Malaysian Journal of Mathematical Sciences 13(S) April: 147–156 (2019)
Special Issue: 3rd International Conference on Mathematical Sciences and Statistics (ICMSS2018)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Extension of Parikh Matrices to Terms and its Injectivity Problem

Chern, Z. J. and Teh, W. C. *

School of Mathematical Sciences, Universiti Sains Malaysia, Malaysia

E-mail: dasmenteh@usm.my
* *Corresponding author*

ABSTRACT

Parikh matrices introduced by Mateescu et al. are very useful in understanding structural properties of words by analyzing their numerical properties. This is due to the information of a word provided by its Parikh matrix is more than its Parikh vector. The study of Parikh matrices is extended in this paper to terms formed over a signature with a binary underlying alphabet. We obtain some numerical properties that characterize when a word is a term. Finally, new M-equivalence preserving rewriting rules are introduced and shown to characterize M-equivalence for our terms, thus contributing towards the injectivity problem.

Keywords: Injectivity problem, M-equivalence, Parikh matrices, subword, terms.

1. Introduction

Parikh's theorem introduced in Parikh (1966) states that the set of Parikh vectors of words in a context-free language is a semilinear set. Parikh matrices introduced in Mateescu et al. (2001) is an extension of the Parikh vectors. Parikh matrices have been widely used in studying (scattered) subword occurrences in words (for example, see Mateescu et al. (2004), Salomaa (2005, 2006)). Two words formed over an ordered alphabet are *M*-equivalent if and only if they share the same Parikh matrix. Although the characterization of *M*-equivalence, also known as the injectivity problem, has been actively investigated (for example, see Atanasiu (2007), Atanasiu et al. (2008, 2002), Fossé and Richomme (2004), Mahalingam and Subramanian (2012), Poovanandran and Teh (2018), Salomaa (2010), Şerbănuţă (2009), Şerbănuţă and Şerbănuţă (2006), Teh (2016a,b), Teh and Atanasiu (2016), Teh et al. (2016)), it remains open even for the ternary alphabet. Meanwhile, for the binary case, the *M*-equivalence preserving rewriting rules defined in Atanasiu et al. (2008) completely characterize its *M*-equivalence.

In this work, a signature to us consists of a set of function symbols and a set of constant symbols such that every function symbol has its own arity. A term over a signature is a word recursively constructed from constant symbols and function symbols. In fact, such a term can be treated as a word formed over an underlying alphabet containing symbols of the signature.

Our work focuses on terms formed over a fixed signature containing a constant symbol and a binary function symbol. We obtain combinatorial properties that characterize when a word is a term over the signature. Analogously, we also introduce rewriting rules, called Rules $E2T$, to determine whether two terms are *M*-equivalent. Our main result shows that Rules $E2T$ is sufficient to characterize *M*-equivalence for our terms.

2. Parikh Matrices

The cardinality of a set X is denoted by $|X|$.

Suppose Σ is a finite alphabet. The set of words over Σ is denoted by Σ^*. The empty word is denoted by λ. Let Σ^+ denote the set $\Sigma^* \backslash \{\lambda\}$. If $v, w \in \Sigma^*$, the concatenation of v and w is denoted by vw. An *ordered alphabet* is an alphabet $\Sigma = \{a_1, a_2, \dots, a_s\}$ with a total ordering on it. For example, if $a_1 < a_2 < \cdots < a_s$, then we may write $\Sigma = \{a_1 < a_2 < \cdots < a_s\}$. On the other hand, if $\Sigma = \{a_1 < a_2 < \cdots < a_s\}$ is an ordered alphabet, then the *underlying* *alphabet* is $\{a_1, a_2, \ldots, a_s\}$. For $1 \le i \le j \le s$, let $a_{i,j}$ denote the word $a_i a_{i+1} \cdots a_j$. Frequently, we will abuse notation and use Σ to stand for both the ordered alphabet and its underlying alphabet, for example, as in "$w \in \Sigma^*$" when Σ is an ordered alphabet. If w is a word, then $|w|$ is the length of w.

