Identity Attributes Metric Modelling based on Mathematical Distance Metrics Models

Felix Kabwe[1], Jackson Phiri[2]
The University of Zambia
School of Natural Sciences
Lusaka, Zambia

Abstract—**Internet has brought a lot of security challenges on the interaction, activities, and transactions that occur online. These include pervasion of privacy of individuals, organizations, and other online actors. Relationships in real life get affected by online mischievous actors with intent to misrepresent or ruin the characters of innocent people, leading to damaged relationships. Proliferation of cybercrime has threatened the value and benefits of internet. Identity theft by fraudsters with intent to steal assets in real space or online has escalated. This study has developed a metrics model based on distance metrics in order to quantify the credential identity attributes used in online services and activities. This is to help address the digital identity challenges, bring confidence to online activities and ownership of assets. The application forms and identity tokens used in the various sectors to identify online users were used as the sources of the identity attributes in this paper. The corpus toolkits were used to mine and extract the identity attributes from the various forms of identity tokens. Term weighting schemes were used to compute the term weight of the identity attributes. Other methods used included Shannon Entropy and the Term Frequency-Inverse Document Frequency scheme (TF*IDF). Standardization of data using data normalization method has been applied. The results show that using the Cosine Similarity Measure, we can identify the identity attributes in any given identity token used to identify individuals and entities. This will help to attach the legitimate ownership to the digital identity attributes. The developed model can be used to uniquely identify an online identity claimant and help address the security challenge in identity management systems. The proposed model can also identify the key identity attributes that could be used to identify an entity in real or cyber spaces.**

Keywords—Mathematical modeling; Cosine Similarity Measure; text frequency; inverse document frequency; cyber space; term weight; internet; digital identity; trust model; normalization; text mining

I. INTRODUCTION

Challenges of identifying internet users associated with valuables that are online have become a serious concern to internet users. The adverse challenges on information security regarding identification of real identity ownership on internet and to services and online activities is of great concern. This research has developed a metrics model based on distance metrics in order to quantify the credential identity attributes used in online services and activities. The model will help in improving cyber security in digital identity management.

This study has reviewed literature that is relevant to the work so as to establish what passed efforts in this area have covered. Areas that have been explored include effects of internet to society and studies that help to understand what identity is from various disciplines. Various forms of identities have been considered which form partial identities, these would have an impact on identity of a person or entity. Consideration of what identity would imply on online services and activities has been looked at so as to have a relevant context in this study. Digital identity is an aspect that is dependent on trust; it is imperative to reflect on trust framework so as to bring to the fore on how the digital identity and trust are inter related. A large part of online activities includes communication of information; we therefore, had to reflect on communication trust model which would be applicable to our study and see the value that it would add to our study. We have reflected on Shannon's Communication trust framework from Shannon's Information theory to guide us in considering digital identity with respect to trust in online activities. Since our work is premised on mathematical modeling, it was imperative that we draw our attention to mathematical modeling and how it could influence our work. This research includes text mining from different documents, the mining would give outcomes that would include errors on data from different backgrounds of the different documents, whose sources are varied. To remove errors which at time would be due to measurement units, noise, and estimations, standardizing of data would be important before we use it in our metrics.

Mathematical modeling has been used in science in finding solutions in real life problems, this study takes interest in mathematical modeling. Using a mathematical model, a solution is being proposed to attend to the challenges that have been encountered in cyberspace concerning digital identity security. The study will use the proposed model on mined data to quantify the identity attributes. We will use the model to verify identification of the owner of digital identity; we will further test the model and establish which identity attributes are key in a given corpus for the identification of an identity claimant.

Literature that was reviewed showed that vector space model uses a storage matrix where columns represent the documents in a collection and rows represent terms in a document. Term frequencies of a given document would help us establish important identity attributes which would identity an entity. Literature indicated that there is a variety of schemes on term weights (attribute importance) which would help us establish whish terms are important in a given token of identity. Some of the schemes (or information retrieval methods) include Shannon's entropy and Term Frequencies - Inverse Document Frequency (TF-IDF). Our interest is to develop a digital identity model that would supply trusted digital identities. The other literature that was reviewed was that on multifactor authentication systems on identity attributes metrics models. This was to help us consider efforts that have been used in the past on augmented efforts. Literature on International Standards regarding identity attributes and identity tokens to appreciate the value this would have on our work was considered. This was to establish the international standards that affect identity attributes and identity tokens which are subject of our study.

