within at most k symbols, where $k = 2 \times$ (the length of the longest word) - 1. Expressed alternatively : As a code in which a complete code word can be identified as soon as its last symbol is received. To achieve this, the set of code words must satisfy the condition that a set of symbols corresponding to a valid code word can occur neither in another code word nor within the catenation of two code words.

In this paper we show that the family of comma-free codes is a proper subfamily of infix codes. In fact a comma-free code can contain only primitive words. We obtained a characterization of this particular kind of codes.

2. Notations and Preliminaries

For a word $u \in X^*$, we let $|u|$ denote the length of the word u and for any two languages $A,B \subseteq X^*$, let AB be the set $AB = \{xy \mid x \in A, y \in B\}$. We call a word, $u \in X^+$, *primitive* if $u = f^n$, $n \geq 1, f \in X^+$ implies $n = 1$. The set of all primitive words over X will be denoted by Q. It is known that every word $u \in X^+$ is a power of a primitive word and the expression is unique. Thus if $u = f^n, f \in Q$, then we call f the *primitive root* of u. For a word $x = a_1 a_2 \cdots a_n$, $a_i \in X$, let the mirror image of x to be $\bar{x} = a_n a_{n-1} \cdots a_1$. (see [2], [3]).

Definition 2.1. Let X be an alphabet. A language $L \subseteq X^+$, $L \neq \varnothing$, is

(a) a *prefix code* if $L \cap LX^+ = \varnothing$;

(b) an *outfix code* if for all $x,y,u \in X^*$, $xy \in L$ and $xuy \in L$ together imply $u = 1$.

(c) an *infix code* if for $x,y,u \in X^*$, $u \in L$ and $xuy \in L$ together imply $xy = 1$.

For the properties of prefix codes, outfix codes and infix codes see [3].

We need the following lemmas in the sequel :

Lemma 2.1. (see [2]) *Let $u,v \in X^+$ with $u \neq 1$, $v \neq 1$. If $uv = vu$, then u and v are powers of a common word.*

Lemma 2.2. (see [5]) *Let $L \subseteq X^+$. Then L is a prefix code if and only if $L(A \cap B) = LA \cap LB$ for all $A, B \subseteq X^*$*

The term comma-free codes has been studied by several researchers. Especially the properties of maximal comma-free codes. Here we express the comma-free codes by a set relation.

Definition 2.2. Let X be an alphabet and let $L \subseteq X^+, L \neq \emptyset$. L is called a *comma-free code* if $L^2 \cap X^+LX^+ = \emptyset$.

Proposition 2.3. *A comma-free code is an infix code and hence a code.*

Proof. Suppose $L \subseteq X^+$ is a comma-free code, i.e., $L^2 \cap X^+LX^+ = \emptyset$. If L is not an infix code, then there exist $x,y \in X^*$, $u \in L$ such that $xy \neq 1$ and $xuy \in L$. Then $xuyxuy \in L^2 \cap X^+LX^+$, a contradiction. This shows that a comma-free code is an infix code. Q.E.D.

By definition every singleton set is an infix code. But this is not the case for comma-free codes. In fact we have the following.

Proposition 2.4. *Let $u \in X^+$. Then $\{u\}$ is a comma-free code if and only if u is a primitive word.*

Proof. ($\Rightarrow$) Suppose u is not a primitive word and let $u = f^n, f \in Q, n \geq 2$. Then $f^nf^n = ff^nf^{n-1} \in \{u^2\} \cap X^+uX^+$ and $\{u\}$ is not a comma-free code.

($\Leftarrow$) Suppose $\{u\}$ is not a comma-free code. Let $uu = xuy, x,y \in X^+$. Clearly, $|u| > |x|$ and $|u| > |y|$. Then $u = xx'$, $u = y'y$ for some $x', y' \in X^+$. It follows that $uu = xx'y'y = xuy$ and $u = x'y'$. Therefore, $u = xx' = x'y' = yy'$ and $|x| = |y'|$, $|x'| = |y|$. This then implies that $x = y'$ and $x' = y$. Thus $u = xx' = xy = yx$ holds. By Lemma 2.1, x and y are powers of a common word and u is not primitive, a contradiction. Q.E.D.

An infix code may not be a comma-free code. For $\{u^2\}$, $u \in X^+$ is an infix code but not a comma-free code.