Definition 2.1. *A word w' is a* subword *of $w \in \Sigma^*$ iff there exist $x_1, x_2, \ldots, x_n, y_0, y_1, \ldots, y_n \in \Sigma^*$, possibly empty, such that*

$$w' = x_1 x_2 \cdots x_n \text{ and } w = y_0 x_1 y_1 \cdots y_{n-1} x_n y_n.$$

A *factor* is a contiguous subword. The number of occurrences of a word u as a subword of w is denoted by $|w|_u$. Two occurrences of u are considered different iff they differ by at least one position of some letter. For example, $|aabab|_{ab} = 5$ and $|baacbc|_{abc} = 2$. By convention, $|w|_\lambda = 1$ for all $w \in \Sigma^*$. The reader is referred to Rozenberg and Salomaa (1997) for language theoretic notions not detailed here.

For any integer $k \ge 2$, let $\mathcal{M}_k$ denote the multiplicative monoid of $k \times k$ upper triangular matrices with nonnegative integral entries and unit diagonal.

Definition 2.2. *Suppose $\Sigma = \{a_1 < a_2 < \cdots < a_s\}$ is an ordered alphabet. The* Parikh matrix mapping*, denoted Ψ_Σ, is the monoid morphism*

$$\Psi_\Sigma : \Sigma^* \to \mathcal{M}_{s+1}$$

defined as follows:
$\Psi_\Sigma(\lambda) = I_{s+1}$; if $\Psi_\Sigma(a_q) = (m_{i,j})_{1 \le i,j \le s+1}$, then $m_{i,i} = 1$ for each $1 \le i \le s+1$, $m_{q,q+1} = 1$ and all other entries of the matrix $\Psi_\Sigma(a_q)$ are zero. Matrices of the form $\Psi_\Sigma(w)$ for $w \in \Sigma^*$ are called Parikh matrices.

Theorem 2.1 (Mateescu et al. (2001))**.** *Suppose $\Sigma = \{a_1 < a_2 < \cdots < a_s\}$ is an ordered alphabet and $w \in \Sigma^*$. The matrix $\Psi_\Sigma(w) = (m_{i,j})_{1 \le i,j \le s+1}$ has the following properties:*

- *$m_{i,i} = 1$ for each $1 \le i \le s+1$;*
- *$m_{i,j} = 0$ for each $1 \le j < i \le s+1$;*
- *$m_{i,j+1} = |w|_{a_{i,j}}$ for each $1 \le i \le j \le s$.*

The *Parikh vector* $\Psi(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_s})$ of a word $w \in \Sigma^*$ is contained in the second diagonal of the Parikh matrix $\Psi_\Sigma(w)$.

Example 2.1. *Suppose* $\Sigma = \{a < b < c\}$ *and* $w = abacc$. *Then*

$$\begin{aligned}\Psi_\Sigma(w) &= \Psi_\Sigma(a)\Psi_\Sigma(b)\Psi_\Sigma(a)\Psi_\Sigma(c)\Psi_\Sigma(c)\\ &= \begin{pmatrix}1&1&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}1&0&0&0\\0&1&1&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\cdots\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&1\\0&0&0&1\end{pmatrix}\\ &= \begin{pmatrix}1&2&1&2\\0&1&1&2\\0&0&1&2\\0&0&0&1\end{pmatrix} = \begin{pmatrix}1&|w|_a&|w|_{ab}&|w|_{abc}\\0&1&|w|_b&|w|_{bc}\\0&0&1&|w|_c\\0&0&0&1\end{pmatrix}.\end{aligned}$$

Definition 2.3. *Suppose* $\Sigma = \{a_1 < a_2 < \cdots < a_s\}$ *is an ordered alphabet.*

(1) Two words $w, w' \in \Sigma^*$ *are M-equivalent, denoted* $w \equiv_M w'$, *iff* $\Psi_\Sigma(w) = \Psi_\Sigma(w')$.