Identity attributes were mined from identity documents and application forms for identity enrolment. Such documents in PDF format were extracted from internet using TalkHelper PDF Converter. Text was then mined from these documents using AntConc 3.5.8, a corpus analysis toolkit for data mining. To remove error, data was normalized for standardization of data.

The proposed model was used for identity attribute quantification and verification. The proposed model was also used to determine term importance in the corpus. Distance metrics has been the basis of our model to quantify the identity attributes. The model would identify attributes that are very key as identifiers of an entity, in other words, these are attributes that can closely identify an entity in online activities. Results of our study have been given and conclusion of the study has been drawn.

II. Literature Review

A. Adverse Effects of Internet

The rapid development of information and communication networks by governments, colleges, enterprises and individuals means that they are employing more and more information systems without clear distinctions of the persons and devices behind their use [1]. It is obvious that the need for identity that would provide complete privacy is vital [1]. It has been established that cybercrime has become one of the fastest growing crimes in the world [2]. Study has showed that computer networks are subject to attacks from malicious sources, with the advent and increasing use of internet attacks are most commonly increasing [3]. In 2007 it was reported that, "in Australia alone the proceeds of identity theft, [was] still one of the largest sources of fraud, [and was] estimated to be nearly $6 billion a year [4]". Identity theft is one of the fastest growing crimes in the world. Security includes protecting individuals, organizations, devices and infrastructure from identity theft, unauthorized data sharing and human rights violations [5]". When devices are lost or stolen, all of the data stored on or accessible from the mobile device may be compromised if access to the device or the data is not effectively controlled [6].

B. Partial Identity

To appreciate identity, we need to consider that a wholesome identity is formed by partial identities. A person may have different identities according to the context in which the identity is applied. For instance, a researcher may be a father, magazine columnist, human right activist, sportsman, politician, philanthropist, friend, and lecturer. He is identified differently, and attributes that make him identified accurately may differ from one context to the other. Fig. 1, illustrates partial identities. A comprehensive identity could be assumed by identifying key characteristics of an individual which we would attribute to be identity attributes.

Identity encompasses all the essential characteristics that make each human unique [3]. An identity of father of this individual may have characteristics of: father of three, kind, loving, hardworking, protective, supportive, merciful, jovial, progressive, etc. The identity of a person comprises a large number of personal properties [3], as indicated above. These properties help to uniquely identify an individual.

C. Digital Identity

It is indicated in [7] that "a digital identity is a virtual representation of a real identity that can be used in electronic interactions with other machines or people". An identity consists of traits, attributes, and preferences upon which one may receive personalized services". E-services require an effective way to manage digital identity information of the users [7].

Windley defines a digital identity as the "data that uniquely describes a subject or an entity and the ones about the subject's relationships to other entities [8]". Further, Windley states that a digital identity is "the persona that an individual presents across all the digital spaces [8]". In [9], we define digital identity as the "electronic representation of personal information of an individual or organization (name, address, phone numbers, demographics, etc.)".

We discover that "in the digital world a person's identity is typically referred to as their digital identity [9]". It is argued in [10] that "identity encompasses all the essential characteristics that make each human unique". Satchell et al. indicated that "identifiers of a respective individual or entity would identify the entity online, from any context of the identity. An identifier uniquely identifies an entity (a person, a computer, an organisation, etc.) within a specific scope [11]". This underscores that digital identification is key in online activities of an entity on internet or computer network.

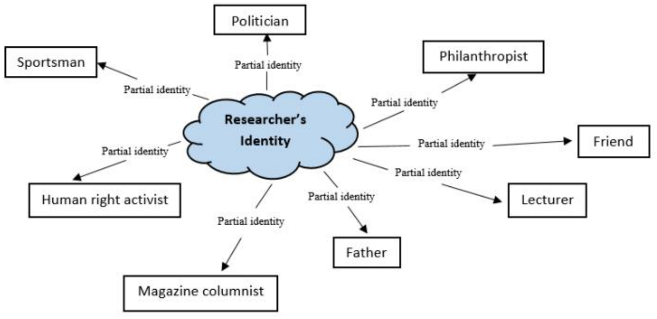


Fig. 1. Illustration of Partial Identity.

ISO/IEC 29003:2013 [12] gives a list of recognized tokens of identification as international standards on identity tokens. The token of identification is meant to fully distinguish the rightful owner of the token. The digital identity has attributes that help to establish an identity of an entity on online activities and services. In [12], we get a list of internationally accepted attributes that can identify an individual online. Identity attributes of a person or entity form a representation of an individual or entity through a given identity token. The relationships between an entity, identity token, digital identity, and identity attributes are presented in Fig. 2.