It is immediate that a subset of a comma-free code is a comma-free code. The following is now clear :

Corollary. *Let $L \subseteq X^+$. If L is a comma-free code, then $L \subseteq Q$.*

Since every singleton set is an infix code, from Proposition 2.4 and the above corollary, we see that the family of comma-free codes is a proper subfamily of the family of infix codes.

Example : Let $X = \{a, b\}$. The language ba^+b is an infinite comma-free code. We can construct comma-free codes in the following ways.

(a) For any $L_1 \subseteq ab^+$ and $L_2 \subseteq b^+a$, the language L_1L_2 is a comma-free code. This is true. For L_1L_2 is a subset of ab^+a and ab^+a is a comma-free code.

(b) Let $L \subseteq X^+$ be a finite languages such that $m = max\{|u| \mid u \in L\}$. The language ba^mLba^m is always a comma-free code.

3. Characterizations of Comma-free Codes

In this section we characterize the comma-free codes. In doing so we need the following terms :

For any $L \subseteq X^+$, let

$L_p = \{x \in X^+ \mid xy \in L \text{ for some } y \in X^+\}$;

$L_s = \{y \in X^+ \mid xy \in L \text{ for some } x \in X^+\}$.

That is, L_p consists of all the proper prefixes of those words in L and L_s consists of all proper suffixes of those words in L.

Proposition 3.1. *Let X be an alphabet and let $L \subseteq X^+$. Then the following are equivalent :*

(1) *L is a comma-free code ;*

(2) *For any $u,v,w \in L$, $x,y \in X^*$, $uv = xwy$ imply $x = 1$ or $y = 1$;*

(3) *For any $u \in L$, $x,y \in X^*$, $xuy \in L^2$ imply $x = 1$ or $y = 1$;*

(4) *L is an infix code and $L \cap L_sL_p = \varnothing$;*

(5) *L is an infix code and $L^2 \cap L_pLL_s = \varnothing$;*

(6) *L is an infix code and $L^n \cap (X^+LX^+L^{n-1}) = \varnothing$, $n \geq 1$;*

(7) *L is an infix code and $L^n \cap (L^{n-1}X^+LX^+) = \varnothing$, $n \geq 1$;*

(8) *L is a comma-free code.*

Proof. The equivalences of (1), (2) and (3) are immediate.

(1) $\Rightarrow$ (4). Suppose L is a comma-free code. By Proposition 2.3, L is an infix code. For the second part, suppose on the contrary that $L \cap L_sL_p \neq \varnothing$ and let $w \in L \cap L_sL_p$. Then $w = xy$ for some $x \in L_p$ and $y \in L_p$. Since $x \in L_p$, $y \in L_p$, we have ux, $yv \in L$ for some $u,v \in X^+$. It follows that $uxyv = uwv \in L^2$ and $L^2 \cap X^+LX^+ \neq \varnothing$, a contradiction. Thus $L \cap L_pL_p = \varnothing$ holds.

(4) $\Rightarrow$ (1). Suppose the condition (4) holds and L is not a comma-free code. Let $u,v,w \in L$ be such that $uv = xwy$ for some $x,y \in X^+$. Since L is an infix code, $u \neq xws$ for all $s \in X^*$ and $v \neq wyr$ for all $r \in X^*$. The remaining case will be $w = w_1w_2$ with $w_1 \in L_s$, $w_2 \in L_p$ and which contradicts the fact that $L \cap L_sL_p = \varnothing$. This show that (4) $\Rightarrow$ (1).

(1) $\Rightarrow$ (5). Trivial.

(5) $\Rightarrow$ (1). Suppose (5) holds and L is not a comma-free code. Let $uv = xwy$ for some $u,v,w \in L$, $x,y \in X^+$. Since L is an infix code, we must have $u = xx'$, $v = y'y$ for some $x', y' \in X^+$. Clearly $x'y' = w$ and $x \in L_p$, $y \in L_s$. It follows that $uv \in L^2 \cap L_pLL_s = \emptyset$, a contradiction.

We now show the equivalences of (1), (6) and (7). If L is an infix code, then L is a bifix code. By Lemma 2.2,

$L^i(L \cap L^i X^+LX^+) = L^{i-1} \cap X^+LX^+$ and $(L \cap X^+LX^+)L^i = L^{i+1} \cap X^+LX^+$ for all $i \geq 1$.

