On Generalized Parikh Matrices for Finite and Infinite Words

Huldah Samuel
Department of Mathematics
Madras Christian College
Tambaram, Chennai – 600 059

ABSTRACT

The concept of the Generalized Parikh Matrices (GPM) of words is introduced in this paper. Some results and properties of the GPM for finite words over Σ are discussed. The definition is extended for infinite words too.

Keywords

Generalized Parikh vectors, line language, Generalized Parikh Matrices.

1. INTRODUCTION

The Parikh mapping, also called Parikh vector introduced in [1] is a significant contribution in the theory of formal languages and this has led to some important results. The Generalized Parikh vector introduced in [2] gives finer details about a word since it gives the exact positions of the letters in the word whereas the Parikh vector of a word over an alphabet counts only the number of occurrences of the letters of Σ in w. An extension of the Parikh mapping, called the Parikh matrix mapping was introduced in [3] based on a certain type of matrices. The Parikh matrix mapping is not injective as many words over an alphabet have the same Parikh matrix. In this paper we introduce Generalized Parikh matrix mapping which is injective and discuss a few properties of the same.

2. PRELIMINARIES

We recall certain notations of Formal Languages here.

Definition 2.1.
Let Σ be a finite alphabet and Σ^* be the set of all words over Σ. For u, w $\in \Sigma^*$, u is said to be a scattered subword of w, if w, which is a sequence of letters contains u as a subsequence.

Example 2.1.
Σ = {a, b}. Let w = aabaabab be a word. Then aa, ab, aaa, abba are all scattered subwords of w.

Definition 2.2.
The collection of all infinite words is denoted by Σ^ω and $\Sigma^\infty = \Sigma^* \cup \Sigma^\omega$.

Definition 2.3.
An alphabet Σ = {a_1, …, a_k} is said to be an ordered alphabet if there is a relation of order < on it. If $a_1 < \cdots < a_k$ then we use the notation Σ = {$a_1 < \cdots < a_k$}.

Definition 2.4.
The Parikh mapping is a monoid morphism $\psi : \Sigma^* \to N^k$ where N denotes non-negative integers and

$$\psi(w) = (|w|_{a_1}, |w|_{a_2}, ..., |w|_{a_k}),$$

where $|w|_{a_k}$ is the number of occurrences of a_k in w.

Parikh matrix for a binary word
Let Σ = {a, b} with a < b. The Parikh matrix mapping ψ is a mapping from Σ^* to M_3 given by $\psi(a) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$\psi(b) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$.

Definition 2.5.
For each u $\in \Sigma^\infty$, 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.

Definition 2.6.
A triangle matrix of order k, is a square matrix M = $(m_{i,j})$ $1 \le i$, $j \le k$ such that $m_{i,j}$ is a non-negative integer for all $1 \le i, j \le k$, $m_{i,j} = 0$ for all $1 \le j < i \le k$.

Notations.
1. The set of all triangle matrices of order k ≥ 1 is denoted by M_k.
2. The set of all triangle matrices is denoted by M.

Definition 2.7.
Let Σ = {$a_1 < \cdots < a_k$} be an ordered alphabet. The Parikh matrix mapping, denoted by ψ_Σ is the monoid morphism $\psi_\Sigma : (\Sigma^*, \cdot, \lambda) \to (M_{k+1}, \cdot, I_{k+1})$ defined by the condition, if $\psi_\Sigma(a_q) = (m_{i,j})$ $1 \le i, j \le k+1$, then for each $1 \le i \le k+1$,
- ❖ $m_{i,i} = 1$,
- ❖ $m_{q,q+1} = 1$
- ❖ and all other elements of the matrix $\psi_\Sigma(a_q)$ are 0.

Definition 2.8.
The Parikh matrix for a word w = $a_1 a_2 ... a_k$ is the product of Parikh matrices of the letters in the word.
i.e., $\psi(w) = \psi(a_1)\psi(a_2) ... \psi(a_k)$.