(2) A word $w \in \Sigma^*$ *is M-unambiguous iff no distinct word is M-equivalent to* w. *Otherwise,* w *is said to be M-ambiguous.*

There are two elementary rules called $E1$ and $E2$ first formally defined in Atanasiu et al. (2008) for deciding whether two words are M-equivalent. The following is Rule $E2$ stated for the binary alphabet, which is the only rewriting rule applicable in this case.

$E2$. Suppose $\Sigma = \{a < b\}$ and $w, w' \in \Sigma^*$. If $w = xabybaz$ and $w' = xbayabz$ for some $x, y, z \in \Sigma^*$, then $w \equiv_M w'$.

Rule $E2$ is sufficient to characterize M-equivalence for the binary alphabet.

Example 2.2. *Let* $w = abbbaab$, $w' = bababab$ *and* $w'' = baabbba$. *By Rule* $E2$, w *(respectively* w'*) is M-equivalent to* w' *(respectively* w''*) with respect to* $\{a < b\}$. *Then,* w *is M-equivalent to* w'' *with respect to* $\{a < b\}$ *due to transitivity of M-equivalence.*

Theorem 2.2 (Atanasiu (2007), Fossé and Richomme (2004))**.** *Suppose* $\Sigma = \{a < b\}$ *and* $w, w' \in \Sigma^*$. *Then* w *and* w' *are M-equivalent if and only if* w' *can be obtained from* w *by finitely many applications of Rule* $E2$. *Hence, any word* w *is M-ambiguous if and only if there are nonoverlapping factors ab and ba in* w.

3. Parikh Matrices on Terms

To us a signature Σ consists of a set of function symbols F and a set of constant symbols C such that each function symbol is assigned an arity. The set of terms over Σ is the set of words over $F \cup C$ that can be recursively constructed by the following rules:

- Every constant symbol is a term.
- If $t_1, t_2, \ldots, t_n$ are terms and f is an n-ary function symbol, then $ft_1t_2 \cdots t_n$ is a term.

In the study of Parikh matrices, we view terms as words formed over the underlying alphabet $F \cup C$. Hence, we study the M-equivalence of our terms with respect to some ordered alphabet with underlying alphabet $F \cup C$.

From now on, fix a signature Σ containing a binary function symbol f and a constant symbol a. Our study focuses only on terms over Σ which are words over $\{f, a\}$. Also, our work studies the M-equivalence of our terms with respect to the ordered alphabet $\{f < a\}$. Since two terms are M-equivalent with respect to $\{f < a\}$ if and only if they are M-equivalent with respect to $\{a < f\}$ due to $|w|_{af} + |w|_{fa} = |w|_a|w|_f$ for any $w \in \Sigma^*$, our results also hold for the ordered alphabet $\{a < f\}$. Henceforth, we will abuse notation and let Σ represent as our fixed signature as well as the ordered alphabet $\{f < a\}$ and its underlying alphabet.

Theorem 3.1. *A word $w \in \Sigma^+$ is a term over Σ if and only if the following properties hold:*

(1) $|w|_a = |w|_f + 1$

(2) $|\alpha|_a \geq |\alpha|_f + 1$ *for any proper suffix* $\alpha \in \Sigma^+$ *of* w.

Proof. We argue by induction on the complexity of terms to show that any term $w \in \Sigma^+$ satisfies properties (1) and (2). Clearly the constant symbol a satisfies properties (1) and (2) and thus the base step holds. For the induction step, suppose $t_1, t_2 \in \Sigma^+$ are terms satisfying properties (1) and (2). We need to prove that ft_1t_2 satisfies the two properties. Since $|t_1|_a = |t_1|_f + 1$ and $|t_2|_a = |t_2|_f + 1$, it follows that $|ft_1t_2|_a = |t_1|_a + |t_2|_a = |t_1|_f + 1 + |t_2|_f + 1 = |t_1|_f + |t_2|_f + 2 = |ft_1t_2|_f + 1$. Hence, property (1) holds for the term ft_1t_2. To prove property (2), suppose α is an arbitrary proper suffix of ft_1t_2. Consider the following cases.