It is clear that (1), (6) and (7) are equivalent.

(1) $\Leftrightarrow$ (8) Since for any $x,y,z,u,v \in X^+$ the condition $xy = uzv$ implies $\bar{y}\bar{x} = \overline{xy} = \overline{uzv} = \bar{v}\bar{z}\bar{u}$, it is clear that (1) is equivalent to (8). Q.E.D.

Proposition 3.2. *Let $L \subseteq X^+$ be an infix code. Then $L^3 \cap X^+L^2X^+ = \emptyset$ if and only if $L^2 \cap L_sLL_p = \emptyset$.*

Proof. ($\Rightarrow$) Immediate.

($\Leftarrow$) Suppose $L^3 \cap X^+L^2X^+ \neq \emptyset$. Then $u_1u_2u_3 = uxyv$ for some u_1, u_2, u_3, x, $y \in L$, $u,v \in X^+$. Since L is an infix code, we have $u_1 = uu'$, $u_3 = v'v$, $u' \in L_s$, $v' \in L_p$. $u_1u_2u_3 = uu'u_2v'v = uxyv$ implies $xy = u'u_2v'$. It then follows that $L^2 \cap L_sLL_p \neq \emptyset$, a contradiction. Q.E.D.

Corollary 3.3. *If $L \subseteq X^+$ is a comma-free code, then $L^2 \cap L_sLL_p \neq \emptyset$.*

4. Some Properties of Comma-free Codes and *n*-Comma-free Codes

Proposition 4.1. *If $L \subseteq X^+$ is a comma-free code, then for any positive integer $n \geq 3$, $L^n \cap X^+L^{n-1}X^+ = \emptyset$.*

Proof. We prove the proposition by induction on n. First we prove that the proposition holds for $n = 3$. Suppose $L^3 \cap X^+L^2X^+ \neq \emptyset$. Then $uvz = xwgy$ for some $u,v,z,g \in L$, $x,y \in X^+$. Clearly $u \neq x$ and $z \neq y$. If $x = uu'$ with $u' \in X^+$, then $uvz = xwgy = uu'wgy$ and $vz = u'wgy$ hold. It follows that $L^2 \cap X^+LX^+ \neq \emptyset$, a contradiction. Similarly $y \neq z'z$ for any $z' \in X^+$. The remaining case is that $u = xx'$ and $z = y'y$ for some $x', y' \in X^+$. We have $uvz = xx'vy'y = xwgy$ and $x'vy' = wg$, which again contradicts the fact that $L^2 \cap X^+LX^+ \neq \emptyset$. Thus $L^3 \cap X^+L^2X^+ = \emptyset$ holds.
Suppose the proposition holds for $n = k - 1$, i.e., $L^{k-1} \cap X^+L^{k-2}X^+ = \emptyset$. If $L^k \cap X^+L^{k-1}X^+ \neq \emptyset$, then there exist $w_1, w_2, \cdots, w_k, u_1, u_2, \cdots, u_{k-1} \in L$ such that $u_1u_2 \cdots u_k = xw_1w_2 \cdots w_{k-1}y$ for some $x,y \in X^+$. It is easy to see that we need to consider the following cases ; (1) $x = u_1u_1'$, (2) $u_1 = xx'$ and $y = u_k'u_k$ and, (3) $u_1 = xx'$ and $u_k = y'y$, where $x', y', u_1', u_k' \in X^+$. The above three conditions will all imply $L^{k-1} \cap X^+L^{k-2}X^+ \neq \emptyset$, a contradiction. Thus by induction we have that $L^n \cap X^+L^{n-1}X^+ = \emptyset$ for all $n \geq 3$. Q.E.D.

The converse of the above proposition is not true as we can see from the following example.

Example : Let $X = \{a, b\}$ and let $L \subseteq X^+$ be such that $L = \{ab^2, b^2ab\}$. Then $L^2 = \{ab^2ab^2, ab^4ab, b^2abab^2, b^2ab^3ab\}$ and $L^3 = \{ab^2ab^2ab^2, ab^2ab^4ab, ab^4abab^2, ab^4ab^3ab, b^2abab^2ab^2, b^2abab^4ab, b^2ab^3abab^2, b^2ab^3ab^3ab\}$. Here $L^3 \cap X^+L^2X^+ = \emptyset$ but $L^2 \cap X^+LX^+ \neq \emptyset$.

