POLYCHRONICITY TENDENCY-BASED ONLINE BEHAVIORAL SIGNATURE

Adeyemi Richard Ikuesan[1,*] · Shukor Abd Razak[2] · Hein S. Venter[1] · Mazleena Salleh[2]

[1] University of Pretoria, Pretoria, Gauteng, South Africa

[2] Department of Computer Science, Universiti Teknologi, Malaysia, Johor Bahru, Malaysia

*Correspondence to: Adeyemi Richard Ikuesan rikuesan@hotmail.com

Abstract

The proliferation of ubiquitous and pervasive computing devices has led to the emergence of research areas like Internet of things, and the Big-Data, which has seen a rise in obfuscation of online identity thus fueling an increase in online anonymity. Online anonymity constitutes a major platform for the exploitation of the potentials of cyber-crime; at the same time, it also inhibits the potential economic power that can be harnessed from the surging Internet population. Methods of online identification, such as usage profiling, demographic profiling, cookie-based identification process, media fingerprinting as well as token-based identification processes, are limited to either system identification or one-to-one identification. Current one-to-one identification mechanisms require huge volumes of templates of known users, and cannot be applied to novel users. This study proposed a psychosocial approach that integrates the human Polyphasia tendency into online identification processes for a one-to-many identification process. To achieve this, the study administered a Polychronic-Monochronic tendency scale measurement instrument to staff members of a research unit in a university, and the server-side network traffic of each respondent was monitored and collected. A logistic model tree was adapted for the one-to-many classification model based on human intrinsic features extracted from the network traffic and Polyphasia dichotomy. High degree of reliable accuracy of >80% was achieved which suggests a reliable model that supports the underlying hypothesis of the proposed model. Based on this accuracy, the approach finds practical relevance in online profiling process for online identification as well as online demographic profiling for e-commerce and e-learning. Furthermore, this approach can be applied to improve recommender systems in areas such as prediction and profile delivery through the extraction of the purpose of online surfing.

Keyword: Polyphasia tendency, logistic model tree, digital time preference signature, one-to-many identification, intrinsic network feature, online psychosocial fingerprint.

I. Introduction

Based on the observation in [1], as of 31st Dec. 2014, there is an estimate of ≈ 3billion Internet users, as against ≈360 million users by 31st Dec 2000. Given the increasing surge of online users, the capability of one-to-one online identification mechanism is gradually shrinking. Anonymization techniques in online interaction as offered by agents, such as the Tor network [2], further complicate the tendency of one-to-one identification process. A pioneering existential study presented in [3] on the online fingerprinting process and the subsequent inferential print assertion by [4]–[7] posits that individual behavioral patterns in online interaction can be adapted for online identification purposes. While these studies provide undeniable mitigation towards reducing online anonymity on the individual basis, the inherent characteristics of a human, such as a relationship between the human traits and navigation patterns, are not captured in any of these studies. Human traits in this regard refer to human psychological dispositions to preference, broadly referred to as soft biometrics, and are effective discriminant for differentiating individuals and consequently, identifying individuals. These soft biometrics are often neglected in typical technical approaches to the online identification process. A purely technical approach is based on the premise that a system under investigation is used by an individual, and any observed behavior originates from the assumed individual. Such premise will fail in a multi-individual environment characterized by the high probability of repudiation. Most cyber-crimes, corporate espionage, and cyber-bullying tend to increase with increasing repudiation tendencies [8]. The integration of human psychological makeup into the understanding of the online behavioral phenomenon is gradually gaining wider interest [9]. These social-technical approaches integrate a broad quantitative behavioral data into network traffic attributes [9], [10]. The approach has been widely adopted in areas like social media analysis, e-mail and blog pattern analysis, and Internet service usage. One underlying pattern in online behavior which closely aligns with a human tendency is the navigation and task execution pattern. Task execution can be viewed along the continuum of multitasking or unitasking, which is well defined by various extractions of Hall's time orientation theory [11]. Hall's construct of time orientation is defined by the Polychronicity model of culture. According to Hall's assertion [12], Polychronicity deals with:

1. Monochronic orientation (Monochronicity): the concept of time as a linear entity through which tasks are ordered in a strictly sequential manner. Individuals who share such orientation typically prefer to organize task one at a time and consider the process of switching from serial ordering of events as a distraction,
2. Polychronic orientation (Polychronicity): the concept of time as a nonlinear, random and rapidly changing entity that spans present, past and future, such that events are not necessarily sequential in nature. Such individual prefers to do several tasks and processes concurrently by switching between segmentations of time.

However, findings in [13] assert that Hall's temporality of Polychronicity underpins three independent aspects of time orientation which includes time-use preference, time use context, and time tangibility. The study argues that Hall's model of Polychronicity should be considered within a particular aspect of measured time orientation. While some studies have explored time orientation on the Internet, this study to the best of our knowledge is the first to explore the dimension of individual Polychronicity for online user identification processes. A brief discussion on the Internet and Polychronicity is presented in the next section.

II. Polychronicity and the Internet

Cultural anthropology observes time perception as the collective cultural artifacts that depict the generic behavior of people who share the same culture. However, time perception is studied from different research perspectives:

1. In consumer behavior and online marketing, time perception is treated as an economic entity characterized by the concept of time tangibility which views time as a scarce resource.
2. In management and organizational behavior, time perception is treated as a source of cross-cultural difference which is capable of explaining interactivity in the workplace.

These perceptions represent the broad dimensions which constitute the fundamentals of Hall's theory of Polychronicity. The study in [13] further classified these dimensions into time tangibility, time use context and time use preference. The study further asserts that these three dimensions are independent and that they can be described as time tangibility continuum, time context continuum, and Polyphasia continuum (generally termed as Monochronic and Polychronic time-use preference). Time tangibility continuum closely relates to the economic dimension of Polychronicity. On the other hand, Time context continuum closely relates to the multifaceted form of cultural communication. The Polyphasia continuum, however, relates to an individual tendency towards a preference for time structure with reference to the task. At one end of the continuum, there is the preference for sequentially linear formation of task and activity while at the other extreme is the preference for multitasking, parallel and a nonlinear sequence of tasks. Amongst these three dimensions, Polyphasia continuum finds directs relationship with the study on human behavior in online interaction [14]. This is because the Internet enables flexible task sequencing and supports (and can easily influence) multitasking. Based on the theoretical correlations between the internet and human time preference, study in [14] observed a bivariate correlation between Internet skill and Polychronicity using a 2-stage least square regression analysis. Findings in [15] observed a significant relationship between polychronicity and online navigation pattern, based on the assertion that linear and nonlinear human task structuring is inherent in human cultural development as measured by Hall's model of temporality. The study observed that Monochrons prefers linear navigation as opposed to nonlinear navigation preference by Polychrons. Furthermore, finding in [16] suggested that information technology (IT) provides optimal multitasking platforms, such that a trait-based preference for organizing computer related activities differs intuitively from individuals at different ends of the Polychronicity continuum. The notion of trait-based Polychronicity is further x-rayed from two differential periscopes: broad/stable trait, situation-specific/dynamic trait. Based on this logic, the study developed computer Polychronicity construct, which is observed to be significantly correlated with perceived usefulness of IT. Similarly, finding in [17] opines that online browsing pattern is inherently temporal which can be contrasted to the human temporal tendency. The study further argued that the existing studies on human Polychronicity sought to establish congruence between individual temporal orientation and preference in online related tasks. This logic further posits that there is a difference in the sensitivity of individuals to time-related issues. One area of time sensitivity in relation to the Internet is the human perception of delay in web application and downloads. The study in [12] observed that Polychrons are significantly less disoriented towards web-download delays in contrast to Monochrons. This finding [12] further lend credence to the asserted characteristic structure of linearity of Monochrons and nonlinearity of Polychrons, in [15]. The findings attributed fewer disorientation characteristics of Polychrons to the ability of Polychrons to engage in multiple online activities, which limit the perceived impact of download delay and consequently prevent the realization of delay in download time. Similarly, [15] observed a distinguishable pattern of navigation linearity between Polychrons and Monochrons. Monochrons tend to exhibit a sequentially linear navigation pattern, while Polychrons tend to exhibit nonlinear, circular navigation pattern. These studies can be described in terms of how an individual actually experiences and manages tasks with respect to time, or the synchronization and planning process within a particular group of individual that shares common demography, such as culture. Studies on Polychronicity and the Internet primarily centers on the former. The current study follows similar dimensions. However, some fundamental distinctions between this study and prior studies are identified as follows:

Fundamental Distinctions	**Prior Studies**	**Proposed study**
Research focus	Behavioral studies on time perception as it relates to the Internet is based on exploration of differences between Monochrons and Polychrons based on behavioral linearity, download delay, and Internet usage skill	Explores the probability of the existence of a distinguishable pattern between individuals along the Polychronicity continuum
Research Methods	Focus on the correlation between polychronicity and observed variables of the research focus. Thus, statistical significance and regression coefficients are observed. Most studies adopt self-report measurement instrument for both dependent and independent variables. [15] explored real-time weblog based on the assumption that Polychronicity is limited to cultural settings which neglected individual tendency in measuring Polychronicity.	Based on self-reported Polychronicity and Server-side network traffic of respondents. Furthermore, network features considered in this study are based on human-centric network features which are capable of revealing inherent navigation behavior, visitation pattern as well as request pattern. Collectively, these features contrast linearity and nonlinearity, sequential pattern and randomization
Result Evaluation	Based on statistical significance and regression coefficients.	Based on machine learning classification metrics. Hence, the result obtained in current study presents a more robust and integrative measure for dichotomy distinctions along the Polychronicity continuum.