Case 1. *α is a proper suffix of t_2.*

We are done as t_2 satisfies property (2).

Case 2. *$\alpha = t_2$.*

In this case, $|t_2|_a = |t_2|_f + 1$.

Case 3. *α is a proper suffix of t_1t_2.*

Since t_2 satisfies property (1) and t_1 satisfies property (2), it follows that $|\alpha|_a \geq |\alpha|_f + 1$.

Case 4. *$\alpha = t_1t_2$.*

In this case, $|t_1t_2|_a = |t_1|_a + |t_2|_a = |t_1|_f + 1 + |t_2|_f + 1 = |t_1t_2|_f + 2$.

Conversely, suppose w is a nonempty word satisfying properties (1) and (2). This time we argue by induction on the length of w. The base step holds since the only word of length 1 that satisfies property (1) is a and that is clearly a term. Consider the induction step. Let t_2 be the unique proper suffix of w such that $|t_2|_a = |t_2|_f + 1$ and t_2 has the maximal length among such proper suffixes. By property (2), the last letter of w must be a, thus a is one such proper suffix. Let t_1 be the unique word such that $w = ft_1t_2$. Note that t_2 satisfies properties (1) and (2). By the induction hypothesis, it follows that t_2 is a term. Since w satisfies property (1), it follows that t_1 also satisfies property (1). Assume t_1 does not satisfy property (2) and thus there exists a proper suffix $x \in \Sigma^+$ of t_1 such that $|x|_a < |x|_f + 1$. Consider the proper suffix xt_2 of w. Since t_2 is the longest proper suffix of w such that $|t_2|_a = |t_2|_f + 1$, it follows that xt_2 must satisfy $|xt_2|_a > |xt_2|_f + 1$, a contradiction as this is not possible with $|x|_a < |x|_f + 1$ and $|t_2|_a = |t_2|_f + 1$. Hence, t_1 also satisfies property (2) and thus t_1 is a term by the induction hypothesis. Since t_1 and t_2 are terms, it follows that $w = ft_1t_2$ is a term. □

The following Rules *E2T* is used to determine whether two terms are *M*-equivalent. Suppose $w, w' \in \Sigma^+$ such that w is a term over Σ.

> *E2T*1. If $w = xafyfaz$ and $w' = xfayafz$ for some $x, z \in \Sigma^+$, $y \in \Sigma^*$ and $|z|_a \geq |z|_f + 2$, then w' is a term and $w \equiv_M w'$.
>
> *E2T*2. If $w = xfayafz$ and $w' = xafyfaz$ for some $x, z \in \Sigma^+$, $y \in \Sigma^*$ and $|x|_f \geq |x|_a + 1$, then w' is a term and $w \equiv_M w'$.

The Rules $E2T$ are sound as this follows from the soundness of Rule $E2$ and the characterization for terms as in Theorem 3.1.

Example 3.1. *Consider the words* $w = faffaffaaaa$, $w' = ffafafafaaa$ *and* $w'' = ffaafffaaaa$. *By Rule* $E2T1$ *(respectively, Rule* $E2T2$*), term* w *(respectively* w'*) is M-equivalent to term* w' *(respectively* w''*) with respect to* Σ. *Then,* w *is M-equivalent to* w'' *due to transitivity of M-equivalence.*

Remark 3.1. *Suppose* $w, w' \in \Sigma^+$ *are terms over* Σ. *If* w' *can be obtained from* w *by an application of Rule* $E2T1$, *then* w *can be obtained from* w' *by an application of Rule* $E2T2$ *and vice versa.*

Lemma 3.1. *Suppose* $w, w' \in \Sigma^+$ *are terms over* Σ *such that* $w \equiv_M w'$. *For every* $1 \leq k \leq |w|$, *finitely many applications of Rule* $E2T1$ *can be applied to* w *and* w' *to obtain* w'' *and* w''' *respectively such that* w'' *and* w''' *agree up to suffix of length* k.

