数理解析研究所講究録
1093巻 1999年 68-73

A Characterization of Min-Wise Independent Permutations Families

Yoshinori Takei[*] Toshiya Itoh[†]

Abstract. A Min-Wise Independent Permutation Family is an efficient tool to estimate similarity of documents. We present a characterization of Exact MWIPFs by *size uniformity*, which represents certain symmetry of the string representation of a family. Also, we present a general construction strategy which produce *any* Exact MWIPF using this characterization.

1 Introduction

1.1 Background

The notion of Min-Wise Independency is recently defined, motivated by the need for efficient calculation of "Resemblance"[1], which is an effective criterion of similarlity of two documents. Min-Wise Independency is a property defined for a family of permutations on $[n] = \{1, 2, \ldots, n\}$. In the paper [2], they presented definitions of various level of Min-Wise Independency with some construction of permutation families. The most fundamental and tight class of Min-Wise Independency is:

Definition 1.1 (Exact MWIPF [2]). We say that $F \subset S_n$ =(the set of all permutations on $[n]$) is an Exact Min-Wise Independent Permutation Family if the following holds:

$$\forall X \subset [n]\,(X \neq \phi),\, \forall x \in X \quad [\ \Pr_{\pi \in_{\mathcal{U}} F}[\min \pi(X) = \pi(x)] = 1/\|X\|\ ], \tag{1.1}$$

where $\pi \in_{\mathcal{U}} F$ means that π is chosen uniformly at random from F.

The rest of the paper focus on Exact Min-Wise Permutation Families and we shall omit the word "Exact". In the proof of [2] Theorem 6, they obtained more explicit condition:

Theorem 1.2 (equivalent condition [2]). *A subset F of S_n is min-wise independent if and only if the following holds:*

$$\forall 0 \leq k < n \forall X \subset [n]\,(\|X\| = n - k)\,\forall x \in X \tag{1.2}$$
$$[\ \Pr_{\pi \in_{\mathcal{U}} F}[\pi([n] \setminus X) = \{1, 2, \ldots, k\} \textit{ and } \pi(x) = k + 1] = 1/((n-k)\tbinom{n}{k})\ ].$$

Lower and upper bounds of their size are:

Theorem 1.3 (lower bound of size [2], **Theorem 1**). *For any integer $n > 0$, let $\mathcal{C} \subseteq S_n$ be a family of min-wise independent permutations. Then $\|\mathcal{C}\|$ is a multiple of* $\mathrm{lcm}(n, n-1, \ldots, 2, 1)$ *and hence* $\|\mathcal{C}\| \geq e^{n - o(n)}$.

Theorem 1.4 (an optimal construction in the sense of size [6], **Theorem 3.3**). *For any integer $n > 0$, there exists a family of min-wise independent permutations $\mathcal{F} \subseteq S_n$ such that* $\|\mathcal{F}\| = \mathrm{lcm}(n, n-1, \ldots, 2, 1)$.

1.2 Main Results

In Section 2, we reformulate the equivalent condition of Min-Wise Independency (Theorem 1.2) as *size uniformity*, using a certain string representation of permutations. In Section 3, we use the characterization to generalize the construction of Theorem 1.4 to a strategy which produce various Min-Wise Independent Permutation Families (Theorem 3.5), and show that the strategy is enough general to produce *any* Min-Wise Independent Permutation Family (Theorem 3.6).

[*] Department of Information Processing, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan. ytakei@kis.ip.titech.ac.jp.

[†] Department of Information Processing, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan. titoh@ip.titech.ac.jp.

1.3 Notations and Definitions

In this paper, MWIPF means Exact Min-Wise Permutation Family.

