Parikh Matrices of Words

K.G. Subramanian* and Ang Miin Huey
School of Mathematical Sciences
Universiti Sains Malaysia, 11800, Penang, Malaysia

Abstract

The notion of Parikh matrix of a word over an ordered alphabet was introduced by Mateescu et al. in 2000, giving rise to a very interesting and effective tool in the study of certain numerical properties of a word, based on subwords (also called scattered subwords). Although the concept of a Parikh matrix has been recently introduced, a lot of research has taken place since then on investigating algebraic as well as language-theoretic properties of words based on associated Parikh matrices. The purpose of this article is to discuss the concept of a Parikh matrix and some of its interesting basic properties.

1. Introduction

The Norwegian mathematician Axel Thue (1863-1922) is considered to be the first to systematically make a study of combinatorial problems on words. The work of Thue can be considered to be the beginning of a new branch of Mathematics called "Combinatorics on words", which is a growing area of Discrete Mathematics with applications in many fields. The references [5-8, 10-12] are excellent monographs, surveys and tutorials on the topic of combinatorics on words and its applications.

A word is a finite or an infinite sequence of symbols taken from a finite set called an alphabet. We refer to a finite word simply as a word. For example the word $abba$ is over the alphabet $\Sigma=\{a,b\}$ and has an interesting feature of being a palindrome. The notion of a Parikh vector of a word introduced in [16] has been a significant contribution in the theory of formal languages as this notion has given rise to important results such as the semilinearity of the set of Parikh vectors of a context-free language.

* This article is based on an invited talk given at the International Conference on Research Trends in Applications of Mathematics at Dr MGR Janaki College of Arts and Science for Women, Chennai , Jan. 29-30, 2010.

The Parikh vector expresses a numerical property of a word by counting the number of occurrences of the symbols in the word. But many words over an alphabet can have the same Parikh vector and so information is lost while changing words into Parikh vectors. For example, if the alphabet is $\Sigma=\{a,b\}$ then the words $abaab, aaabb$ have the same Parikh vector (3, 2).

The notion of a Parikh matrix introduced in [15] based on a certain type of matrices is an extension of the Parikh vector. With every word w over an ordered alphabet, a Parikh matrix can be associated and it is a triangular matrix, with 1's on the main diagonal and 0's below it but the entries above the main diagonal provide information on the number of certain subwords (also called scattered subwords) in w. An interesting aspect of the Parikh matrix is that it has the classical Parikh vector as the second diagonal above the main diagonal. Although the Parikh matrix is still not injective, two words with the same Parikh vector have in many cases different Parikh matrices and thus the Parikh matrix gives more information about a word than a Parikh vector does. Since the introduction of this interesting notion of a Parikh matrix, many works investigating properties of words based on these matrices have appeared. See for example [1-4, 9, 13-15, 17-27].

This article, aimed at an interested and uninitiated reader in the topic of Parikh matrices, starts with an introduction to the concept of a Parikh matrix of a word and discusses some of the interesting basic properties of the Parikh matrix, besides providing a detailed list of references on this topic, although not exhaustive. The authors acknowledge having used the references [1,3,9,13-15,17,27] for the exposition of this article.

2. Parikh Matrix

Let Σ be an alphabet. The set of all words over Σ is denoted by Σ^* and the empty word by λ. For a word $w \in \Sigma^*$, $|w|$ denotes the length of w. A word u is a subword of a word w, if there exist words $x_1 \cdots x_n$ and $y_0 \cdots y_n$, (some of them possibly empty), such that $u = x_1 \cdots x_n$ and $w = y_0 x_1 y_1 \cdots x_n y_n$. For example if w = ***abbaabab*** is a word over the alphabet $\{a,b\}$, then ***babb*** is a subword of w. (In the literature subwords are also called "scattered subwords"). The number of occurrences of the word u as a subword of the word w is denoted by $|w|_u$. Two occurrences of a sub-word are considered different if they differ by at least one position of some letter. In the word w = *abbaabab*, the number of occurrences of the word *babb* as a subword of w is 4 i.e. $|w|_{babb} = 4$.

We recall (mostly informally) the definition of a Parikk matrix [11], which is a generalization of the Parikh vector [12]. We mostly restrict our attention to a binary alphabet $\Sigma=\{a,b\}$ and binary words over Σ. We also assume the alphabet $\Sigma=\{a,b\}$ to be ordered in the sense that $a < b$ and if the alphabet is $\Sigma=\{a,b,c\}$, then we assume $a < b < c$.