Proof. We argue by induction. Since w and w' are terms, the last letter of each must be a and thus their suffixes of length 1 agree. Hence, the base step holds. For the induction step, by the induction hypothesis, we can obtain w'' from w and w''' from w' by finitely many applications of Rule $E2T1$ such that w'' and w''' agree up to suffix of length k.

Case 1. $w'' = ux\alpha$ *and* $w''' = vx\alpha$ *for some* $x \in \Sigma$ *and* $u, v, \alpha \in \Sigma^+$ *such that* $|\alpha| = k$.

Clearly, $x\alpha$ *is the common suffix of length* $k+1$ *of* w'' *and* w'''.

Case 2. $w'' = ufa^j\alpha$ *and* $w''' = vf\alpha$ *for some* $u, v, \alpha \in \Sigma^+$ *and positive integer* j *such that* $|\alpha| = k$.

Since $w'' \equiv_M w'''$, *by the right invariance of M-equivalence, it follows that* $ufa^j \equiv_M vf$ *and thus* ufa^j *is M-ambiguous. By Theorem 2.2,* ufa^j *contains nonoverlapping factors* af *and* fa. *Hence,* $u = xafy$ *for some* $x, y \in \Sigma^*$ *and thus* $w'' = xafyfa^j\alpha$. *Since* w''' *is a term, it follows that* $|f\alpha|_a \geq |f\alpha|_f + 1$ *and thus* $|\alpha|_a \geq |\alpha|_f + 2$. *Hence, we can apply Rule* $E2T1$ *to* w'' *and obtain the term* $w_1 = xfayafa^{j-1}\alpha$.

Now we repeat the process. Let $u_1 = xfaya$ *and thus* $w_1 = u_1fa^{j-1}\alpha$. *Since* $w'' \equiv_M w_1$, *by transitivity,* $w_1 \equiv_M w'''$. *By the right invariance of M-equivalence,* $u_1fa^{j-1} \equiv_M vf$. *Hence,* u_1fa^{j-1} *is M-ambiguous and thus by Theorem 2.2, there are nonoverlapping factors* af *and* fa *in* u_1fa^{j-1}. *Therefore,* $u_1 = x_1afy_1$ *for some* $x_1, y_1 \in \Sigma^*$ *and thus* $w_1 = x_1afy_1fa^{j-1}\alpha$. *Similarly, we can then apply Rule* $E2T1$ *to* w_1 *and obtain the term* $w_2 = x_1fay_1afa^{j-2}\alpha$.

Hence, after a total of j many applications of Rule $E2T1$ to w'', we will obtain a term $w_j = \beta f \alpha$ for some $\beta \in \Sigma^+$. Clearly, $f\alpha$ is the common suffix of length $k+1$ of w_j and w'''.

Case 3. *$w'' = uf\alpha$ and $w''' = vfa^j\alpha$ for some $u, v, \alpha \in \Sigma^+$ and positive integer j such that $|\alpha| = k$.*

This is similar to Case 2. □

Theorem 3.2. *Suppose $w, w' \in \Sigma^+$ are terms over Σ. Then w and w' are M-equivalent if and only if w' can be obtained from w by finitely many applications of Rules $E2T$.*

Proof. The backward direction is straightforward as Rules $E2T$ are sound.

Conversely, suppose $w, w' \in \Sigma^+$ are M-equivalent terms over Σ. By Lemma 3.1, there exists two terms $w'', w''' \in \Sigma^+$ that can be obtained from w and w' respectively by finitely many applications of Rule $E2T1$ such that they agree up to suffix of length $|w|$. This implies that $w'' = w'''$. By Remark 3.1, finitely many applications of Rule $E2T2$ can be applied to w''' to obtain w'. Hence, w' can be obtained from w by finitely many applications of Rules $E2T$. □

Therefore, Theorem 3.2 shows that Rules $E2T$ is sufficient to characterize M-equivalence for our terms.