We interpret $\mathcal{S}_n$ as the set of strings:

$$\mathcal{S}_n = \{\pi = \langle x_1, \dots, x_i, \dots, x_n\rangle \mid x_i \in [n],\ x_i \neq x_j\ (i \neq j)\} \tag{1.3}$$

where, $\pi = \langle x_1, \dots, x_i, \dots, x_n\rangle$ represents the permutation $\pi : [n] \ni x_i \mapsto i \in [n]$. We put $\mathcal{H}_{n,k} := \{H \subset [n] \mid \|H\| = k\}$, the collection of all size-k subsets of $[n]$. For each $H \in \mathcal{H}_{n,k}$ and $x \in [n]$, define subsets $L_k(H)$ and $M_k(x)$ of $\mathcal{S}_n$ as:

$$L_k(H) := \{\langle x_1, \dots, x_k, \dots, x_n\rangle \in \mathcal{S}_n \mid \{x_1, \dots, x_k\} = H\} \tag{1.4}$$

$$M_k(x) := \{\langle x_1, \dots, x_k, \dots, x_n\rangle \in \mathcal{S}_n \mid x_k = x\}. \tag{1.5}$$

Then a subset F of $\mathcal{S}_n$ decomposes into sum of disjoint subsets:

$$F = \coprod_{H \in \mathcal{H}_{n,k}} (F \cap L_k(H)) = \coprod_{H \in \mathcal{H}_{n,k}} \coprod_{\xi \in [n] \setminus H} (F \cap L_k(H) \cap M_{k+1}(\xi)). \tag{1.6}$$

We define the set of first-k substrings of $\mathcal{S}_n$

$$\mathcal{S}_{n,k} := \{\pi = \langle x_1, \dots, x_i, \dots, x_k\rangle \ \mid x_i \in [n],\ x_i \neq x_j\ (i \neq j)\} \quad (0 \leq k \leq n) \tag{1.7}$$

and define the first-k substring oprators for $0 \leq k \leq \ell \leq n$:

$$\varphi^\ell_k : \mathcal{S}_{n,\ell} \ni \langle x_1, \dots, x_k, \dots, x_\ell\rangle \mapsto \langle x_1, \dots, x_k\rangle \in \mathcal{S}_{n,k}. \tag{1.8}$$

2 Characterization of MWIPF by Size Uniformity

Using the above definitions, we obtain a reformulation of Theorem 1.2, which characterizes MWIPF by certain symmetry of size of subsets in the string representation:

Theorem 2.1. *A subset F of $\mathcal{S}_n$ is a MWIPF if and only if*

$$\|F \cap L_k(H) \cap M_{k+1}(\xi)\| = \|F\|/((n-k)\tbinom{n}{k}) \quad \text{for all } 0 \leq k < n, H \in \mathcal{H}_{n,k}, \xi \in [n] \setminus H \tag{2.9}$$

Though it is merely a direct translation of Theorem 1.2, it gives an explicit goal when we try to construct a MWIPF. So we shall emphasize this characteristic of MWIPFs, calling it *size uniformity.*

3 A General Construction of MWIPF

3.1 Informal Description of the Strategy

We sketch basic strategy of the general construction(see Figure 1). First, we start with the replicated null strings $C \cdot \langle\rangle$, where C is the cardinality of the set to produce. Then we iterate stages consist of classifying and appending, as in the former construction [6], but allowing more general appending rather than the cyclic appending. Here $\mathrm{AP}^H_{n,k}[C]$ denotes a map which appends each element of $[n] \setminus H$ to each string of $\mathcal{F}^H_{n,k}$. In Subsection 3.4 we will discuss what is admissible as $\mathrm{AP}^H_{n,k}[C]$ to output a MWIPF. Then we regard that for given constant n and C (under the condition C is possible value as cardinality of a MWIPF), when we fix a sequence of admissible subroutines $(\mathrm{AP}^H_{n,k}[C])^{0 \leq k < n}_{H \in \mathcal{H}_{n,k}}$, we get a construction of MWIPF . In Theorem 3.5, we shall show the strategy produces a MWIPF for all combination of admissible appending maps.

Set theoretical notations $\mathcal{F}, \mathcal{G}$ are too informal since they contain multiple elements. In what follows, we reformulate each procedure to avoid confusion.