Definition 2.9.
A language L $\subset \Sigma^\infty$ is called a line language if there exists a line ℓ in R^2 such that L = {x $\in \Sigma^\infty$: p(x) lies on ℓ}. Then ℓ is called the language line of L.

3. GENERALIZED PARIKH MATRICES FOR FINITE WORDS

In this section we introduce the concept of Generalized Parikh Matrices (GPM) using the Generalized Parikh Vectors of words.

Definition 3.1.

The Generalized Parikh Matrix is a morphism $\psi : (\Sigma^*, \cdot, \lambda) \rightarrow (M_{k+1}, \cdot, I_{k+1})$ defined by the following conditions:

Let $u = a_{i_1} a_{i_2} ... a_{i_n} \in \Sigma^*$ where $a_{i_j} \in \Sigma = \{a_1, a_2, ..., a_k\}$ and $\psi(u) = \psi(a_{i_1}) \psi(a_{i_2}) ... \psi(a_{i_n})$ where each $\psi(a_{i_j}) = M_{ij}$ such that $M_{ij} = (m_{i,j})$ with $1 \le i, j \le (k+1)$,

❖ $m_{i,i} = 1$ for each $1 \le i \le (k+1)$
❖ $m_{j,j+1} = r$ where r is the i_jth coordinate of the GPV of u
❖ all other elements of the matrix are 0.

Example 3.1.

Let $\Sigma = \{a < b\}$ be an ordered alphabet. Then the Generalized Parikh matrix (GPM) of the word w = aba is

$$\psi(aba) = \psi(a)\psi(b)\psi(a)$$

$$= \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{2^2} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2^3} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{1}{2} + \frac{1}{2^3} & \frac{1}{2^3} \\ 0 & 1 & \frac{1}{2^2} \\ 0 & 0 & 1 \end{pmatrix}.$$

Example 3.2.

$$\psi(baab) = \psi(b)\psi(a)\psi(a)\psi(b)$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2^2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2^3} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{2^4} \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{1}{2^2} + \frac{1}{2^3} & \frac{1}{2^6} + \frac{1}{2^7} \\ 0 & 1 & \frac{1}{2} + \frac{1}{2^4} \\ 0 & 0 & 1 \end{pmatrix}.$$

Note.

The elements of the super diagonal of the GPM of a word w are the GPVs of the letters of the alphabet taken in the given order.

In the above Examples 2.1 and 2.2 the entries $m_{1,2}$, $m_{2,3}$ and $m_{1,3}$ are the GPVs of a's, b's and scattered subwords ab's in the words.

We now give a characterization theorem for Generalized Parikh Matrices.

Theorem 3.1.

Given an upper triangular square matrix $\begin{pmatrix} 1 & r_1 & r_2 \\ 0 & 1 & r_3 \\ 0 & 0 & 1 \end{pmatrix}$ where each r_1, r_2, r_3 are of the form $\frac{r}{2^n}$, (n is +ve integer), this matrix corresponds to a word $w \in \Sigma^*$, where $\Sigma = \{a < b\}$ if it satisfies the following conditions: (i) $r_1 + r_3$ is of the form $\frac{2^n - 1}{2^n}$ (ii) If $r_1 = \frac{x}{2^n}$, $r_2 = \frac{k}{2^t}$, $r_3 = \frac{y}{2^m}$ then t = n + m and k = 0 or $\le$ xy.

Proof.

(i) The GPM for a word is an upper triangular square matrix, whose super diagonal elements are the coordinates of the GPV of the word. Thus if r_1 and r_3 are the positions of a and b

in the word resp. then by a property of line languages, we have $r_1 + r_3 = \frac{2^n - 1}{2^n}$.

