GENERALIZED PARIKH VECTORS AND PARTIAL LINE LANGUAGES

HULDAH SAMUEL, V. RAJKUMAR DARE

Abstract: The concept of Parikh vectors are defined to give an algebraic condition for Context free Languages. The Generalized Parikh Vectors are defined for words using the position of the letters in a word. Its combinatorial properties are studied in [2]. Partial words were introduced in the context of gene (or protein) comparison which are strings of symbols from a finite alphabet that may have a "do not know" symbol. In this paper we define the concept of Generalized Parikh Vector for Partial words and ω-Partial words.

Keywords: Generalized Parikh vector, line languages, partial words, ω-partial words.

Introduction: Partial words were introduced by Berstel and Boasson[5] in the context of gene (or protein) comparison which are strings of symbols from a finite alphabet that may have a "do not know" symbol. While a word can be described as a total function, a partial word can be described by a partial function. Some combinatorial properties of partial words have been investigated and many more are in progress. In this paper we define the Generalized Parikh vectors (GPV) for partial words and relate it to line languages.

Preliminaries: We recall some basic notions concerning words.

Definition 2.1: Let Σ be a finite non empty set of symbols called an alphabet. The symbols in Σ are called letters.

Any finite string over Σ is called a word overΣ.

For example, Σ = {a, b} then x = baaaba is a word overΣ.

An infinite word over Σ is an infinite string over Σ. For example, u = aabba ... is an infinite word. The set of all words over Σ is denoted by Σ^*. The collection of all infinite words is denoted by Σ^ω and $\Sigma^\infty = \Sigma^* \cup \Sigma^\omega$.

Notation.

The length of a word x over Σ is denoted as |x| and $|x|_a$ is the number of a's in the word x, where a∈Σ . In particular |λ| = 0.

Definition 2.2: Let u, v ∈Σ^+ be two words. Then u is a factor of v if v = x u y for some x, y ∈Σ^*.

Definition 2.3: A word u is a subword of a word x if there exists words $x_1, ...,x_n$ and $y_0, ..., y_n$, some of them possibly empty such that $u = x_1 ... x_n$, and $x = y_0 x_1 y_1 ... x_n y_n$.

Definition 2.4. [1]

If $\Sigma = \{a_1, a_2, ...,a_k\}$ then the Parikh mapping is a monoid morphism ψ : $\Sigma^* \to N^k$ where N denotes non-negative integers and $\psi(x) = (|x|_{a_1}, |x|_{a_2}, ..., |x|_{a_k})$.

Example.

Let Σ = {a, b} and x = aabbb.

Then ψ(x) = (2, 3).

Definition 2.5. [2]: For each u ∈Σ^∞, the generalized Parikh vector denoted by p(u) is given by

$p(u) = (p_1, p_2, ..., p_n)$ where $p_i = \sum_{j \in A_i} \frac{1}{2^j}$ where $A_i \subset N$ and A_i contains all the positions where a_i occurs in u.

Definition 2.6: Let |Σ| = 2. A language L ⊂Σ^* is called a line language if there exists a line ℓ in R^2 such that L = {x ∈Σ^∞ : p(x) lies on ℓ}. Then ℓ is called the language line of L.

Partial Words - its GPV and Line Languages: We now recall the definition of partial words and define GPV for partial words.

Definition 3.1. [5]: A partial word x is defined as a partial function from {1, ..., |x|} to Σ. The positions where x(n) [the nth letter of x] is not defined for n < |x|, are called holes of x. D(x) denotes the domain of x, and the set of all holes of x are denoted by H(x) where

H(x) = {{1, ..., |x|} \ D(x)}.

A word over Σ is a partial word over Σ with an empty set of holes.

The collection of all partial words is denoted by $\Sigma\Diamond^*$

Notation:

The symbol ◊ is used to represent a hole.

Definition 3.2: For any partial word u over Σ, |u| denotes its length including its holes.

For example, If x= ab◊ abb◊ ba then |x| = 9.

Definition 3.3: A strict partial word over Σ is a partial words over Σ with at least one hole.

Definition 3.4: An infinite partial word u over Σ is a partial map u : N →Σ. For 1 ≤i<∞, if u(i) is defined, then we say that i belongs to the domain of u, otherwise we say that u(i) is not defined and i belongs to the set of holes of u. We write an infinite partial word as ω-partial word. An ω-word over Σ is an ω-partial word over Σ with an empty set of holes.

The collection of all ω-partial words is denoted by $\Sigma_\Diamond^\omega$ and $\Sigma_\Diamond^\infty = \Sigma_\Diamond^* \cup \Sigma_\Diamond^\omega$.

Definition 3.5: A strict ω-partial word x over Σ is an ω-word with H(x) ≥ 1.

Example 3.1: u = aabab◊ab^ω is a strict ω-partial word with D(u) = {1, 2, 3, 4, 5, 7, 8, 9, ...} and H(u) = 6.

We now define the Generalized Parikh Vector for partial words and infinite partial words.

Definition 3.6: For a partial word x ∈Σ^∞, the Generalized Parikh vector denoted by $P_\Diamond(x)$, is defined by $P_\Diamond(x) = (p_1, p_2, ..., p_n)$ where $p_i = \sum_{j \in A_i} \frac{1}{2^j}$ where $A_i \subset N$ and A_i contains all the positions where a_i occurs in x.

Example 3.2.

Let ω = a◊bb◊a then

$$P_{\Diamond}(\omega) = \left(\frac{1}{2} + \frac{1}{2^6}, \frac{1}{2^3} + \frac{1}{2^4}\right).$$

Definition 3.7: A language $L \subset \Sigma^{\infty}$ is called a partial line language if there exists a line ℓ in R^2 such that, $L = \{x \in \Sigma_{\Diamond}^{\infty} : p_{\Diamond}(x)$ lies on $\ell\}$. Then ℓ is called the partial language line.