Figure 1: General strategy

```
Initial:  Let $\mathcal{F}_{n,0} := \mathcal{F}^{\phi}_{n,0} := \{C \cdot \langle\rangle\}$    (Replicating Procedure)
for       $k = 0$ to $n-1$ do
Stage $k$:
          Classifying Procedure: Classify strings $\langle x_1, \ldots, x_k\rangle$ among $\mathcal{F}_{n,k}$
          by their contents as sets $\{x_1, \ldots, x_k\}$, i.e.
            for each $H \in \mathcal{H}_{n,k}$ do
                 Let $\mathcal{F}^{H}_{n,k} := \mathcal{F}_{n,k} \cap \varphi^{n}_{k}(L_k(H))$.
            end
            for each $H \in \mathcal{H}_{n,k}$ do
                 Appending Procedure:
                 Let $\mathcal{G}^{H}_{n,k+1} := \mathrm{AP}^{H}_{n,k}[C](\mathcal{F}^{H}_{n,k})$
            end
          Let $\mathcal{F}_{n,k+1} := \bigcup_{H \in \mathcal{H}_{n,k}} \mathcal{G}^{H}_{n,k+1}$.
end
output $\mathcal{F}_n = \mathcal{F}_{n,n}$.
```

3.2 Formulation of Replicating

For an arbitrary subset S of $\mathcal{S}_{n,k}$, let $\mathfrak{M}(S)$ denote the free $\mathbb{Z}$-module generated by S:

$$\mathfrak{M}(S) := \{\mathbf{x} = \sum\nolimits_{\sigma \in S} a(\sigma)\sigma \mid a(\sigma) \in \mathbb{Z}\} \tag{3.10}$$

When all of coefficents of an element $\mathbf{x} \in \mathfrak{M}(S)$ are nonnegative, it is identified with a so-called "multiset" and each coefficent $a(\sigma)$ represent multiplicity of σ in $\mathbf{x}$. If all of its coefficients are either 0 or 1, they represent the characteristic function of a "genuine" subset of S. To represent these situations, we define subsets of $\mathfrak{M}(\mathcal{S}_{n,k})$ for $p, q \in \mathbb{Z}$,

$$\mathsf{MULT}[p,q]_k := \{\mathbf{x} = \sum\nolimits_{\sigma \in \mathcal{S}_{n,k}} a(\sigma)\sigma \mid a(\sigma) \in \mathbb{Z}, p \le a(\sigma) \le q\}, \tag{3.11}$$

then $\mathsf{MULT}[0,\infty]_k$[1] ($\mathsf{MULT}[0,1]_k$) represents the collection of all multisets over $\mathcal{S}_{n,k}$ (resp. genuine subsets of $\mathcal{S}_{n,k}$).

In these cases the sum of all coefficients coincides with the cardiality of its (multi)set interpretation. So we define the weight of an element of $\mathfrak{M}(\mathcal{S}_{n,k})$ as:

$$w_k(\sum_{\sigma \in \mathcal{S}_{n,k}} a(\sigma)\sigma) := \sum_{\sigma \in \mathcal{S}_{n,k}} a(\sigma) \tag{3.12}$$

The substring operator φ^{ℓ}_{k} induces the linear map:

$$\begin{array}{cccccc} \Phi^{\ell}_{k}: & \mathfrak{M}(\mathcal{S}_{n,\ell}) & \twoheadrightarrow & & & \mathfrak{M}(\mathcal{S}_{n,k}) \\ & \Cup & & & & \Cup \\ & \sum\limits_{\tau \in \mathcal{S}_{n,\ell}} b(\tau)\tau & \mapsto & \sum\limits_{\tau \in \mathcal{S}_{n,\ell}} b(\tau)\varphi^{\ell}_{k}(\tau) & = & \sum\limits_{\sigma \in \mathcal{S}_{n,k}} \left(\sum\limits_{\tau : \varphi^{\ell}_{k}(\tau) = \sigma} b(\tau) \right) \sigma \end{array} \tag{3.13}$$

for $0 \le k \le \ell \le n$. Obviously Φ^{ℓ}_{k} preserves weight:

$$w_{\ell}(\mathbf{x}) = w_k(\Phi^{\ell}_{k}(\mathbf{x})) \quad (\mathbf{x} \in \mathfrak{M}(\mathcal{S}_{n,\ell})) \tag{3.14}$$

[1] In this case, the inequality on $a(\sigma)$ is should be read as $0 \le a(\sigma) < \infty$.