The **Parikh vector** is a mapping Ψ from Σ^* to $\mathrm{N}\times\mathrm{N}$ where $\Sigma=\{a,b\}$ and N is the set of natural numbers including zero, such that for a word w in Σ^*, $\Psi(w)=(|w|_a, |w|_b)$ with $|w|_a$ denoting the number of occurrences of the letter a in w. For example for the word w = *abbaabab* the Parikh vector is (4,4).

The notion of a Parikh matrix is an extension of the Parikh vector.

Let M_3 denote 3×3 upper triangular matrices with non-negative integer entries and the main diagonal entries being 1s and all entries below the main diagonal being zeros. For example the 3×3 matrix $M=\begin{pmatrix}1 & 4 & 17\\ 0 & 1 & 4\\ 0 & 0 & 1\end{pmatrix}$ is an element of M_3.

Parikh matrix of a binary word

Let $\Sigma=\{a,b\}$ with $a < b$. The Parikh matrix mapping Ψ_2 is a mapping from Σ^* to M_3 given by

$$\Psi_2(a)=\begin{pmatrix}1 & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}, \Psi_2(b)=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 1\\ 0 & 0 & 1\end{pmatrix}$$

and $\Psi_2(uv)=\Psi_2(u)\Psi_2(v)$, $u,v \in \Sigma^*$ where multiplication of matrices is the operation on the right side of this equation. For a word *w*, the matrix $\Psi_2(w)$ is called the Parikh matrix of *w*.

As an illustration of the computation of the Parikh matrix of a word, consider $w = abbaabab$. Then

$$\Psi_2(w)=\Psi_2(abbaabab)= \Psi_2(a)\Psi_2(b)\Psi_2(b)\Psi_2(a)\Psi_2(a)\Psi_2(b)\Psi_2(a)\Psi_2(b)$$

$$=\begin{pmatrix}1&1&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&1\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&1\\0&0&1\end{pmatrix}\begin{pmatrix}1&1&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&1&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&1\\0&0&1\end{pmatrix}\begin{pmatrix}1&1&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&1\\0&0&1\end{pmatrix}$$

$$=\begin{pmatrix}1&4&9\\0&1&4\\0&0&1\end{pmatrix}$$

is the Parikh matrix of $w = abbaabab$, which is an upper triangular matrix with 1s in the main diagonal. Note that the word $w = abbaabab$ has four *a*, four *b* and nine subwords *ab*. Also the Parikh vector (4, 4) occurs in the second diagonal above the main diagonal in the Parikh matrix of *w*. In fact we notice that for a binary word *w*

$$\Psi_2(w)=\begin{pmatrix}1 & |w|_a & |w|_{ab}\\0&1&|w|_b\\0&0&1\end{pmatrix}.$$

Parikh Matrix of a word over an arbitrary ordered alphabet:

In fact, the notion of a Parikh matrix is defined for words over an arbitrary ordered alphabet although we have considered here only binary words over two letters. In fact if the alphabet is $\Sigma=\{a,b,c\}$ with $a < b < c$, then the Parikh matrix of a word over $\Sigma=\{a,b,c\}$ is a 4×4 upper triangular matrix such that

$$\Psi_3(w)=\begin{pmatrix}1 & |w|_a & |w|_{ab} & |w|_{abc}\\0&1&|w|_b&|w|_{bc}\\0&0&1&|w|_c\\0&0&0&1\end{pmatrix}.$$

3. Properties of Parikh matrices

We now begin to discuss some of the basic properties of Parikh matrices of binary words.

A Characterization:

In the following Theorem, the entries in the second and third diagonals of a Parikh matrix are characterized. This yields a complete characterization of Parikh matrices of binary words as there are no more diagonals in the Parikh matrix of a binary word.

Theorem 1:

Let $M = \begin{pmatrix} 1 & m_{12} & m_{13} \\ 0 & 1 & m_{23} \\ 0 & 0 & 1 \end{pmatrix}$ be the Parikh matrix of a word over $\Sigma = \{a, b\}$ with $a < b$.

Then

i) m_{12}, m_{23} on the second diagonal of M can be arbitrary non-negative integers

ii) m_{13} can be arbitrary integers but satisfying the condition $0 \le m_{13} \le m_{12}m_{23}$.

Remark:

Not every upper triangular matrix in M_3 can be Parikh matrix of a binary word. For example the matrix $M = \begin{pmatrix} 1 & 4 & 17 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix}$ which is in M_3 is not a Parikh matrix of any word as it can be seen from Theorem 1, that the entry in the first row, third column can at the most be 16.

M-equivalent or amiable words:

The Parikh matrix mapping is not injective. In other words, two or more words can have the same Parikh matrix. For example the following five words over the alphabet $\Sigma = \{a, b\}$ with $a < b$

baabaab, baaabba, abbaaab, abababa, aabbbaa have the same Parikh matrix

$$\begin{pmatrix} 1 & 4 & 6 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$$

Two words having the same Parikh matrix are called M-equivalent or amiable. The five words mentioned above are M-equivalent.