Definition 3.8: A partial line language is said to be an ω-partial line language if it contains only ω-partial words.

Example 3.3: $L = (ba)^* \Diamond (ba)^{\omega}$ is a ω-partial line languagesince itlies on the line y = 2x.

Theorem 3.1: The partial line language corresponding to the partial language line $x + y = \frac{1}{2^{n+1}}$, $n \geq 0$ is $L = \{(\Diamond^n a \Diamond^{\omega}) \cup (\Diamond^n b \Diamond^{\omega})\}$.

Proof.

Let $\Sigma = \{a, b\}$. The partial language line $x + y = \frac{1}{2}$ contains words $a \Diamond^{\omega}$ and $b \Diamond^{\omega}$ only. For n = 1, we find that the partial line language $x + y = \frac{1}{2^2}$ contains $L = \{(\Diamond a \Diamond^{\omega}) \cup (\Diamond b \Diamond^{\omega})\}$. It is easy to verify that $(\Diamond^n a \Diamond^{\omega}) \cup (\Diamond^n b \Diamond^{\omega})$ lies on the line $x + y = \frac{1}{2^{n+1}}$. Hence the theorem.

Theorem 3.2: If $\ell : x + y = \frac{2^n + 1}{2^{n+1}}$: n = 1, 2, ... then the corresponding partial line language is $L = \{\Sigma \Diamond^{n-1} \Sigma \Diamond^* : n\}$.

All words of the partial line language $L = \{\Diamond \Sigma \Diamond^{n-1} \Sigma \Diamond^{\omega}\}$: n = 1, 2, ... will lie on the partial language line $x + y = \frac{2^n + 1}{2^{n+1}}$.

Remarks:

1. All partial words have their GPV's in the region bounded by the lines x = 0, y = 0 and x + y = 1. This region is called PL-region (Figure 3.1) where no strictly partial word lies on the line x+y = 1.
2. Partial words are densely packed in the PL region.

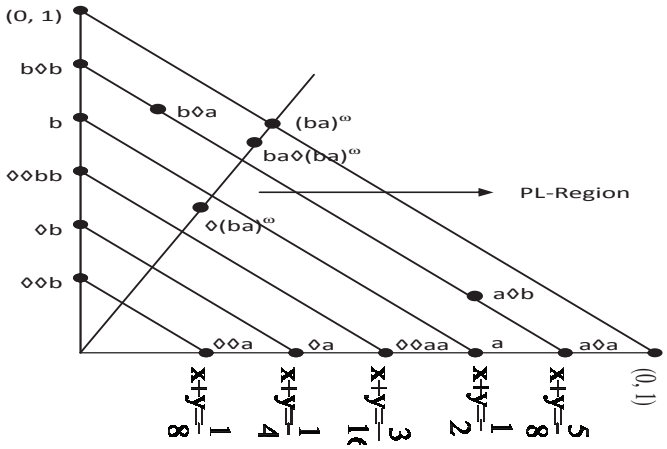

Fig. 3.1

Definition 3.9: All partial words lying in the region bounded by x = 0, y = 0 and any line ℓ form the triangle language corresponding to the line ℓ.

Theorem 3.4: The triangle language corresponding to the partial language line $x + y = \frac{1}{2^n}$ is $L = \{\Diamond^m \Sigma \Diamond^{\omega} : m > n\}$: m, n = 1, 2, ...

Proof.

No words lie on x = 0, y = 0. The partial words lying on the line $x + y = \frac{1}{2^n}$ would be of the form $\Sigma \Diamond^{\omega}$. Since the partial language lines corresponding to $x + y = \frac{1}{2^2}$, $x + y = \frac{1}{2^3}$, ... are $\Diamond \Sigma \Diamond^{\omega}$, $\Diamond^2 \Sigma \Diamond^{\omega}$, ... respectively, all these languages will lie inside the lines bounded by x = 0, y = 0 and $x + y = \frac{1}{2^n}$. Thus the triangle language corresponding to $x + y = \frac{1}{2^n}$ is $\{\Diamond^m \Sigma \Diamond^{\omega} : m > n\}$.

Conclusion: The Parikh vector indicates the number of occurrences of each letter in a word whereas the Generalized Parikh vector gives the exact positions of each letter in a word. In this paper we have defined the Generalized Parikh vector for partial words and have studied some properties of its line languages.

References:

1. R.J. Parikh, "On context-free languages," J. Assoe. Comp. Math., 13, 1966, pp. 570–581.
2. R. Siromoney, V.R. Dare, "A generalization of Parikh vectors for finite and infinite words," Lecture Notes in Computer Science, 206, Springer Verlag, 1985.
3. K. Sasikala, T. Kalyani, V.R. Dare and P.J. Abisha, "Line languages," Electronic Notes in Discrete Mathematics, 12, 2003.
4. Huldah Sathyaseelan, V. Rajkumar Dare, " On GeneralizedParikh Vector," Proceedings of National Conference on Discrete Mathematics and its Applications,NCDMA2007, pp.160-165.
5. Francine Blanchet – Sadri, "Algorithmic combinatorics on Partial words," Chapman & Hall / CRC.

* * *

Huldah Samuel, /Department of Mathematics/ Madras Christian College/ Tambaram/Chennai – 600 059/huldahs@yahoo.com

V. Rajkumar Dare/Department of Mathematics/ Madras Christian College/ Tambaram/ Chennai – 600 059/rajkumardare@yahoo.com