Remark 3.2. *In fact, Theorem 3.2 tells us that for every M-equivalent terms w, w' over Σ, we can always obtain w' from w by applying finitely many applications of Rule $E2T1$ followed by Rule $E2T2$.*

4. Conclusion

This paper presents a new direction in the study of Parikh matrices where we focus on special words that are terms. Rules $E2T$ introduced prove to completely characterize M-equivalence for our terms. Hence, our work contributes to the injectivity problem for terms over our fixed signature. As a continuation, we are working on the characterization of M-unambiguous terms and the preservation of M-equivalence for our terms under the shuffle operation analogous to what is done in Atanasiu and Teh (2002). Our study of Parikh matrices for terms can also be extended to terms formed over any signature with a ternary underlying alphabet.

Acknowledgements

Both authors gratefully acknowledge the financial support for this research by Research University Grant No. 1001/PMATHS/8011019 of Universiti Sains Malaysia.

References

Atanasiu, A. (2007). Binary amiable words. *Internat. J. Found. Comput. Sci.*, 18(2):387–400.

Atanasiu, A., Atanasiu, R., and Petre, I. (2008). Parikh matrices and amiable words. *Theoret. Comput. Sci.*, 390(1):102–109.

Atanasiu, A., Martín-Vide, C., and Mateescu, A. (2002). On the injectivity of the Parikh matrix mapping. *Fund. Inform.*, 49(4):289–299.

Atanasiu, A. and Teh, W. C. (2016). A new operator over Parikh languages. *Internat. J. Found. Comput. Sci.*, 27(6):757–769.

Fossé, S. and Richomme, G. (2004). Some characterizations of Parikh matrix equivalent binary words. *Inform. Process. Lett.*, 92(2):77–82.

Mahalingam, K. and Subramanian, K. G. (2012). Product of Parikh matrices and commutativity. *Internat. J. Found. Comput. Sci.*, 23(1):207–223.

Mateescu, A., Salomaa, A., Salomaa, K., and Yu, S. (2001). A sharpening of the Parikh mapping. *Theor. Inform. Appl.*, 35(6):551–564 (2002).

Mateescu, A., Salomaa, A., and Yu, S. (2004). Subword histories and Parikh matrices. *J. Comput. System Sci.*, 68(1):1–21.

Parikh, R. J. (1966). On context-free languages. *J. Assoc. Comput. Mach.*, 13:570–581.

Poovanandran, G. and Teh, W. C. (2018). On M-equivalence and strong M-equivalence for Parikh matrices, *Internat. J. Found. Comput. Sci.*, 29(1):123–137.

Rozenberg, G. and Salomaa, A. (1997). *Handbook of formal languages. Vol. 1.* Springer-Verlag, Berlin.

Salomaa, A. (2005). Connections between subwords and certain matrix mappings. *Theoret. Comput. Sci.*, 340(2):188–203.

Salomaa, A. (2006). Independence of certain quantities indicating subword occurrences. *Theoret. Comput. Sci.*, 362:222–231.

Salomaa, A. (2010). Criteria for the matrix equivalence of words. *Theoret. Comput. Sci.*, 411(16):1818–1827.

Şerbănuţă, V. N. (2009). On Parikh matrices, ambiguity, and prints. *Internat. J. Found. Comput. Sci.*, 20(1):151–165.

Şerbănuţă, V. N. and Şerbănuţă, T. F. (2006). Injectivity of the Parikh matrix mappings revisited. *Fund. Inform.*, 73(1):265–283.

Teh, W. C. (2016a). Parikh matrices and Parikh rewriting systems. *Fund. Inform.*, 146:305–320.

Teh, W. C. (2016b). Parikh matrices and strong M-equivalence. *Internat. J. Found. Comput. Sci.*, 27(5):545–556.

Teh, W. C. and Atanasiu, A. (2016). On a conjecture about Parikh matrices. *Theoret. Comput. Sci.*, 628:30–39.

Teh, W. C., Atanasiu, A., and Poovanandran, G. (2018). On strongly M-unambiguous prints and Şerbănuţă's conjecture for Parikh matrices, *Theoret. Comput. Sci.*, 719:86–93.