The current study explores the probability of the existence of digital time-usage preference signature in online interaction which can be adapted to distinguish online users along the time-usage preference (Polychronicity) continuum. To achieve the aim of this study, the methodology presented in Figure 1 is adopted, while details of the processes are discussed in the next section. The methodology comprises three interdependent phases. Phase-1 entails the identification of organization and the subsequent process of data acquisition: network traffic and the data for Polyphasia continuum. Polyphasia continuum as measured by the polychromic-Monochronic tendency scale (PMTS) is identified and carefully assessed in the second phase of the methodology. The third phase, Phase-3, involves the extraction of human-centric features, the definition of a class for subsequent supervised machine learning process, and exploration of applicable supervised machine learning algorithms. Supervised machine learning is used in this study because the respective classes/categories of the user along the Polyphasia continuum is known a priori.

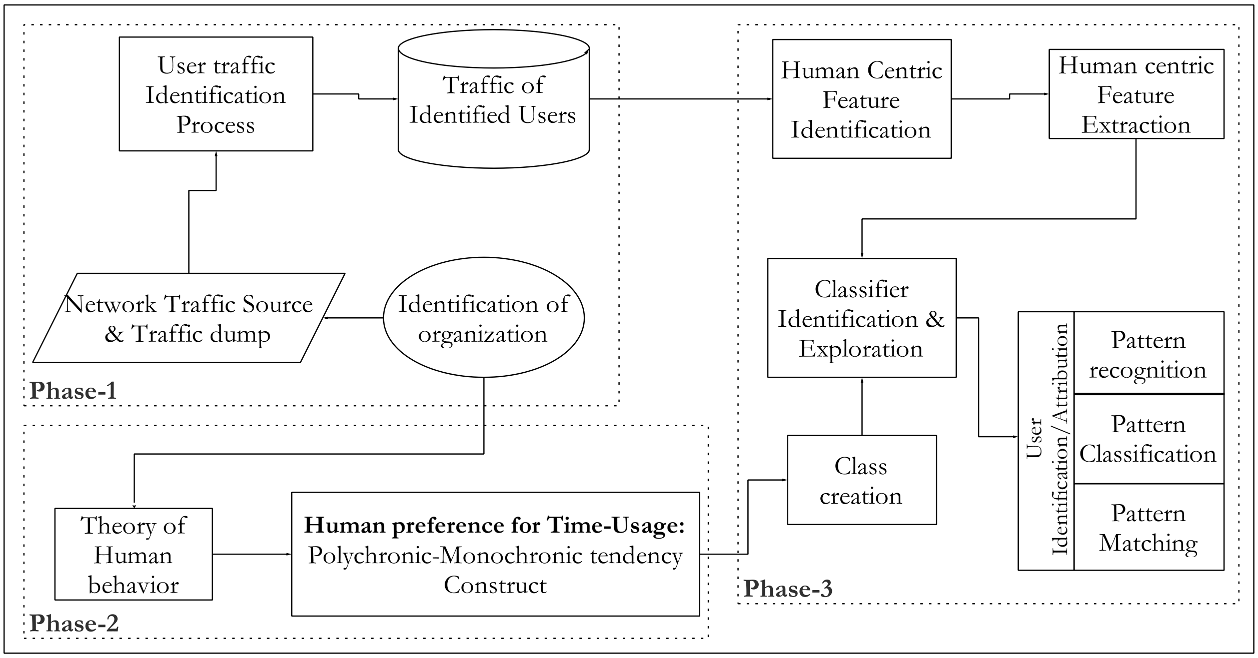

Figure 1: Methodology for the proposed User Identification model

III. Research Methods

The term Polyphasia as introduced by [13] is a strict term of one of the dimensions of Hall's Polychronicity which refers to the time usage preference of an individual. Polyphasia tendency as a construct has been widely applied to the measure of individual Polychronicity. This study adopts the Polychronic-Monochronic tendency scale (PMTS) developed in the study in [18] as an instrument to measure individual Polyphasia tendency. The choice of PMTS is due to its observed reliability (as reflected in the measured goodness of fit indices such as CFI, GFI, AVE, and RMSEA). Furthermore, it does not depend on situation-specific tendency which suggests replicability. PMTS also measures individual Polychronicity as a reflective model with strong validity test (discriminant and nomological). PMTS measures individuals along the Polyphasia continuum ranging from 1, which indicates highly monochromic tendency to 5, which indicates highly Polychronic tendency. In order to extract the dichotomous class of Polyphasia continuum, this study assumes that a typical Polyphasia data follows a normal distribution. Since percentile represents area under the normal distribution curve, this study adopted a quartile dichotomy as depicted in Figure 2. The adapted cut-off in Figure 2 is independent of the distribution of the data. This is to prevent data-centric dichotomy and also to aid experimental repeatability. A similar but less granular dichotomy is adopted in [14].

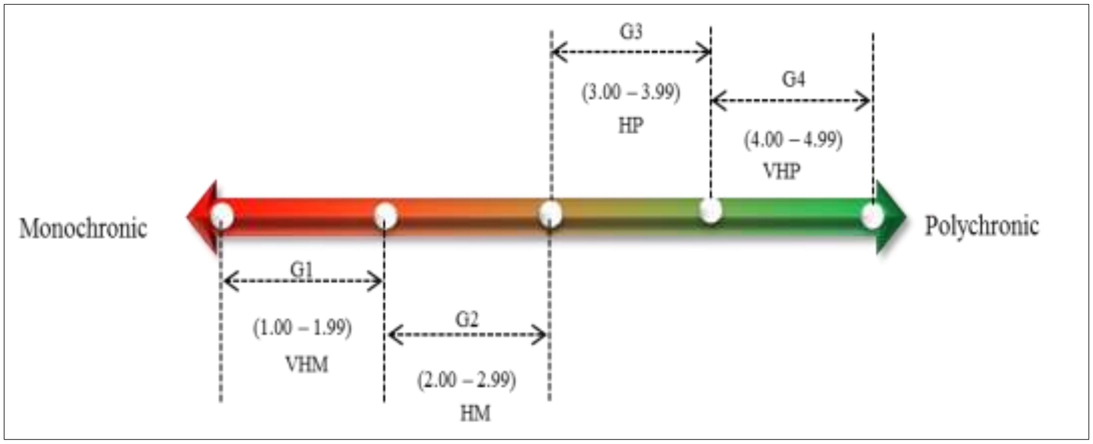

Figure 2: Cutoffs for Polyphasia Dichotomy Derivation

The term dichotomy is used in the loosed sense in this manuscript to mean categorization or classification of an individual into four different regions along the Polyphasia continuum. As shown in Figure 2, the adapted Polyphasia dichotomy has four classes – very high Monochrons, high Monochrons, high Polychrons and very high Polychrons, whereby each represents a generic dichotomy applicable to any Polyphasia respondents. To recruit respondents for this study, two basic requirements were established. First, the respondent must be an actively employed staff member of an organization with the capability of frequent client-server communication using the Internet service. Second, the respondent must actively communicate using the client-server communication process of the organization using login identity that is strictly used by only the respondent. Based on these two criteria, a Research Management Centre of a research University in Malaysia was selected. Initial consent forms were distributed to all staff members of the organization. A total of 66-respondents returned a completed consent form indicating their readiness to participate in the study, amongst which a total of 55-respondents submitted a completely filled 5-items PMTS measurement instrument. 43-respondents satisfied the criteria for inclusion in this study. PMTS construct description and statistical summary of the response of the 43-respondents are presented in Table 1. Details of the path diagram which shows the correlation coefficient of the model is presented in the supplementary file.

Table 1: PMTS Description and statistical summary of response

Item	**PMTS Description**	**Parameter (N=43)**		**Value**
1	Poly-mono behavior preference	Minimum		1
2	Reported poly-mono Behavior	Maximum		5
3	Comfort with poly-mono behavior	Mean		3.36
4	Liking to juggle simultaneous activities	Median		3.6
5	Poly-mono behavior as most efficient way to use time	Std. Deviation		0.93
Class	**PMTS Dichotomy**	Variance		0.86
1.00	Very High Monochronicity (VHM)	Skewness		-0.41
2.00	High Monochronicity (HM)	Percentiles	25th	1.8
3.00	High Polychronicity (HP)		50th	3.6
4.00	Very High Polychronicity (VHP)		75th	4

The distribution of the response as captured in the percentiles showed that 25% of the respondents fall between the VHM and HM dichotomy at score value of 1.8. Conversely, the distribution indicates that about 25% of the respondents are within the VHP dichotomy at score ≥ 4. The server-side network traffic of the 43-respondents was collected from April 2014 to December 2014. A heuristic methodology was developed to clean the raw log file of requested URL and to extract relevant human-centric features. The heuristics consider web requests that originate as a result of human action, in contrast to requests initiated by system or network facility on behalf of the individual. The heuristics was applied to individual requests and the following human-centric features were extracted based on a 30-minutes session boundary which is the generally accepted session duration [5], [19]. Two categories of network traffic characteristics feature space were extracted. The first category is a unidimensional time series which was adapted for extracting online vocabulary signature of each dichotomy. The second category comprises behavioral characteristic features, such as web request pattern, web page visitation pattern as well as session characteristics. The second category data was used to observe the probability of distinction among all observed dichotomies.