A word $w \in \{a,b\}^*$ with $a < b$ is called M-unambiguous, if there is no word $w' \in \{a,b\}^*$ so that $\Psi_2(w) = \Psi_2(w')$. Otherwise w is called M-ambiguous. For example the word *baabaab* is M-ambiguous.

A characterization of M-ambiguous binary words:

For a binary words, there is a simple characterization of M-ambiguous words. This is stated in the following Theorem.

Theorem 2 A word $w \in \{a,b\}^*$ with $a < b$ is M-ambiguous if and only if w has the factors *ab* and *ba* in non-overlapping positions.

For example in the word *a**ab**b**ba**a*, the factors ab, ba occur in non-overlapping positions (indicated in bold letters). This word is M-ambiguous and it is known that there are four other words *baabaab, baaabba, abbaaab, abababa* having the same Parikh matrix as *aabbbaa.*

Inverse of a Parikh matrix:

Let $\Sigma = \{a,b\}$ with $a < b$. The mirror image *mi*(w) of a word $w = c_1 c_2 \cdots c_{k-1} c_k$, $c_i \in \{a,b\}, 1 \le i \le k$, is defined by $w = c_k c_{k-1} \cdots c_2 c_1$. For example, if $w = abbaaab$, then $mi(w) = baaabba.$

If $\Psi_2(w) = (m_{ij})_{1 \le i,j \le 3}$ is the Parik matrix of a word w, then the alternate Parikh matrix of w is the matrix $\overline{\Psi}_2(w) = (m^*_{ij})_{1 \le i,j \le 3}$, $m^*_{ij} = (-1)^{i+j} m_{ij}, 1 \le i, j \le 3.$

The next theorem brings out the connection between the inverse of the Parikh matrix of a word w and the alternate Parikh matrix of the mirror image of w.

Theorem 3 Let $\Sigma=\{a,b\}$ with $a<b$ and $w\in\{a,b\}^*$ be a word. Then the inverse of the Parikh matrix of w is given by

$$[\Psi_2(w)]^{-1} = \overline{\Psi_2}(mi(w)).$$

As an illustration the inverse of the Parikh matrix $\begin{pmatrix}1 & 4 & 6\\ 0 & 1 & 3\\ 0 & 0 & 1\end{pmatrix}$ of the word *abbaaab* is the alternate Parikh matrix of *mi(abbaaab)* = *baaabba*, which is $\begin{pmatrix}1 & -4 & 6\\ 0 & 1 & -3\\ 0 & 0 & 1\end{pmatrix}$ as *baaabba* has also the same Parikh matrix $\begin{pmatrix}1 & 4 & 6\\ 0 & 1 & 3\\ 0 & 0 & 1\end{pmatrix}$.

Ratio-property:

For words over $\Sigma=\{a,b\}$ with $a<b$, define $u\equiv v$, if there exist words x, y, z such that $u = xabybaz$; $v = xbayabz$. The relation $\equiv$ is an equivalence relation.

Theorem 2 can now be reworded as follows:

Theorem 4

Let $\Sigma=\{a,b\}$ with $a<b$. For words u,v over Σ the Parikh matrices $\Psi_2(u)$ and $\Psi_2(v)$ are equal if and only if $u \equiv v$.

Two words w_1, w_2 over $\Sigma=\{a,b\}$ with $a<b$ are said to satisfy the ratio property if $p_i = sq_i, i=1,2$ and s is a constant, where the Parikh matrices of w_1, w_2 are respectively

$$\begin{pmatrix}1 & p_1 & p_{12}\\ 0 & 1 & p_2\\ 0 & 0 & 1\end{pmatrix}, \begin{pmatrix}1 & q_1 & q_{12}\\ 0 & 1 & q_2\\ 0 & 0 & 1\end{pmatrix}$$

We then write $w_1 \sim_r w_2$.

We can now relate the equivalence relations $\equiv$ and $\sim_r$.

Theorem 5 Let $\Sigma=\{a,b\}$ with $a<b$. Let w_1, w_2 be any two words over Σ such that $w_1 \sim_r w_2$. If x, y are any two words over Σ such that $x \sim_r w_1$, $y \sim_r w_2$, then

$$\Psi_2(xy)=\Psi_2(yx).$$

4. Conclusion

The topic of Parikh matrices is a promising area of research related to combinatorics on words. In this article some of the properties of Parikh matrices of binary words were discussed. When the alphabet has more than two symbols there are many challenging problems (similar to or different from the binary case) that have been explored and that await investigation. The interested reader can consult the references listed and also other references in them for such interesting problems. It should be mentioned here that another generalization of Parikh vector that reflects the positions of letters in a word is considered in [28].