3.3 Formulation of Classifying

The direct decompsition formula (1.6) is extended to $\mathfrak{M}(\mathcal{S}_{n,k})$:

$$\mathfrak{M}(\mathcal{S}_{n,k}) = \bigoplus_{H\in\mathcal{H}_{n,k}} \mathfrak{M}(\varphi_k^n(L_k(H))) = \bigoplus_{H\in\mathcal{H}_{n,k}} \bigoplus_{\xi\in[n]\setminus H} \mathfrak{M}(\varphi_{k+1}^n(L_k(H)\cap M_{k+1}(\xi))). \tag{3.15}$$

We define the projections:

$$\begin{array}{rccc} \Psi_k^H: & \mathfrak{M}(\mathcal{S}_{n,k}) & \twoheadrightarrow & \mathfrak{M}(\varphi_k^n(L_k(H))) \\ & \rotatebox{90}{$\in$} & & \rotatebox{90}{$\in$} \\ & \sum\limits_{\sigma\in\mathcal{S}_{n,k}} a(\sigma)\sigma & \mapsto & \sum\limits_{\sigma\in\varphi_k^n(L_k(H))} a(\sigma)\sigma\ . \end{array} \tag{3.16}$$

and

$$\begin{array}{rccc} \psi_{k+1}^{H,\xi}: & \mathfrak{M}(\mathcal{S}_{n,k+1}) & \twoheadrightarrow & \mathfrak{M}(\varphi_{k+1}^n(L_k(H)\cap M_{k+1}(\xi))) \\ & \rotatebox{90}{$\in$} & & \rotatebox{90}{$\in$} \\ & \sum\limits_{\sigma\in\mathcal{S}_{n,k+1}} a(\sigma)\sigma & \mapsto & \sum\limits_{\sigma\in\varphi_{k+1}^n(L_k(H)\cap M_{k+1}(\xi))} a(\sigma)\sigma, \end{array} \tag{3.17}$$

respectively. Taking the projection Ψ_k^H is exactly "classifying" in this setting. To represent the notion of size uniformity, we add some more definitions. For $C\in\mathbb{Z}>0$ and each $0\le k<n$, $H\in\mathcal{H}_{n,k}$, $\xi\in[n]\setminus H$, define subsets of $\mathfrak{M}(\mathcal{S}_{n,k+1})$ as:

$$\mathsf{UNIF}[C]_{k+1}^{H,\xi} := \{\mathbf{x}\in\mathfrak{M}(\mathcal{S}_{n,k+1}) \mid w_{k+1}(\psi_{k+1}^{H,\xi}(\mathbf{x})) = C/((n-k)\tbinom{n}{k})\}. \tag{3.18}$$

Then put

$$\begin{aligned} \mathsf{UNIF}[C]_{k+1} &:= \bigcap\nolimits_{H\in\mathcal{H}_{n,k}} \bigcap\nolimits_{\xi\in[n]\setminus H} \mathsf{UNIF}[C]_{k+1}^{H,\xi}, && (3.19) \\ \mathsf{UNIF}[C]_{\le k} &:= \bigcap\nolimits_{0\le i\le k} (\Phi_i^k)^{-1}(\mathsf{UNIF}[C]_i), && (3.20) \end{aligned}$$

respectively. Here we put $\mathsf{UNIF}[C]_0 := \{C\cdot\langle\rangle\}$.

Using these settings, we obtain a reformulation of Theorem 2.1.

Proposition 3.1. *An* $\mathbf{x}\in\mathfrak{M}(\mathcal{S}_n)$ *represent a MWIPF without multiple elements of size* C *if and only if* $\mathbf{x}\in\mathsf{MULT}[0,1]_n\cap\mathsf{UNIF}[C]_{\le n}$.