A. Web Request Characteristics

Individual request pattern was observed through the inter-request characteristic-pattern observed in each session. Inter-request time is the time difference between two consecutive requests within a session. Statistical properties of web request characteristics as defined in [20], which include mean, standard deviation, variance, kurtosis, and skewness of individual web request were extracted from each session. A total of 10 human-centric features were extracted from the web request characteristics. The inter-request (also referred to as flight time) pattern of each user was processed as a unidimensional time series, which can be described as follows: suppose the time for initial request is t_i and the time elapsed between when a respondent submits a request and when the request gets to the server is given as δt, the flight time between the requests is defined by the expression in (1).

$$\mathrm{F}_i \;\square\; t_{i+1} - t_i \tag{1}$$

The time series of each respondent in the same dichotomy is further examined for common subsequence (which is generally referred to as sequitur) among individuals in each dichotomy. In order to extract the sequitur, a symbolic aggregate approximation (SAX) technique is applied to the individual time series. Common subsequence in SAX is based on three parameters: piecewise approximate aggregation (PAA), sliding window size (SWS), and alphabet size (AS). Parameter optimization for extraction of optimal common subsequence for an individual user is obtained using a GrammarViz2.0 tool [21]. In order to extract the sequitur signature in each dichotomy, two logics present applicability as depicted in Table 2. The observed unique sequitur (if it exists) is then defined as the signature for that dichotomy.

Table 2: Fundamental logic for Vocabulary Signature

Step	**LOGIC-I**	**LOGIC-II**
1	Combine the inter-request time series of all respondents in the same dichotomy such that each dichotomy is a unidimensional time series	Compute sequitur from the inter-request time series data of individual respondents in each dichotomy
2	Compute sequitur for each dichotomy	Compute the similarity among individual respondents in the same dichotomy
3	Compute the dissimilarity among the dichotomies	Compute dissimilarity among the dichotomies

The intersection of Set (set theory) is used to observe similarity of the sequitur among individuals in each dichotomy, while the difference of set is used as the dissimilarity measure of sequitur among dichotomies. Assume $\cap_{VHP}$, $\cap_{HP}$, $\cap_{HM}$ represents the extraction of vocabulary from dichotomies HVP, HP and HM respectively, the probability of the existence of a unique vocabulary for each dichotomy can be expressed using Bayes theorem of the form depicted in (2).

$$Signature = \begin{cases} \boldsymbol{P}\left(\cap \boldsymbol{VHP} \mid (\cap \boldsymbol{HP} \times \cap \boldsymbol{HM})\right) = \left(\boldsymbol{P}\left((\cap \boldsymbol{HP} \times \cap \boldsymbol{HM}) \mid \cap \boldsymbol{VHP}\right) \times \boldsymbol{P}(\cap \boldsymbol{VHP})\right) / \boldsymbol{P}(\cap \boldsymbol{HP} \times \cap \boldsymbol{HM}) \text{ for } P_r \text{ of } VHP \\ \boldsymbol{P}\left(\cap \boldsymbol{HP} \mid (\cap \boldsymbol{VHP} \times \cap \boldsymbol{HM})\right) = \left(\boldsymbol{P}\left((\cap \boldsymbol{VHP} \times \cap \boldsymbol{HM}) \mid \cap \boldsymbol{HP}\right) \times \boldsymbol{P}(\cap \boldsymbol{HP})\right) / \boldsymbol{P}(\cap \boldsymbol{VHP} \times \cap \boldsymbol{HM}) \text{ for } P_r \text{ of } HP \\ \boldsymbol{P}\left(\cap \boldsymbol{HM} \mid (\cap \boldsymbol{HP} \times \cap \boldsymbol{VHP})\right) = \left(\boldsymbol{P}\left((\cap \boldsymbol{HP} \times \cap \boldsymbol{VHP}) \mid \cap \boldsymbol{HM}\right) \times \boldsymbol{P}(\cap \boldsymbol{HM})\right) / \boldsymbol{P}(\cap \boldsymbol{HP} \times \cap \boldsymbol{VHP}) \text{ for } P_r \text{ of } HM \end{cases} \tag{2}$$

The probability presented in (2) illustrates the computation of the posterior probability of a particular dichotomy from the given prior probability and the conditional probability, based on mutual exclusivity.

B. Visitation Pattern

The University Centre operates a two server, load-balancing client-server communication architecture. This implies that the possible number of the probable web page is bounded by the total number of web pages in the two servers as represented in (3).

$$URL_{Total} = \sum_{i=1}^{s}\left[\sum_{j=1}^{N} URL_j\right] \tag{3}$$

where s = total number of server, N = number of unique URL on each server. This study assumes that the individual web request pattern obeys power law as asserted in [22], [23]. Visit characteristics considered in this study include aggregation of visit within a session, the rate of revisit per session and the session length with respect to visit aggregation. These are depicted in (4), (5), and (6) respectively.

$$V_{agg} = \frac{\sum_{i=1}^{n}(URL\ per\ session)_i}{\sum_{j}^{N}(URL\, under\, observation)_j} \tag{4}$$

$$R_{vs} = \frac{\sum_{i=1}^{n}(URL\ per\ session)_i}{S_d}, S_d \Rightarrow session\ duration = \int_{j=1}^{n} t_j dt, \leq 30 minutes \tag{5}$$

$$S_{agg} = \frac{\sum_{j}^{N}(URL\, under\, observation)_j}{S_d} \tag{6}$$

The notion of rate of the visit is in conformity with expression in (3), based on the logic that the probable URLs visitable by an individual are limited to the observable URLs on the server. In addition, this presupposes that the interest-driven model and priority queue model of probable request patterns [23] are captured in the bounded URL distribution such that all observed users share similar working conditions, and the major observable distinction can be revealed through observation of the innate behavioral composition. Three features were derived from the visitation pattern. In addition, session duration and the total number of request per session were also derived. A total of 15-features were extracted from the network traffic.

The observed duration of server-side data collection was divided into pattern observation (training) phase and pattern validation phase. 21-weeks and 15-weeks were adopted for model training and validation phases respectively. In order to observe the distinction among the observed dichotomies, six supervised machine learning algorithms are explored. The choice of the selected six classifiers is based on initial exploration of applicable classifiers on the extracted features. The six classifiers include logistic regression model, logistic model tree (LMT), C4.5 (J48) decision tree, Reduce Error pruning decision tree (REPTREE), Decision table Naïve Bayes (DTNB) and Partial decision tree (PART). Discussion of these classification algorithms can be found in [24]–[26]. The process adapted for classification exploration in this study is similar to the defined process in [25] as illustrated in Figure 3. The exploration process involves the search of applicable classifiers capable of establishing discriminative boundaries among classes in the dataset based on the informative capability of the computed features. The process starts with the pre-processing of the data. This process involves data cleaning, the extraction of a sequence of request and sessionization of the request, based on the defined session threshold. The next stage involves splitting the dataset into training and testing samples. This is followed by a classifier exploration process. The default baseline for the exploration process is based on the highest class prior-probability which can be measured by the ZeroR algorithm in Waikato Environment for Knowledge Analysis

(WEKA®) workbench. The WEKA software is adapted for the classifier exploration process in this study. WEKA software is a Java-based open source software which has gained wider adoption in pattern classification and machine learning process due to its robustness [26] and the within-script automation capability [27]. The experimental process implemented in this study was based on an accuracy obtained using 10-fold cross validation and 10-iterative process to prevent overfitting. Conventional performance evaluation measures are adapted to evaluate each overall performance of each classifier. These include accuracy, Kappa statistics, root mean square error (RMSE), Precision, recall, F-measure and Area Under the receiver operating characteristic Curve (AUC) [26]–[28].