We note that every comma-free code is an anti-reflective language in the sense that for any $x,y \in X^+$, $xy \in L$ implies $yx \notin L$. Thus if $u \in Q$ and v is a cyclic permutation of u, then $\{u, v\}$ is not a comma-free code. However, the language $L = \{a^n b^n \mid n \geq 1\}$ is anti-reflective but not comma-free, where $a, b \in X, a \neq b$.

In general the catenation of two comma-free codes may not be a comma-free code. Nevertheless, for a given finite comma-free code L, we can always find a word u such that uL is a comma-free code.

In fact if $L = \{u_1, u_2, \cdots, u_n\}$ is a finite comma-free code and $m = max\{|u| \mid u \in L\}$, then for the word $u = a^{2m}b, a \neq b \in X, uL$ is clearly a comma-free code.

We could have more general setting. In fact we have the following :

Proposition 4.2. *For any finite comma-free code L, there exist an infinite language $A \subseteq X^+$ such that AL is a comma-free code.*

Proof. Let $L \subseteq X^+$ be a finite comma-free code such that $m = max\{|u| \mid u \in L\}$. Let $A = \{ab^{2m+n}a \mid n \geq 1\}$. Then clearly AL is a comma-free code. Q.E.D.

Like n-code considered by M. Ito and others, we now consider n-comma-free codes. An *n-comma-free code* is a language $L \subseteq X^+$ such that every n elements of L is a comma-free code.

Lemma 4.3. *A language $L \subseteq X^+$ is a 3-comma-free code if and only if L is a comma-free code.*

Proof. Immediate. Q.E.D.

Therefore, the only interesting n-comma-free code is a 2-comma-free code. By Proposition 2.4, we see that a language $L \subseteq X^+$ is a 1-comma-free code if and only if L consists of only primitive words.

Proposition 4.4. *Every 2-comma-free code is an infix code.*

Proof. Let L be a 2-comma-free code. Assume L is not an infix code. Then there exists $u \in L$ and $x,y \in X^*$, $xy \neq 1$ such that $xuy \in L$. This implies that $u, xuy \in L$ and $uxuy, xuyu \in L^2$, a contradiction. Therefore, L is an infix code. Q.E.D.

A word $u \in X^+$ is said to be *nonoverlapping* if $u = vx = yv$, $v,x,y \in X^*$ implies $v = 1$. A language $L \subseteq X^*$ is *nonoverlapping* if every word u contained in L is nonoverlapping.

We now have the following :

Proposition 4.5. *Let $L \subseteq Q$ be a nonoverlapping language. If L is an infix code, then L is a 2-comma-free code.*

Proof. Since $L \subseteq Q$, by Proposition 2.4 L is 1-comma-free code. Now suppose L is not a 2-comma-free code. Then there exist $u,v \in L(u \neq v)$ such that $\{u, v\}$ is not a comma-free code. By definition, $uv = xuy$ or $uv = x'vy'$ for some $x, x', y, y' \in X^*$.

Suppose $uv = xuy$. Then since $\{u, v\}$ is an infix code, we must have $u = xr$ for some $r \in X^+$. Thus $uv = xrv = xuy$ and u is not nonoverlapping, a contradiction.

Similarly, the case $uv = x'vy'$ also will lead to a contradiction. This shows that L is a 2-comma-free code. Q.E.D.

References

[1] Ito, M., Jurgensen, H., Shyr, H.J and Thierrin, G., n-Prefix-Suffix Languages, Report 167 Department of Computer Sciences, The University of Western Ontario, Ontario, 1987.

[2] Lyndon, R.C. and Schützenberger, M.P., The equation $a^M = b^N c^P$ in a free group, Michigan Math. J. 9(1962) 289 - 298

[3] Shyr, H.J., Free Monoids and Languages, Lecture Notes, Department of Mathematics, Soochow University, Taipei (1979)

[4] Colomb, S.W., Gordon, B. and Welch, L.R., Comma-free codes, Canada. J. Math. 10 (1958) 202 - 209

[5] Szilard, A.L., On Closure Properties of Language Operations, Thesis (M.A.) The University of Western Ontario (1968)

C.Y. Hsieh, S.C. Hsu and H.J. Shyr

Institute of Applied Mathematics
National Chung-Hsing University
Taichung, Taiwan 300