3.4 Formulation of Appending

An "Appending" is a corresponding rule from inputs $\mathbf{x} = \sum_{\sigma\in\varphi_k^n(L_k(H))} a(\sigma)\sigma \in \mathfrak{M}(\varphi_k^n(L_k(H)))$ to outputs $\mathbf{y} = \sum_{\tau\in\varphi_{k+1}^n(L_k(H))} b(\tau)\tau \in \mathfrak{M}(\varphi_k^n(L_k(H)))$ such that $\Phi_k^{k+1}(\mathbf{y}) = \mathbf{x}$. But we are interested in rules only those which are admissible as processes to produce MWIPF. They need not care about inputs violating uniformity of ealier stages, while they should guarantee that their outputs satisfy size uniformity with respect to this stage. This leads us to the following definition of classes of maps:

Definition 3.2 ($\mathcal{AP}[C]_k^H$).

$$\mathcal{AP}[C]_k^H := \left\{ \begin{array}{l} \mathrm{AP}_k^H : \Psi_k^H(\mathsf{UNIF}[C]_{\le k}\cap\mathsf{MULT}[0,\infty]_k) \to \mathfrak{M}(\varphi_{k+1}^n(L_k(H))) \\ \left| \begin{array}{l} \mathrm{AP}_k^H\left(\sum_{\sigma\in\varphi_k^n(L_k(H))} a(\sigma)\sigma\right) = \sum\limits_{\substack{\xi\in[n]\setminus H \\ \sigma\in\varphi_k^n(L_k(H))}} b(a,\sigma,\xi)\langle\sigma\xi\rangle, \\ \text{(i) } \forall\xi\in[n]\setminus H, \forall\sigma\in\varphi_k^n(L_k(H))\,[b(a,\sigma,\xi)\in\mathbb{Z}_{\ge0}], \\ \text{(ii) } \forall\xi\in[n]\setminus H\left[\sum_{\sigma\in\varphi_k^n(L_k(H))} b(a,\sigma,\xi) = C/((n-k)\tbinom{n}{k})\right], \\ \text{(iii) } \forall\sigma\in\varphi_k^n(L_k(H))\left[\sum_{\xi\in[n]\setminus H} b(a,\sigma,\xi) = a(\sigma)\right] \end{array} \right. \end{array} \right\} \tag{3.21}$$

Figure 2: CycleAp$[C]_k^H$

$$\begin{array}{cccc}
\hline
 & \sigma_1 & \xi_0 & \\
a(\sigma_1)\Big\{ & \sigma_1 & \xi_1 & \uparrow \\
 & \sigma_1 & \xi_2 & \\
 & \vdots & \vdots & \\
\vdots & \vdots & \xi_{n-k-1} & C/\binom{n}{k} \\
 & \vdots & \xi_0 & \\
 & \vdots & \vdots & \\
a(\sigma_{n!/(n-k)!})\Big\{ & \sigma_{n!/(n-k)!} & \xi_{n-k-2} & \downarrow \\
 & \sigma_{n!/(n-k)!} & \xi_{n-k-1} & \\
\hline
\end{array}$$

By definitions (3.11),(3.18) and (3.13) the above conditions (i),(ii) and (iii) are equivalent to:

$$\begin{array}{lll}
\text{(i)} \;\Leftrightarrow\; \text{(a):} & \mathrm{AP}_k^H(\mathbf{x}) \in \mathsf{MULT}[0,\infty]_{k+1}, & \\
\text{(ii)} \;\Leftrightarrow\; \text{(b):} & \psi_{k+1}^{H,\xi}(\mathrm{AP}_k^H(\mathbf{x})) \in \mathsf{UNIF}[C]_{k+1}^{H,\xi} & \text{for all } \xi \in [n]\setminus H, \\
\text{(iii)} \;\Leftrightarrow\; \text{(c):} & \Phi_k^{k+1}(\mathrm{AP}_k^H(\mathbf{x})) = \mathbf{x}. &
\end{array} \qquad (3.22)$$

In addition, we define the following subclass to exclude rules those which produce "MWIPF with multiple elements":

Definition 3.3 ($\mathcal{RAP}[C]_k^H$).

$$\mathcal{RAP}[C]_k^H := \{A \in \mathcal{AP}[C]_k^H \mid A \text{ satisfies (i),(ii),(iii) and}$$
$$\text{(iv) } [\mathbf{x} \in \mathsf{MULT}[0,(n-k)!]_k \Rightarrow A(\mathbf{x}) \in \mathsf{MULT}[0,(n-k-1)!]_{k+1}]\} \qquad (3.23)$$

Example 3.4 (CycleAp$[C]_k^H$). For all $0 \leq k < n$ and $H \in \mathcal{H}_{n,k}$, the following map CycleAp$[C]_k^H$ is an element of $\mathcal{RAP}[C]_k^H$ when $C \in \mathbb{Z}_{\geq 0}$ is divisible by $(n-k)\binom{n}{k}$.