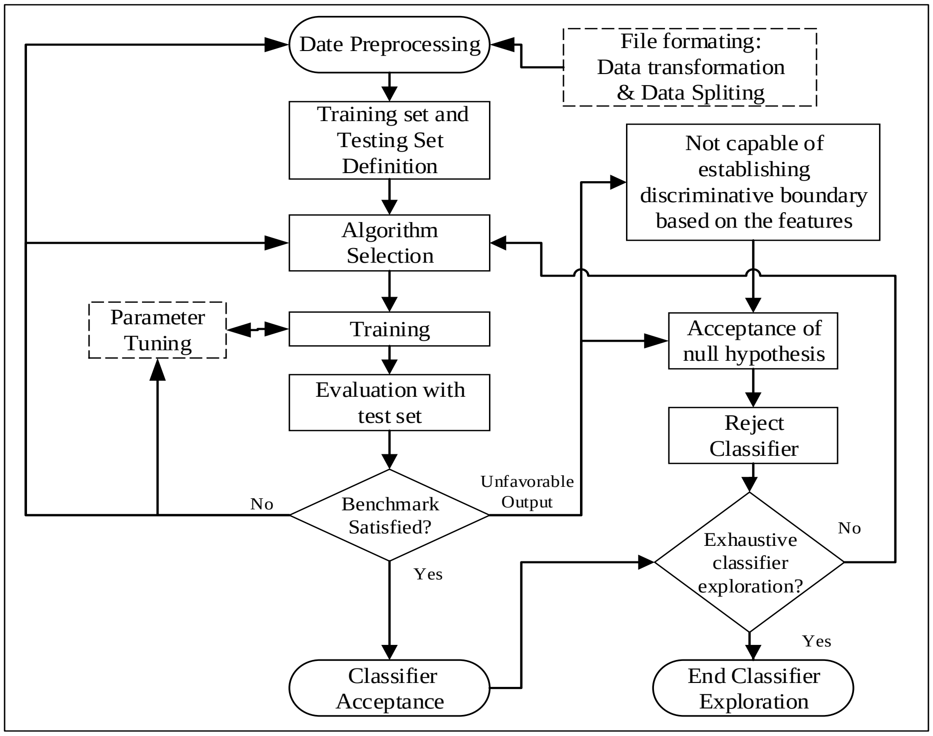

Figure 3: Procedure for Classifier Exploration

IV. Result and Analysis

The result of this study is presented in two dimensions. The first being the probability of extracting sequitur signature (vocabulary) based on the inter-request pattern among individual in the same dichotomy, and the second dimensions explore the probability of extracting behavioral signature based on the dichotomous categorization, using the default configuration in the WEKA framework. The dichotomy cutoff defined in Figure 2 was applied to the response from the 43-respondents. The four dichotomies –VHM, HM, HP, and VHP– were extracted from the distribution of the respondents. Summary of the network traffic statistics of the observed dichotomy is presented in Table 3.

Table 3: Statistical Summary of Network traffic of each dichotomy

Dataset	Dichotomy	Statistics						
		Respondents	Total Session	Total Request	Min Request	Max Request	Min Session	Max Session
Training Dataset	VHP	15	2162	71136	76	36476	7	366
	HP	15	1954	62021	166	15894	22	359
	HM	11	1675	53664	344	23163	29	341
	VHM	2	252	1984	651	1333	107	145
Validation Dataset	VHP	15	2162	71136	76	36476	7	366
	HP	15	1954	62021	166	15894	22	359
	HM	11	1675	53664	344	23163	29	341
	VHM	2	220	1880	560	1327	108	112

In order to ascertain the inter-cluster dissimilarity and the intra-cluster similarity among the extracted dichotomies, further statistical tests –f-test and t-test– were carried out on each extracted dichotomy. F-test is a measure of equal variance of observed variables, while t-test is a measure of the equal mean of variables. The results of the f-test and t-test, as presented in Table 4, strongly support the reliability of the extracted dichotomies. However, the VHM cluster was dropped among the dichotomies due to its relatively low sample size (number of respondents).

Table 4: Summary of Statistics of f-test and t-test

Variables	VHP	HP	HM	VHM
VHP				
HP	A,R			
HM	A,R	A,R		
VHM	*R*,*R*	A,R	A,R	

A= acceptance of null hypothesis R= rejection of null hypothesis.
The sequence A,R represents f-test, t-test respectively. Where A implies an acceptance of null hypothesis of an f-test which implies that there is no statistically significant difference in the variance of the variables under investigation, and R implies the rejection of the null hypothesis of a t-test which indicate that there is a statistically significant difference between the mean of the observed variables

The t-test was carried out based on an assumed equal variance as justified by the rejection of the null hypothesis of the f-test in all variables, except at the intersection of VHP and VHM. This is important because equality in variance closely relates to the notion of intra-cluster similarity within each dichotomy. Similarly, acceptance of the t-test closely relates to the notion of inter-cluster dissimilarity among the different dichotomies. Based on the observed statistical validity, the probability of a vocabulary signature and behavioral signature is thus explored. This is presented and discussed in the next section of this manuscript.

A. Probability of Vocabulary Signature

The extraction of common subsequence using SAX sequitur was achieved based on the 32:4:4 parameter configuration for SWS, PAA and AS respectively. This configuration gave the optimal common subsequence for each individual in each dichotomy. The vocabulary signature explored in this study was based on *LOGIC-II*, as defined in Table 2. The result of the application of *LOGIC-II* is shown in Table 5.

Table 5: Summary of vocabulary signature

Parameters	**VHP**	**HP**	**HM**
Total number of respondents	15	15	11
Actual respondents Used	10	11	6
Minimum Sequitur threshold	44	47	63
Maximum Sequitur Observed	595	820	564
Total Number of Sequitur	2380	3068	1911
Number of similar sequiturs	4	5	4
Number of Unique Sequitur	2	4	3

Total Number of sequitur refers to step-1, number of similar sequitur refers to step-2 and number of unique sequitur refers to step-3 of LOGIC-II, as defined in Table 2.

The results in Table 5 indicate the existence of vocabulary signature. The actual number of respondents represent the number of respondents used in extracting sequitur from the total number of respondents as shown in the distribution depicted in Figure 4. This is based on a threshold selection function defined in (7). The choice of 40-rules was based on the distribution of the observed sequitur. This implies that a respondent whose number of sequitur is less than 40-rules is discarded in the signature formation process.

$$Sequitur\,threshold = Number\,of\;\,Rules \geq 40\,rules\,of\;\,common\,subsequence \tag{7}$$

This further ensures that sufficient numbers of rules are compared in each dichotomy. On the other hand, it also implies that respondents with a higher number of rules are reduced to the size of the respondents with the minimum rule.
The total number of sequitur for each dichotomy is based on the summation of all sequiturs of individual respondents in the dichotomy, in conformity to *LOGIC-II* of Table 2. The number of similar sequiturs refers to the common sequitur (set-intersection) among individual respondents in the same dichotomy. Conversely, the number of unique sequiturs is the set-difference in the number of common sequitur in one dichotomy that is not present in other dichotomies.

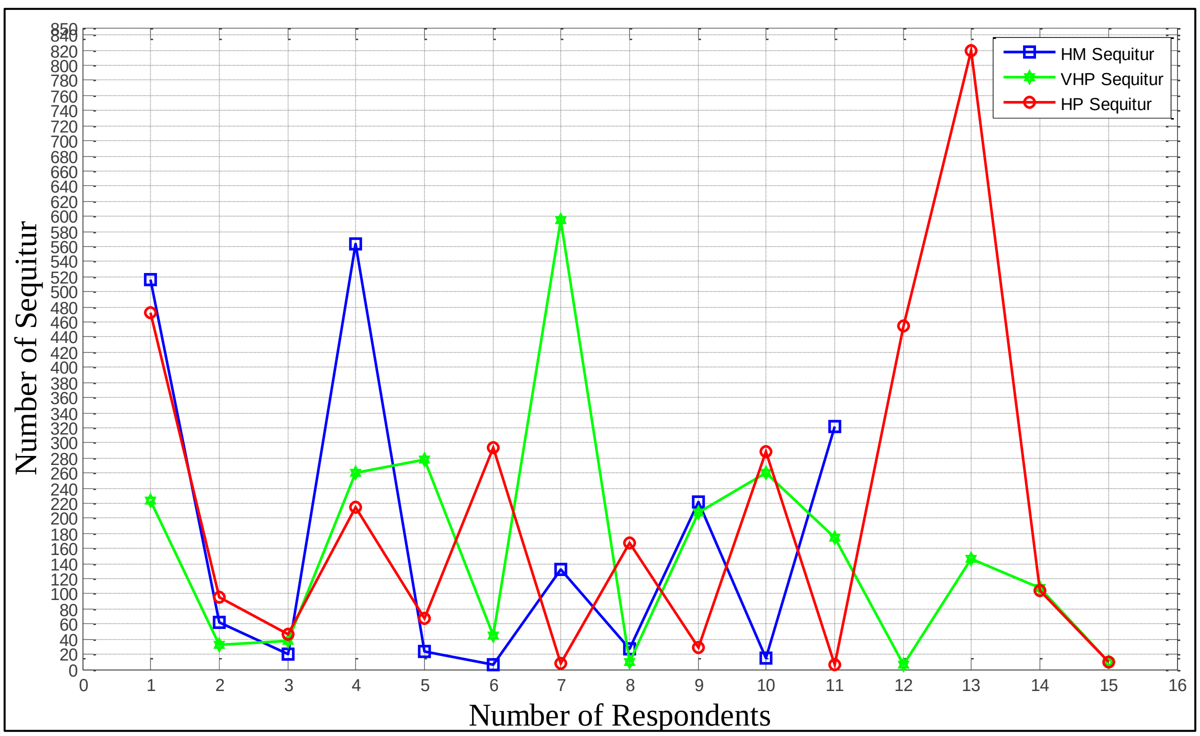

Figure 4: Distribution of Extracted Sequitur

Based on the unique sequitur extracted from each dichotomy, there is the probability of the existence of digital vocabulary signature using human request pattern. The number of similar sequiturs in each dichotomy is considerably small (4:5:4 for VHP: HP: HM respectively) relative to the total number observed. As a means to further explore the existential probability of digital time-use preference signature, the next section observed the behavioral composition and the likelihood of such existence based on the combination of human intrinsic features which are capable of revealing behavioral tendency, using machine learning approach.

B. Probability of Behavioral signature

The procedure defined in Figure 3 is explored in this section based on the extracted features. The features comprise 15 human-centric descriptions in online interaction. Class membership is defined based on the number of dichotomies extracted from the Polychronic Monochronic tendency scale as presented in Table 3. The six classifiers identified in Section 3 were explored on the training and validation datasets. Results of the exploration are divided into two subsections based on the dataset. The performance of the classifiers on the training dataset is presented in Table 6.