Acknowledgement The authors gratefully acknowledge support from an FRGS grant and a short term grant of Universiti Sains Malaysia.

References

1. A. Atanasiu, Binary amiable words, *Intern. J. Found. Comput. Sci.,* 18 (2007) 387-400.
2. A. Atanasiu, R. Atanasiu and I. Petre, Parikh matrices and amiable words, *Theo. Comp. Sci.*, 390 (2008) 102-109.
3. A. Atanasiu, C. Martin-Vide, A. Mateescu, On the injectivity of the Parikh matrix mapping, *Fundam. Inform.* 46 (2001) 1-11.
4. A. Atanasiu, C. Martin-Vide, A. Mateescu, Codifiable languages and the Parikh matrix mapping, *J. Univ. Comp. Sci.*, 7 (2001) 783-793.
5. J. Berstel, Axel Thue's works on repetitions in words, *Fourth Conference on Formal power series and algebraic combinatorics*, 1992.
6. J. Berstel and J. Karhumaki, Combinatorics of words – A Tutorial, *Bulletin of EATCS,* 79 (2003) 178-228.
7. J. Berstel and D. Perrin, The origins of combinatorics on words, *European J. Combinatorics,* 28 (2007) 996-1022.
8. C. Choffrut and J. Karhumaki, Combinatorics of Words, In *"Handbook of Formal Languages"* (Eds. G. Rozenberg and A. Salomaa), Springer-Verlag, 1997.
9. S. Fosse, G. Richomme, Some characterizations of Parikh matrix equivalent binary words, *Inf. Process. Letters*, 92 (2004) 77-82.

10. J. Karhumaki, Combinatorics on words: A new challenging topic, In: Abel, M. (Ed.), *Proceedings of FinEst, Estonian Mathematical Society,* Tartu, (2004) 64-79.
11. M. Lothaire, Combinatorics on Words, Addison-Wesley, 1983.
12. M. Lothaire, *Algebraic Combinatorics on words*, Cambridge University Press, 2002.
13. A. Mateescu, Algebraic aspects of Parikh matrices, *Lecture Notes in Comp. Sci.* 3113 (2004) 170-180.
14. A. Mateescu, A. Salomaa, Matrix indicators for subword occurrences and ambiguity, *Int. J. Found. Comput. Sci.* 15 (2004)277-292.
15. A. Mateescu, A. Salomaa, K. Salomaa, S. Yu, A Sharpening of the Parikh Mapping, *Theoret. Informatics Appl.*, 35 (2001) 551-564.
16. R.J. Parikh, On context-free languages. *J. Assoc. Comput. Mach.* **13** (1966) 570-581.
17. A. Salomaa, Connections between subwords and certain matrix mappings, *Theor. Comput. Sci.* 340 (2005) 188-203.
18. A. Salomaa, On the injectivity of Parikh matrix mappings, *Fundam. Inform.* 64 (2005) 391-404.
19. A. Salomaa, Independence of certain quantities indicating subword occurrences, *Theo. Comp. Sci.* 362 (2006) 222-231.
20. A. Salomaa, Subword balance in binary words, languages and sequences, *Fundam. Inform.*, 75 (2007) 469-482.
21. A. Salomaa, Upper triangular matrices and subword occurrences, *Scientiae Mathematicae Japonicae Online,* e-2009-32.
22. A. Salomaa, Criteria for matrix equivalence of words, To appear in *Theo. Comp. Sci.*
23. A. Salomaa and S. Yu, Subword conditions and subword histories, *Inform. Comput.,* 204 (2006) 1741-1755.
24. A. Salomaa, and S. Yu, Subword occurrences, Parikh matrices and Lyndon Images, *TUCS Tech. Report* 929, 2009.
25. T.-F. Serbanuta, Extending Parikh matrices, *Theo. Comp. Sci.* 310 (2004) 233-246.
26. V.N. Serbanuta and T.-F. Serbanuta, Injectivity of the Parikh matrix mappings revisited, *Fundam. Inform.* 73 (2006) 265-283.
27. G. Siromoney, R. Siromoney, K.G. Subramanian and V.R. Dare, Generalized Parikh vector and public key cryptosystems, In "A Perspective in Theoretical Computer Science, Commemorative Volume for Gift Siromoney," Ed: R. Narasimhan, World Scientific (1989) 301-323.
28. K.G. Subramanian, A.M. Huey and A.K. Nagar, On Parikh matrices, *Int. J. Found. Comp. Sci.,* 20 (2009) 211-219.