(ii) If the GPV of the word is $(r_1, r_3) = \left(\frac{x}{2^n}, \frac{y}{2^m} \right)$ then the GPV of the scattered subwords ab in the word is the product of $\left(\sum \frac{1}{2^i} \right) \left(\sum \frac{1}{2^j} \right)$ for $\forall$ i < j where $\sum \frac{1}{2^i}$ is the positions of a, and $\sum \frac{1}{2^j}$ is the positions of b. Hence it follows directly that t = n + m. The value of k is 0 or $\le$ xy. We consider the three following cases:

Case (1) If in the word w, all a's preceed b then all the scattered subwords ab of the word will be the product of the positions of every a and every b and hence k = xy.

Case (2) If in the word w, some b's are followed by a after which there are no other b's then in the scattered subword ab, the positions of a which follows b are not considered for the GPV of ab. Hence k < xy.

Case (3) If the word begins with b or a sequence of b's and is followed by a's, and there are no b's following a in the word then there are no scattered subwords ab hence k = 0.

Theorem 2.2.

If w_1 and w_2 are words over $\Sigma = \{a, b\}$ where the generalized Parikh matrices for w_1 and w_2 are $\psi(w_1)$ and $\psi(w_2)$ respectively denoted as follows:

$$\psi(w_1) = \begin{pmatrix} 1 & p_1 & r_1 \\ 0 & 1 & q_1 \\ 0 & 0 & 1 \end{pmatrix} \qquad \psi(w_2) = \begin{pmatrix} 1 & p_2 & r_2 \\ 0 & 1 & q_2 \\ 0 & 0 & 1 \end{pmatrix}$$

Then the Generalized Parikh Matrix for the catenation of the words w_1 and w_2 is

$$\psi(w_1 w_2) = \begin{pmatrix} 1 & p_1 + \frac{1}{2^{|w_1|}} p_2 & r_1 + \frac{1}{2^{2|w_1|}} r_2 + p_1 \frac{1}{2^{|w_1|}} q_1 \\ 0 & 1 & q_1 + \frac{1}{2^{|w_1|}} q_2 \\ 0 & 0 & 1 \end{pmatrix}$$

Proof.

Let w_1 and w_2 be words over Σ^+, $\Sigma = \{a, b\}$. Let the GPV of the words w_1 and w_2 be $p(w_1) = (p_1, q_1)$ and $p(w_2) = (p_2, q_2)$ where p_i are the positions of a's and q_i are the positions of b's in the words w_1 and w_2 for i = 1, 2 respectively. Then by the above note, we have the GPV of the word $w_1 w_2$ is

$$p(w_1 w_2) = \left(p_1 + \frac{1}{2^{|w_1|}} p_2, q_1 + \frac{1}{2^{|w_1|}} q_2 \right).$$

4. GENERALIZED PARIKH MATRICES FOR INFINITE WORDS

The concept of generalized parikh matrices introduced in the previous section was defined using generalized parikh vectors (GPV) of that finite word

We now define the generalized parikh matrices for infinite words (GPMI) using the following vectors.

Definition 3.1.

Let $\Sigma = \{a, b\}$. Let $w = a_1 a_2 \ldots a_n$ be a word, $w \in \Sigma^*$. Then we define the following vectors for w.

$p_1(w) = (p_a, p_b)$ where $p_a = \sum_{j \in A} \dfrac{1}{2^j}$ where $A \subset N$ and A contains all the positions where a occurs in the word w.

$p_2(w) = (p_{aa}, p_{ab}, p_{ba}, p_{bb})$ where $p_{a_i a_j} = \sum_{k \in A_i} \dfrac{1}{2^k}$, $A_i \subset N$ where

A_i contains all the positions of a_i where the factors $a_i a_j$ occurs.

Properties:

- For $\Sigma = \{a, b\}$ and $w \in \Sigma^*$, $p_1(w) = (p_a, p_b)$ and $p_2(w) = (p_{aa}, p_{ab}, p_{ba}, p_{bb})$ we observe that for $p_1(w)$,

  $p_a + p_b = \dfrac{2^{|w|} - 1}{2^{|w|}}$.

  In $p_2(w)$, $\sum_{1 \le i,j \le 2} p(a_i a_j) = \dfrac{2^{|w|-1} - 1}{2^{|w|-1}}$.