Table 6: Experimental Result of Training Dataset

Classifier	Accuracy	Kappa Stats	RMSE	Precision	Recall	F-Measure	AUC
ZeroR	39.71(0.05)	0.00(0.00)	0.47(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.50(0.00)
Logistic	41.34(0.99)	0.05(0.02)	0.47(0.00)	0.42(0.04)	0.17(0.02)	0.24(0.02)	0.55(0.02)
DTNB	70.20(2.92)	0.54(0.05)	0.35(0.01)	0.74(0.03)	0.65(0.05)	0.69(0.04)	0.88(0.02)
PART	68.23(4.19)	0.51(0.07)	0.37(0.02)	0.76(0.09)	0.69(0.11)	0.71(0.06)	0.89(0.03)
J48	76.35(1.51)	0.64(0.02)	0.34(0.01)	0.82(0.02)	0.79(0.02)	0.81(0.02)	0.92(0.01)
LMT	**84.30(1.72)**	**0.76(0.03)**	**0.29(0.01)**	**0.86(0.02)**	**0.84(0.02)**	**0.85(0.02)**	**0.94(0.01)**
REPTree	77.83(2.39)	0.66(0.04)	0.32(0.02)	0.80(0.03)	0.80(0.03)	0.80(0.03)	0.94(0.01)

9809 instances, 10-fold cross validation, 10-iterations, Confidence (0.0001) two-tailed

The experimental results shown in Table 6 shows that the LMT classifier exhibit superior classification accuracy over all other classifiers. The baseline classifier, ZeroR, obtained an accuracy equivalent to the highest class prior probability. The class prior probability for VHP, HP, and HM classes are 379.71%, 33.12%, and 27.1% respectively. Therefore, the LMT classifier achieved an average accuracy of 84.30% which is higher than other classifiers. The reliability of the achieved accuracy is supported by the values of F-measure (averaged at 0.85) and AUC (averaged at 0.94). F-measure and AUC range from the poor performance of 0 to superior performance of 1. Similarly, the value of RMSE and Kappa statistics averaged at 0.29, and 0.76 respectively and lend credence to the reliability of the accuracy of LMT classifier. For RMSE, a value closer to 0 indicates reliable accuracy. REPTree performed closer to LMT with an average accuracy of 77.83%. Other performance metrics, particularly AUC, indicates that the obtained accuracy has a high reliability. J48 and DTNB classifier also demonstrate high accuracy relative to the baseline accuracy. In order to ascertain the consistency of the performance of the classifiers, further experimental process is carried out on the validation data, as shown in Table 7.

Table 7: Experimental Result of Validation Dataset

Classifier	Accuracy	Kappa Stats	RMSE	Precision	Recall	F-Measure	AUC
ZeroR	37.33(0.06)	0.00(0.00)	0.47(0.00)	0.37(0.00)	1.00(0.00)	0.54(0.00)	0.50(0.00)
Logistic	38.79(1.63)	0.05(0.03)	0.47(0.00)	0.39(0.02)	0.64(0.03)	0.48(0.02)	0.56(0.02)
DTNB	52.46(2.95)	0.28(0.05)	0.43(0.01)	0.52(0.03)	0.72(0.04)	0.60(0.02)	0.72(0.03)
PART	79.33(3.98)	0.69(0.06)	0.31(0.02)	0.80(0.07)	0.87(0.08)	0.83(0.04)	0.93(0.02)
J48	**85.96(1.71)**	**0.79(0.03)**	**0.28(0.02)**	**0.85(0.02)**	**0.93(0.02)**	**0.89(0.02)**	**0.94(0.01)**
LMT	**85.20(1.72)**	**0.78(0.03)**	**0.28(0.01)**	**0.85(0.02)**	**0.91(0.02)**	**0.88(0.02)**	**0.95(0.01)**
REPTree	**82.39(2.81)**	**0.73(0.04)**	**0.30(0.02)**	**0.82(0.03)**	**0.90(0.03)**	**0.86(0.02)**	**0.95(0.01)**

5791 instances, 10-fold cross validation, 10-iterations, Confidence (0.0001) two-tailed

The results presented in Table 7 are based on the class prior probability of 37.33%, 33.74% and 28.92% for VHP, HP and HM classes respectively. The baseline classifier is equal to the highest class prior probability with an accuracy of 37.33% (accuracy based on chance). J48 and LMT achieved relatively similar accuracy averaged at 85.96% and 85.20% respectively. Other performance metrics, particularly the AUC and F-measure, lend credence to the similarity of the performance of J48 and LMT. The result of the validation dataset substantiates the result obtained in the training data set in Table 6. Based on the validation dataset, J48 shows superior classification accuracy, higher Kappa statistics, and higher F-measure. However, the LMT classifier exhibits consistency in classification accuracy and reliability based on the AUC value averaged at 0.95. A plot of the AUC value of the classifiers, presented in Figure 5, shows that LMT exhibits consistent classification accuracy which is more effective than the J48 classifier. A detailed dissection of the accuracy of LMT classifier based on the validation data set is presented in Figure 6. The LMT classifier achieved 91.4% accuracy in classifying the VHP dichotomy, based on a class prior probability of 37.33%. This translates into accuracy strength of 54.07%. In other words, the LMT classifier can efficiently distinguish online individual with an accuracy of 8.63 instances out of every 10 instances. Similarly, HP and HM can be correctly classified in 7.16 and 7.15 instances respectively, out of every 10 instances. The accuracy strength (out of every 10 instances) presented in Figure 6 is based on the instance calculation expression presented in (8).

$$IA_{10} = \left\{ \left(Accuracy - Prior\ Prob\right) \Big/ \left(100 - prior\ prob\right) \right\} \times 10 \qquad (8)$$

The instance accuracy (IA) calculation expression indicates the capability of a classifier to accurately identify an instance (per 10-instance in this case). The result proves that the LMT classifier is capable of distinguishing online subjects along the VHP, HP and HM dichotomies.

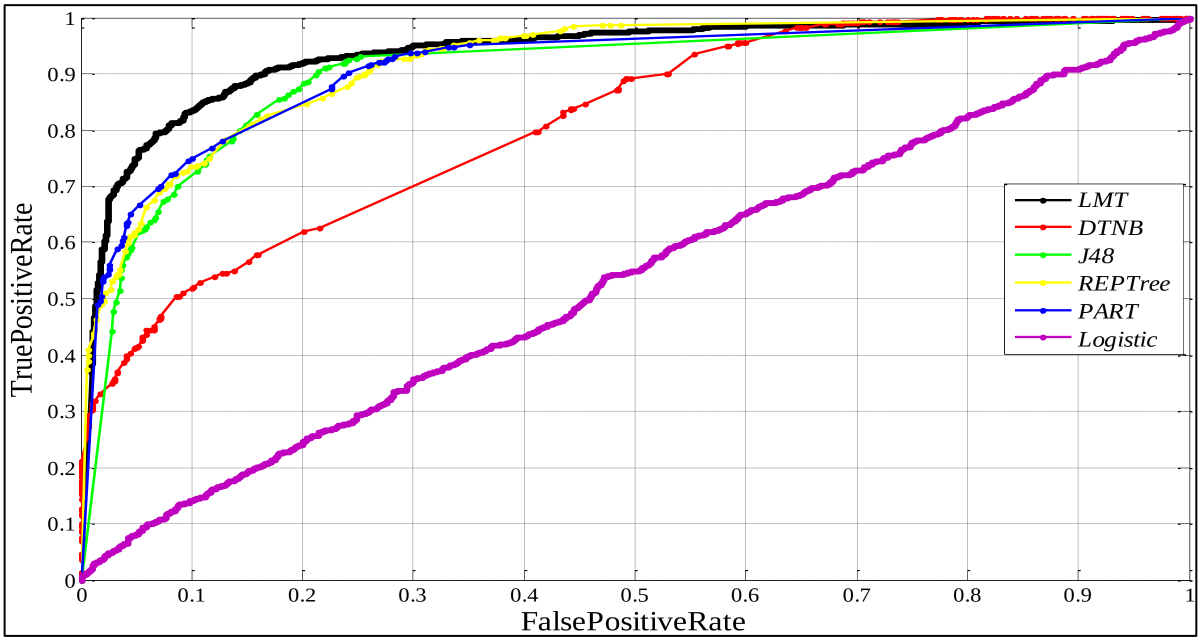

Figure 5: Comparative Analysis of AUC of Explored Classifiers

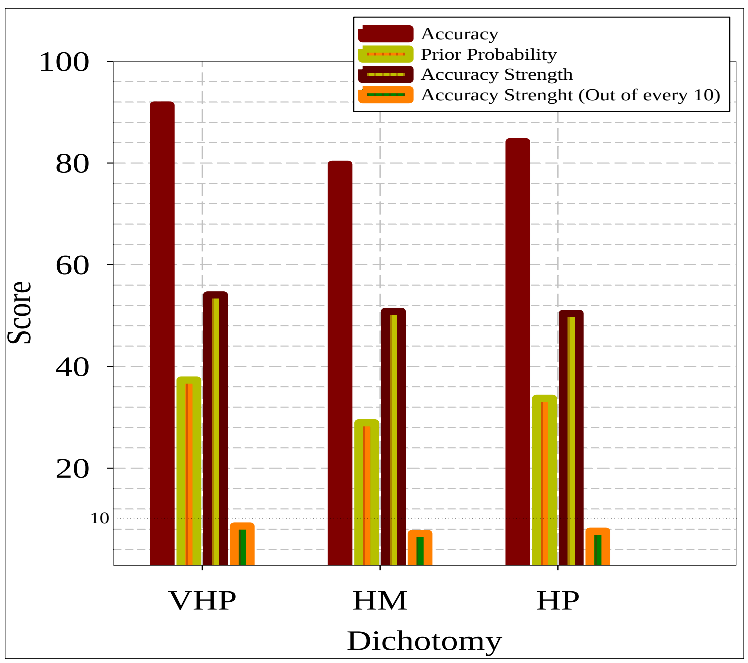

Figure 6: Analysis of Accuracy of LMT classifier

V. Discussion

The result shown in Table 6 and Table 7 consistently demonstrates the accuracy of the logistic model tree. A logistic model tree (LMT) classifier is a hybrid classifier that integrates a linear logistic regression model into a decision tree (DT) classification mechanism. Classification is achieved by generating decisions with logistic models at its leaves, and the prediction estimate is obtained by the use of posterior class probability. The integration of DT into LMT demonstrates its superiority over the linear regression model when applied to the highly multidimensional data set that requires ease of human interpretability. The performance of DTNB classifier is inferior to LMT classifier in this study.