Description of CycleAp$[C]_k^H$. Fix a numbering of $\varphi_k^n(L_k(H))$ and a numbering of $[n]\setminus H$ respectively; $\varphi_k^n(L_k(H)) = \{\sigma_1,\dots,\sigma_{n!/(n-k)!}\}$ and $[n]\setminus H = \{\xi_0,\dots\xi_{n-k-1}\}$ (numbering from 0 of the later set for convenience). For the input $\mathbf{x} = \sum_{\sigma\in\varphi_k^n(L_k(H))} a(\sigma)\sigma$ list each of σ_i with multiplicity $a(\sigma_i)$ (possibly zero) in a column. By the assumption that input $\mathbf{x}$ is in $\Psi_k^H(\mathsf{UNIF}[C]_{\leq k}) \subset \Psi_k^H(\mathsf{UNIF}[C]_k)$, the number of rows is $\sum_i a(\sigma_i) = w_k(\mathbf{x}) = \sum_{x\in H} w_k(\psi_k^{H\setminus\{x\},x}(\mathbf{x})) = k \cdot C/((n-(k-1))\binom{n}{k-1}) = C/\binom{n}{k}$.

Then create the second column by the cyclic sequence $\xi_0,\xi_1,\dots,\xi_{n-k-1},\xi_0,\xi_1,\dots$, i.e., fill the rth row of the second column by $\xi_{(r \bmod (n-k))}$ (Figure 2). Note that the cycle $\xi_0,\dots,\xi_{n-k-1}$ repeats exactly $C/((n-k)\binom{n}{k})$ times and ends with $\xi_{n-k-1}$, since the number of row $C/\binom{n}{k}$ is divisible by the period $n-k$. For each of σ_i and ξ_j, set the occurence of pair σ_i,ξ_j to $b(a,\sigma_i,\xi_j)$, and set $\mathbf{y} = \sum_{\xi\in[n]\setminus H,\sigma\in\varphi_k^n(L_k(H))} b(a,\sigma,\xi)\langle\sigma\xi\rangle$ to the output $\mathrm{AP}_k^H(\mathbf{x})$. Then it is easy to see that CycleAp$[C]_k^H$ $\in \mathcal{RAP}[C]_k^H$. □

3.5 The General Construction

Now we are ready to describe the general strategy to produce multiplicty-free MWIPFs of specific size:

Theorem 3.5. *Let $n > 0$ be an integer and let $0 < C \leq n!$ be an integer which is a multiple of* $\mathrm{lcm}(n, n-1, \dots, 1)$. *Fix a sequence of appending maps* $(\mathrm{AP}_k^H \in \mathcal{RAP}[C]_k^H)_{H\in\mathcal{H}_{n,k}}^{0\leq k<n}$. *Define the sequence* $(\mathbf{x}_k \in \mathfrak{M}(\mathcal{S}_{n,k}))_{k=0}^n$ *as follows:*

$$\begin{array}{ll}
(1): & \mathbf{x}_0 := C \cdot \langle\rangle \\
(2): & \mathbf{x}_{k+1} := \sum\limits_{H\in\mathcal{H}_{n,k}} \mathrm{AP}_k^H(\Psi_k^H(\mathbf{x}_k)) \quad (0 \leq k < n)
\end{array} \qquad (3.24)$$

then, $\mathbf{x}_k \in \mathsf{UNIF}[C]_{\leq k} \cap \mathsf{MULT}[0,(n-k)!]_k$ *for all* $(0 \leq k \leq n)$*. Especially,* $\mathbf{x}_n$ *is an element of* $\mathsf{MULT}[0,1]_n \cap \mathsf{UNIF}[C]_{\leq n}$*, i.e.,* $\mathbf{x}_n$ *represent an Exact MWIPF of size* C *without multiple elements.*

Proof. Note that $C/((n-k)\binom{n}{k}) \in \mathbb{Z}$ for all k (cf. [6] Lemma 3.1 or [2] Theorem 6). If $\mathbf{x}_k \in \mathsf{UNIF}[C]_{\leq k} \cap \mathsf{MULT}[0,(n-k)!]_k$, then it is easily checked that $\mathbf{x}_{k+1} \in \mathsf{UNIF}[C]_{\leq k+1} \cap \mathsf{MULT}[0,(n-(k+1))!]_{k+1}$ by the definition of $\mathcal{RAP}[C]_k^H$. □

3.6 Completeness of the Construction

Indeed, the strategy produce any MWIPF under suitable conmbination of appending maps.