- For an infinite word w, the sum of the coordinates of $p_i(w)$ for all integers i, is equal to 1.
  For example, if $w \in \Sigma^\infty$ and $\Sigma = \{a, b\}$ then $p_1(w) = (p_a, p_b)$ where $p_a + p_b = 1$.

- For a word $w \in \Sigma^*$, $p_1(w) = (x, y)$ the word w lies on the line $x + y = \dfrac{2^{|w|} - 1}{2^{|w|}}$.

  For $p_2(w) = (x_1, x_2, x_3, x_4)$, the word $w \in \Sigma^*$ lies in the hyperplane $x_1 + x_2 + x_3 + x_4 = \dfrac{2^{|w|-1} - 1}{2^{|w|-1}}$.

- The sum of the coordinates of $p_1(w)$ is $\sum_{i=1}^{|w|} \dfrac{1}{2^i}$ and that of

  $p_2(w)$ is $\sum_{i=1}^{|w|-1} \dfrac{1}{2^i}$.

- For $\Sigma = \{a, b\}$, $p_i(w)$ is a 2^i-coordinate vector.
  i.e., $p_1(w)$ is a 2-coordinate vector.
  $p_2(w)$ is a $2^2 = 4$-coordinate vector.

Definition 3.2.

Let $\Sigma = \{a_1, a_2, \ldots, a_n\}$ and $a_1 < a_2 < \ldots < a_n$ be the ordering of the alphabet. Then the generalized parikh matrices for a word $w \in \Sigma^\infty$ denoted by GPMI, is an upper triangular square matrix of order (n+1) satisfying the following conditions;
entries $m_{i,i+1}$ are p_a and p_b of $p_1(w)$, for i = 1, 2 respectively.

entries $m_{i,i+2} = \begin{cases} p_{a_i a_j} \text{ of } p_2(w) & \text{with } i \neq j \text{ and } i < j, \\ 0 & \text{otherwise.} \end{cases}$

Example 3.1.
Let $w = (ab)^\omega$ over $\Sigma = \{a, b\}$. Then

$$GPMI((ab)^\omega) = \begin{pmatrix} 1 & p_a & p_{ab} \\ 0 & 1 & p_b \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^5} + \cdots \infty & \frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^5} + \cdots \infty \\ 0 & 1 & \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \cdots \infty \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{2}{3} & \frac{2}{3} \\ 0 & 1 & \frac{1}{3} \\ 0 & 0 & 1 \end{pmatrix}$$

Example 3.2.

$$GPMI(a^n b^\omega) = \begin{pmatrix} 1 & \frac{2^n - 1}{2^n} & \frac{1}{2^n} \\ 0 & 1 & \frac{1}{2^n} \\ 0 & 0 & 1 \end{pmatrix}$$

Example 3.3.

$$GPMI(ba^\omega) = \begin{pmatrix} 1 & \frac{1}{2^2} + \frac{1}{2^3} + \cdots \infty & 0 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}$$

5. CONCLUSION

The Generalized Parikh Matrix mapping introduced in this paper, is injective and found to give finer details of a word. A characterization theorem for GPM and the properties of the same have been dealt with in this paper. It will be of interest to examine the problem of finding properties of GPM for infinite words.

6. ACKNOWLEDGEMENT

7.REFERENCES

[1] R.J. Parikh, On context-free languages, J. Assoe. Comput. Mach., 13 (1966) 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] A. Mateescu, A. Salomaa, K. Salomaa, S. Yu, A sharpening of the Parikh mapping, Theoret. Informatics. Appl., 35 (2001), 551−564.

[4] K. Sasikala, T. Kalyani, V.R. Dare and P.J. Abisha, Line languages, Electronic Notes in Discrete Mathematics, 12, 2003.

[5] K.G. Subramanian, Ang Miin Huey and Atulya K. Nagar, On Parikh matrices, International Journal of Foundation of Computer Science, 20(2) (2009), 211−219