DTNB is an integration of Naïve Bayes algorithm into decision table mechanism. An initial experiment based on Naïve Bayes shows a very poor performance. Naïve Bayes classifier assumes that all attributes in the dataset are independent. The capability of LMT classifier to infer larger structural knowledge from the high dimensional data set can be attributed to its superiority over DTNB. PART is a rule-based induction algorithm which builds decision tree by avoiding global optimization in order to reduce the time and processing complexities. PART uses the separate-and-conquer approach of RIPPER and combines it with the decision tree mechanism of C4.5 by removing all instances from the training dataset that are covered by this rule and proceeding recursively until no instance in the dataset remains. PART builds a partial decision tree for the current set of instances by choosing Leafs with the largest coverage as the new rule. Similarly, the Reduced Error Pruning Tree (REPTree) classifier applies regression tree logic and generates multiple trees in altered iterations by sorting values of numeric attributes once, based on information gain principle (which measures the expected reduction in entropy). REPTree prune trees based on reduced-error pruning with the back fitting method, and integration of C4.5 mechanism for missing value by splitting each corresponding instances into fractional instances. A J48 decision tree is a java coded version of C4.5 decision tree implemented in the WEKA workbench. The C 4.5 decision tree learning is an inductive inference process which recursively partitions instances of a given attribute space. Classification of an instance is done by constructing nodes that form root tree using singular incoming edges to link nodes while supporting multiple outgoing edges through the predefined discrete function of input attribute value. The performance of the LMT classifier shows superiority over J48 in terms of the measured parameters in Table 4 and Table 5, particularly in terms of the AUC. This is further shown in the average number of correctly classified and incorrectly classified instances in Table 8. A granular observation of the resultant model shows that LMT generated smaller size of rules for classification relative to J48. However, the time taken to build the LMT model is significantly higher than the time taken to build the J48 decision tree model.

Table 8: Comparative Analysis of Performance of Classifiers

Classifier	Avg. Correct	Avg. Incorrect	Avg. Time	Number of rules/trees
ZeroR	389.5	591.4	0.01	Not applicable
Logistic	405.51	575.39	4.93	Not applicable
DTNB	688.62	292.28	41.04	83
PART	669.24	311.66	19.38	251
J48	748.92	231.98	2.04	1201(601)
LMT	826.91	153.99	62.8	735(368)
REPTree	763.39	217.51	0.52	889

10-fold cross validation, 10-iterations, Confidence (0.0001) two-tailed

As shown in Table 8, the number of rules generated for classification is not a direct indicator of the effectiveness of a classifier. This is evident in the number of rules generated by LMT (with 735 rules), J48 (with 1201 rules), REPTree (with 889 rules) and PART (with 251 rules) in decreasing order of performance. A sample of the classification process of the LMT classifier is represented in a partial decision tree and the logistic model as presented in Figure 7. The result shows that individuals exhibit structural patterns that integrate the following feature combinations on the PMTS scale: when an individual satisfies the decision condition specified by the partial decision tree (line 2- line 28 of Figure 7), the individual can be represented by the leaf node of the logistic model LM_130:15/75 (256). This implies that an individual can be categorized into HP, HM or VHP using the classification equation presented in (9).

$$\begin{cases} HP = 6.87 + \left[Fligth\,Kurt\right] \times -0.01 + \left[Rateof\ Visit\right] \times 22.74 + \left[RateofVisitperSession\right] \times 35.5 \\ HM = -3.91 + \left[Fligth\,Kurt\right] \times 0.02 + \left[TotalNo\right] \times -0.02 + \left[IntervalKurt\right] \times 0.02 + \left[RateofVisit\right] \times 8 + \\ \qquad \left[RateVisitperSession\right] \times -14.06 + \left[RateofVisitCountperSession\right] \times -0.63 \\ VHP = -8.36 + \left[Flight\ Kurt\right] \times 0.02 + \left[Total\ No\right] \times 0.02 + \left[RateofVisit\right] \times -117.83 + \left[RateVisitperSession\right] \times 1.17 \end{cases} \tag{9}$$

These rules demonstrate how the LMT classifier can be used to classify an individual on the Internet into an HP- a dichotomy, HM- dichotomy or VHP- dichotomy.

Partial Decision Tree of LMT classifier

```
Rate of Visit Count per Session <= 1.592275
| Rate of Visit Count per Session > 1.592275
| Rate of Visit Count per Session <= 7.636594
| | Rate of Visit Count per Session > 7.636594
| | Rate of Visit Count per Session <= 8.850932
| | | Session Duration <= 1374
| | | | Flight SD <= 9.956741: LM_119:15/90 (17)
| | | | Flight SD > 9.956741
| | | | | Session Duration <= 1059
| | | | | | Session Duration <= 816
| | | | | | | Session Duration <= 315
| | | | | | | | Flight SD <= 36.920994
| | | | | | | | | Interval Mean <= 18: LM_120:15/165 (14)
| | | | | | | | | Interval Mean > 18: LM_121:0/150 (3)
| | | | | | | | Flight SD > 36.920994: LM_122:15/150 (8)
| | | | | | | Session Duration > 315
| | | | | | | | Flight Kurt <= 18.829262
| | | | | | | | | Rate of Visit Count per Session <= 8.477752: LM_123:15/165 (34)
| | | | | | | | | Rate of Visit Count per Session > 8.477752: LM_124:15/165 (16)
| | | | | | | | Flight Kurt > 18.829262: LM_125:0/135 (3)
| | | | | | Session Duration > 816: LM_126:15/120 (70)
| | | | | Session Duration > 1059
| | | | | | Rate of Visit Count per Session <= 8.20904
| | | | | | | Session Duration <= 1112: LM_127:15/135 (8)
| | | | | | | Session Duration > 1112: LM_128:15/135 (30)
| | | | | | Rate of Visit Count per Session > 8.20904: LM_129:15/120 (23)
| | | Session Duration > 1374: LM_130:15/75 (256)
```

HP = Class 0 :

```
6.87 +
[Session Duration] * 0 +
[Flight Kurt] * -0.01 +
[Total No] * 0 +
[Interval Mean] * 0 +
[Interval Skew] * 0 +
[Rate of Visit] * 22.74 +
[Rate Visit per Session] * 35.58 +
[Rate of Visit Count per Session] * 0
```

HM = Class 1 :

```
-3.91 +
[Session Duration] * 0 +
[Flight Skew] * 0 +
[Flight Kurt] * 0.02 +
[Total No] * -0.02 +
[Interval Mean] * 0 +
[Interval SD] * 0 +
[Interval Var] * 0 +
[Interval Skew] * 0 +
[Interval Kurt] * 0.02 +
[Rate of Visit] * 8 +
[Rate Visit per Session] * -14.06 +
[Rate of Visit Count per Session] * -0.63
```

VHP = Class 2 :

```
-8.36 +
[Session Duration] * 0 +
[Flight Skew] * 0 +
[Flight Kurt] * 0.02 +
[Total No] * 0.02 +
[Interval Mean] * 0 +
[Interval SD] * 0 +
[Interval Var] * 0 +
[Rate of Visit] * -117.83 +
[Rate Visit per Session] * 1.17 +
[Rate of Visit Count per Session] * 0
```

Figure 7: Partial Decision Tree of LMT

Since this study could not extract classes for VHM dichotomy, it is logical to assume that novel signature of individuals who belong to such dichotomy on the PMTS continuum are not captured in the present study. However, this result is particularly insightful because it uses a DT for the classification process. A DT offers several advantages over other types of classifiers. This includes ease of interpretation by humans, ease of representation and application. Furthermore, a DT does not require a prior assumption about the structure of the data but builds knowledge based on the structure of the data.