Theorem 3.6. *Let* n *be a positive integer and let* $C > 0$ *be a multiple of* $\operatorname{lcm}(n, n-1, \ldots, 1)$*. If* $\mathbf{y} \in \mathfrak{M}(\mathcal{S}_n)$ *represents a MWIPF without multiple elements of size* C*, then there exist a sequence of appending map* $(\mathrm{AP}_k^H \in \mathcal{RAP}[C]_k^H)_{H \in \mathcal{H}_{n,k}}^{0 \leq k < n}$ *such that:*
if we define $\mathbf{x}_k \in \mathfrak{M}(\mathcal{S}_{n,k})(0 \leq k \leq n)$ *as in Theorem 3.5, then* $\mathbf{x}_n = \mathbf{y}$*.*

Proof. For each k and H, define

$$\mathrm{AP}_k^H(\mathbf{z}) := \begin{cases} \sum_{\xi \in [n] \setminus H} \psi_{k+1}^{H,\xi}(\mathbf{y}_{k+1}) & \text{if } \mathbf{z} = \Psi_k^H \circ \Phi_k^n(\mathbf{y}) \\ \mathrm{CycleAp}[C]_k^H(\mathbf{z}) & \text{otherwise.} \end{cases} \tag{3.25}$$

and check that $\mathrm{AP}_k^H \in \mathcal{RAP}[C]_k^H$. Then check $\mathbf{x}_k = \Phi_k^n(\mathbf{y})$ inductively. □

4 Concluding Remarks

In this paper, we reformulated a characterization of MWIPF as *size uniformity*, then presented a strategy to produce subsets of $\mathcal{S}_n$, which conform the output set to size uniformity. The strategy gives a construction of MWIPF when one fix a choise of appending maps. On the other hand, by suitable choise of appending maps, it produce any MWIPF. Thus the strategy is a surjection from the set of all combination of appending maps to the set of all MWIPFs. This means that we may understand that characteristic of a MWIPF is consist of local one (characteristic of each appending map) and global one (combination of appending maps).

For practical apprications such as estimating document similarity, smaller family size and effecient sampling are more desiable than exactness of Min-Wise Independency. In [2], they presented a number of possible relaxized versions of MWIPF (k-restricted, approximated, biased distribution, etc.) for this purpose. Indyk [4] presented a construction strategy of approximately Min-Wise Independent Permutation Families based on families of hash-functions , which is useful for the derandomization of the RNC algorithm [3]. Then it would be a problem: "For these relaxied versions of MWIPFs, are there analogs of the generic construction strategy?"

References

[1] Broder, A.Z., On the Resemblance and Containment of Documents," *Proc. of Compression and Complexity of Sequences*, pp.21-29 (1998).

[2] Broder, A.Z., Charikar, M., Frieze, A.M., and Mitzenmacher, M., "Min-Wise Independent Permutations," *Proc. of the Thirtieth Annual ACM Symposium on Theory of Computing*, pp.327-336 (1998).

[3] Broder, A.Z., Charikar, M., and Mitzenmacher, M., "A Derandomization Using Min-Wise Independent Permutations" *RANDOM'98*, pp. 15-24 (1998).

[4] Indyk, P. "A Small Approximately Min-Wise Independent Family of Hash Functions", *SODA '99*, pp. 454-456 (1999).

[5] Shinozaki, T., Itoh, T. "A Polynomial Time Sampling Algorithm for an Optimal Family of Min-Wise Independent Permutations", These proceedings.

[6] Takei, Y., Itoh, T., Shinozaki, T. "An Optimal Construction of Exactly Min-Wise Independent Permutations", *Technical Report of IEICE*, COMP98-62, pp.89-98 (1998).