As a way to further explore the probability of higher accuracy, ensemble meta-classifiers are further explored using three classifiers that performed consistently higher in the validation phase presented in Table 7, as the base classifiers. Bagging, LogitBoost and AdaBoost classes of algorithms are bootstrapping aggregation techniques widely used for constructing ensemble technique, and Bagging technique is considered superior in situations which involve substantial classification noise [29], [30]. This supposition is also observed in this study. Bagging meta-classifier was observed to show significant improvement among the explored classifiers. The average result of the Bagging meta-classifier is presented in Table 9. The result shows a slight improvement on the LMT and J48 as a base classifier. However, a significant improvement was observed with REPTree classifier, from an average of 82.39% accuracy in Table 7 to 85.59% accuracy in Table 9.

Table 9: Evaluation based on Bagging Meta-classifier Technique

Metrics/Dataset	Baseline		Bagging-J48		Bagging-LMT		Bagging-REPTree	
	TD	***VD***	***TD***	***VD***	***TD***	***VD***	***TD***	***VD***
Accuracy	39.71	37.33	77.45	**86.73**	85.65	**85.85**	81.79	**85.59**
Kappa Stats	0.00	0.00	0.66	0.80	0.78	0.79	0.72	0.78
RMSE	0.47	0.47	0.32	0.26	0.27	0.27	0.30	0.27
Recall	0.00	0.37	0.82	0.85	0.87	0.85	0.84	0.84
F-Measure	0.00	0.54	0.81	0.89	0.86	0.8	0.84	0.88
AUC	0.50	0.50	0.95	0.96	0.97	0.96	0.96	0.96

TS: Training Dataset, VD: Validation data set Baseline classifier: ZeroR (highest class prior probability)

To the best of our knowledge, this is the first study (as an extension of the preliminary study in [31]) that explored online Polychronic tendency based on the integration of a generic demographic dichotomy and human-centric network pattern classification. The result obtained from this study could have practical implication for online identification process as well as online demographic profiling process (as further discussed below). The result from the vocabulary signature suggests the existence of online browsing pattern which is consistent for an individual on different dichotomy on the PMTS continuum. This thus implies that the inter-request patterns of an online-surfer may be adapted to the online identification process to enhance the reliability of authentication mechanism. Similarly, results from the classification of human-centric features lend support to the existence of the digital PMTS pattern. Taken together, the results imply that human Polychronic-Monochronic digital tendencies exist in the similitude of human attributional characteristics, such as in physiological and psychological signature. Authentication in today's online platform struggles with identification mechanism which often requires password or token pattern memorization, captcha enforcement, dictionary phrase complexities as well as behavior verification mechanism. Vocabulary signatures may be integrated as a complementary mechanism to enhance security, especially in a multi-factor approach to authentication. Furthermore, the current online mechanisms for profiling of online behavior (such as recommender and E-learning systems), suffer setbacks in identifying and classifying novel behavioral pattern. In addition, the mechanisms require a huge volume of databases for the template matching process. The findings from this study present a complementary platform for classifying online users by using generic patterns of different dichotomies. One area of practical application is in the integration into user attribution process in the digital forensic investigation. This is further illustrated in Figure 9. User attribution during a digital investigation can harness the reliability of the digital signature extracted from the Polyphasia dichotomies to complement the behavioral signature of other sources. This can be readily deployed in an insider investigation processes.

As highlighted in Section 1, the aim of this study is to explore the probable existence of digital polyphasia tendency, which can be used as digital marker or signature. The study is not designed to specifically identify individual tendency. However, the extension of the findings to accurately identify an individual remains a question for further studies, which could require the integration of other human-centric characteristics, particularly from a client-side data source. This may also require a larger sample size as well as a higher volume of network traffic. This is evident from the result of the vocabulary signature. A larger sample size is required for a more reliable and accurate signature database. In addition, the frequency of occurrence of vocabulary signature is not observed in this study. Such can be used to further enhance the accuracy of the identification process.

One major limitation of this study is that it could only define three dichotomies. Larger sample size can provide a means to explore more granular dichotomies as against the dichotomy defined in Figure 2. Moreover, the measurement instrument, PMTS, does not provide insight into linear and nonlinear browsing pattern. The integration of linear/nonlinear pattern exploration is a fundamental behavior of online users. A more appropriate measurement instrument which considers human linearity in addition to online Polychronicity is therefore needed to fully understand online Polychronic tendencies.

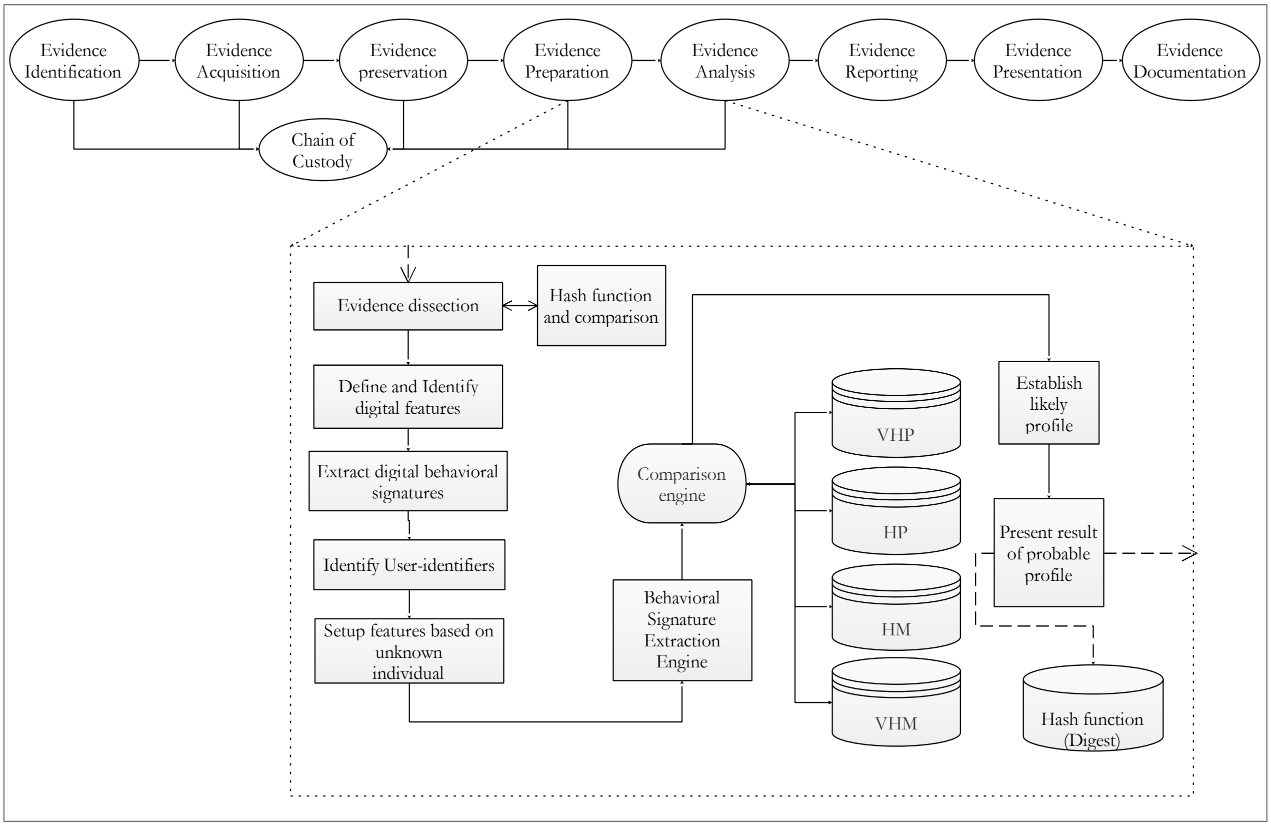

Figure 9: Process of User Attribution in Digital forensics based on Polyphasia tendency

Other areas of probable negligence are the introduction of the sentimental component into Polychronicity. Sentimental analysis of human behavioral preference would further reveal, in addition to individuality, the emotional and psychological state of the individual. Such analysis process could significantly improve profile delivery of online user. A sentimental analysis process involves the decomposition of the representative sentimental attributes of online behavior. These sentimental attributes can then be transformed into sentiment vectors which sums-up to represent individual affective composition. Li and colleagues [32] explored similar approach, albeit, from the perspective of prediction of stock-exchange market behavior. Other aspects of affective analysis that can be explored involve the association of Polychronicity as a behavioral composition for the recognition, understanding, and inference of emotional and opinion-related analysis in online interaction. This synergy has the potential to leverage the innate description of the human time preference for the prediction of online user behavior. Furthermore, this approach can be developed to provide a complementary corpus for automated user-identification in online chat-group platform, in similitude to Sentic-computing process of automatic text and opinion mining as highlighted in the exposition presented in [33].

Moreover, the identification and integration of more discriminative features present a process for the improvement of dissimilarity among different dichotomies. On the behavioral signature, the observed accuracy of LMT, J48, and REPTree Bagging meta-classifier can be further improved. Furthermore, given that the observable behavioral tendencies within the given 21-weeks and 15-weeks data categorization could be influenced by circumstantial factors and other non-measurable dependencies such as work stress, emotional fluctuation as well as seasonal changes, we further evaluate the combination of the 21-weeks and 15-weeks. The resultant combination can then be passed through a resampling filter (without repetition) of training ratio validation (e.g, a 60:40 resampling) for training and testing respectively. Other complex classification processes, such as ensemble classifiers based on stacking, could be further explored to extract higher accuracy for the behavioral signature. Classifiers that can leverage the semantic relationship in the data using posterior and prior probability could also be considered. Such classifiers include the structural minimax probability machine (SMPM). The discriminative power, and the efficiency of the SMPM to reveal the semantic composition of the information in a given data has been thoroughly evaluated in [34]. The perspective of online behavior exploration for understanding human behavior on the Internet presents a new paradigm with the potential to evolve into a medium for achieving a robust personalized solution to online profiling. A proper dissection of the PMTS-identity of an end-user opens researchers to a more fundamental discourse on the underlying causation of network burstiness, the probability of sophisticated artificial networks capable of replicating as well as mimicking human dynamics; in affective computing. Furthermore, it opens researchers to better comprehend the evolution of Internet consumption. In terms of online security, PMTS introduces a complementary platform for online identification mechanism, specifically in the one-to-many authentication process. The integration of PMTS (vocabulary and behavioral) signatures on identification and authentication process further suggest a robust and reliable security mechanisms laced with inherent human characteristics.

VI. Conclusion

This study explored the probability of the existence of digital Polychronic-Monochronic Tendency Scale (PMTS) on the Internet, based on the integration of human-centric network features and the administration of PMTS measurement instrument. Vocabulary and behavioral signatures of dichotomized respondents were used to explore the probability of the existence of the PMTS signature. A symbolic aggregate approximation (SAX) sequitur was adapted for vocabulary signature. The observed sequitur for each dichotomy was further compared to other classes using set difference theory. This study further explored PMTS signature based on behavioral characteristics in online communication. Six classifiers were initially explored, amongst which, the logistic model tree (LMT) classifier performed better. Further exploration based on Bagging ensemble technique on LMT, J48, and REPTree classifiers, further substantiate the observed performance. Behavioral signatures were extracted from each dichotomy. The result obtained in this study confirms the probability of the existence of a digital PMTS in online interaction. As a behavioral trait, human preference along the Polychronic-Monochronic continuum is a viable option for implementation, as a complementary measure for online user attribution. This can, therefore, aid in user authentication, user profiling in electronic media, as well as user investigation in forensics processes.

Acknowledgement

We would like to thank the Ministry of Education, Malaysia for sponsoring this research grant (Vote: R.J130000.7813.4F193), Universiti Teknologi Malaysia and University of Pretoria. The preliminary version of this article has been published in ASC 2017 in conjunction with BIGCOMP 2017 [31]

Reference

[1] M. G. Miniwatts, “Internet Usage Statistics: The Internet Big Picture. World Internet Users and 2015 Population Stats,” 2015. [Online]. Available: http://www.internetworldstats.com/stats.htm. [Accessed: 04-Aug-2015].

[2] Tor-Project, “Tor: Anonymity Online,” 2015. [Online]. Available: https://www.torproject.org/index.html.en. [Accessed: 04-Aug-2015].

[3] B. Padmanabhan and Y. Yang, “Clickprints on the web: Are there signatures in web browsing data?,” *Available SSRN http//ssrn.com/abstract=931057 or http//dx.doi.org/10.2139/ssrn.931057*, 2007.

[4] Y. (Catherine) Yang, “Web user behavioral profiling for user identification,” *Decis. Support Syst.*, vol. 49, no. 3, pp. 261–271, Jun. 2010.

[5] Y. Yang and B. Padmanabhan, “Toward user patterns for online security: Observation time and online user identification,” *Decis. Support Syst.*, vol. 48, no. 4, pp. 548--558, 2010.

[6] D. Herrmann, C. Gerber, C. Banse, and H. Federrath, *Analyzing characteristic host access patterns for re-identification of web user sessions*, ISBN: 978-., vol. 7127 LNCS. Espoo, Finland: Springer, 2010.

[7] M. Abramson and D. Aha, “User Authentication from Web Browsing Behavior,” in *The Twenty-Sixth*

International FLAIRS Conference, 2013, pp. 268–273.
[8] I. Baggili and M. Rogers, “Self-Reported Cyber Crime: An Analysis on the Effects of Anonymity and Pre-Employment Integrity,” *Int. J. Cyber Criminol.*, vol. 3, no. 2, pp. 536–549, 2009.
[9] B. Brown, C. Lampe, K. Rodden, and N. Ducheneaut, “Models, theories and methods of studying online behaviour,” in *CHI’10 Extended Abstracts on Human Factors in Computing Systems*, 2010, pp. 4449–4452.
[10] Y. Amichai-Hamburger, G. Wainapel, and S. Fox, “‘On the Internet no one knows I’m an introvert’: extroversion, neuroticism, and Internet interaction.,” *Cyberpsychol. Behav.*, vol. 5, no. 2, pp. 125–8, Apr. 2002.
[11] Y. Zhang, R. S. R. Goonetilleke, T. Plocher, and S. S. M. S. S. M. Liang, “Time orientation and human performance,” *Hong Kong Univ. Sci. Technol. Kowloon*, no. 1999, pp. 247–252, 2004.
[12] G. M. Rose, R. Evaristo, and D. Straub, “Culture and consumer responses to web download time: A four-continent study of mono and polychronism,” *IEEE Trans. Eng. Manag.*, vol. 50, no. 1, pp. 31–44, 2003.
[13] D. K. Palmer and F. D. Schoorman, “Unpackaging the multiple aspects of time in polychronicity,” *J. Manag. Psychol.*, vol. 14, no. 3/4, pp. 323–345, 1999.
[14] W. Lee, T. Tan, and S. Hameed, “Polychronicity, the Internet, and the mass media: A Singapore study,” *J. Comput. Mediat. Commun.*, vol. 11, pp. 300–316, 2006.
[15] A. Kralisch, M. Eisend, and B. Berendt, “The Impact of Culture on Website Navigation Behaviour,” in *The 11th International Conference on Human-Computer Interaction (HCI International)*, 2005, pp. 1–9.
[16] D. Joshua M., L. S. Lee, and M. Y. Yi, “Time-user preference and technology acceptance: Measure development of computer polychronicity,” *Am. J. Bus.*, vol. 24, no. 2, pp. 23–32, 2009.
[17] N. Capdeferro, M. Romero, and E. Barberà, “Polychronicity: review of the literature and a new configuration for the study of this hidden dimension of online learning,” *Distance Educ.*, vol. 35, no. 3, pp. 294–310, 2014.
[18] J. D. Lindquist and C. Kaufman-Scarborough, “The Polychronic Monochronic Tendency Model: PMTS scale development and validation,” *Time Soc.*, vol. 16, no. 2–3, pp. 253–285, Sep. 2007.
[19] R. Kumar and A. Tomkins, “A characterization of online browsing behavior,” in *Proceedings of the 19th international conference World Wide Web*, 2010, pp. 561–570.
[20] I. R. Adeyemi, A. S. Razak, and M. Salleh, “A psychographic framework for online user identification,” in *International Symposium on Biometrics and Security Technologies (ISBAST)*, 2014, pp. 198–203.
[21] P. Senin, J. Lin, X. Wang, and T. Oates, “GrammarViz 2.0: a Tool for Grammar-Based Pattern Discovery in Time Series,” in *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2014, pp. 468–472.
[22] A. Barabasi, “The origin of bursts and heavy tails in human dynamics,” *Nature*, vol. 435, no. 7039, p. 207–211., 2005.
[23] T. Zhou, X. Han, and B. Wang, “Towards the Understanding of Human Dynamics,” *Sci. matters Humanit. as complex Syst.*, no. 1, pp. 207–233, 2008.
[24] T. T. T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” *Commun. Surv. Tutorials, IEEE*, vol. 10, no. 4, pp. 56–76, 2008.
[25] S. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” *Informatical*, vol. 31, pp. 249–268, 2007.
[26] F. M. Othman, T. Moh, and S. Yau, “Comparison of Different Classification Techniques Using WEKA for Breast Cancer,” in *3rd Kuala Lumpur International Conference on Biomedical Engineering*, 2007, vol. 15, pp. 520–523.
[27] Y. C. Yang, “Web user behavioral profiling for user identification,” *Decis. Support Syst.*, vol. 49, no. 3, pp. 261–271, 2010.
[28] I. R. Adeyemi, S. A. Razak, M. Salleh, and H. S. Venter, “Observing consistency in online communication patterns for user re-identification,” *PLoS One*, vol. 11, no. 12, pp. 1–27, 2016.
[29] E. Bauer, R. Kohavi, P. Chan, S. Stolfo, and D. Wolpert, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,” *Mach. Learn.*, vol. 36, no. August, pp. 105–139, 1999.
[30] T. G. Dietterich, “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees,” *Mach. Learn.*, vol. 40, pp. 139–157, 2000.
[31] I. R. Adeyemi, S. A. Razak, H. S. Venter, and M. Salleh, “High-level online user attribution model based on human Polychronic-Monochronic tendency,” in *2017 IEEE International Conference on Big Data and Smart Computing, BigComp 2017*, 2017, pp. 445–450.
[32] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, “News impact on stock price return via sentiment analysis,” *Knowledge-Based Syst.*, vol. 69, no. 1, pp. 14–23, 2014.
[33] E. Cambria, R. Speer, C. Havasi, and A. Hussain, “SenticNet : A Publicly Available Semantic Resource for

Opinion Mining," *Artif. Intell.*, vol. 10, pp. 14–18, 2010.
[34] B. Gu, X. Sun, and V. S. Sheng, "Structural Minimax Probability Machine," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 7, pp. 1646–1656, Jul. 2017.