Due to the unrestrictive nature of the Internet, without proper identification and authentication, users are becoming more vulnerable to identity fraud and theft. Online identity theft, fraud, and privacy concerns have become a huge issue now; identity theft is big business [13]".

D. Communication Trust Model

Trust is an important ingredient during online communication between entities. However, such communication may not be immune to bad elements scavenging on the internet. It was observed that "[a hacker] could exploit a user's indoor location data to infer a variety of personal information, such as work role, smoker or not, coffee drinker or not, and even age [14]". It is imperative that such risks are eliminated by exploring solutions to such problems. Trust [which is] defined as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another [15]" is a basic constituent of social life [16],[17]. A trust framework model based on "the communication model by Shannon and Weaver [18] was adopted, incorporating the sending and receiving process of an individual according to the three tier approach of data, information, and knowledge. The Shannon and Weaver [18] model is one way directed and stems from the domain of information theory. In this model "a trustor places trust in the trustee [18]". Fig. 3 illustrates the Communication Trust Framework guiding the Shannon and Weaver model. "Communication consists of four major components the sender, the receiver, the message, and the environment. The communication process can generally be distinguished in the three phases sending, transmitting, and receiving. The phases of sending and receiving are concerned with the process of the message formation and comprehension by the sender and the receiver respectively [18]". The study by Memon and Arain shows that "communication requires adequate privacy level [19]" to improve security of information and online services; these could be assets that could be affected in online communication of entities. It was observed that "preserving privacy [in communication] is an important challenge [20]". This is necessary so as to ensure that only those that are entitled to private information or private online services have access to such. Sensitive information and services should be limited to only those that are privy to such assets. However, there has to be "a balance [between] service quality and privacy protection [20]". This would help to maximize the benefits of services and securing the interests of the legitimate digital identity owners to improve service quality amidst security of service.

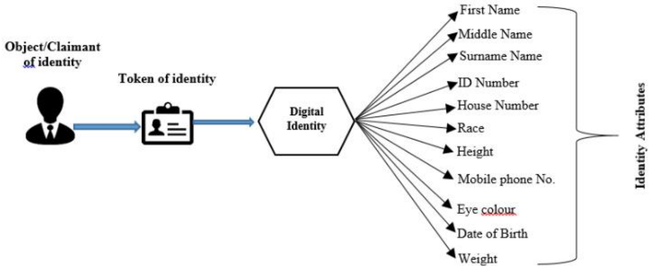


Fig. 2. Representation of Digital Identity by Identity Attributes.

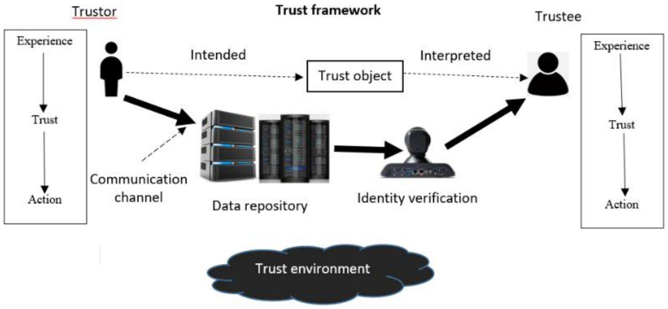


Fig. 3. Communication Trust Framework.

E. Mathematical Modeling

Haines and Crouch (2007) characterize "mathematical modeling as a cyclical process in which real-life problems are translated into mathematical language, solved within a symbolic system, and the solutions tested back within the real-life system [21]". This demonstrates how mathematical modelling can present a mathematical model that would help in solving a real life situation using mathematics. It is the interest of this research to establish a model that would help in presenting a solution to the problem of this research using a mathematical model. "Mathematical models comprise a range of representations, operations, and relations, rather than just one, to help make sense of real-life situations [22]".

F. Data Standardization

"We often want to compare scores or sets of scores obtained on different scales [23]". Standardizing data that comes from different sources would help us to "eliminate the unit of measurement by transforming the data into new scores with a mean of 0 and a standard deviation of 1. Considering that this research has interest to compare with the performance of other metrics, it is prudent that we have a common ground of comparing the performance of the metrics. We transform data "to improve our ability to discover knowledge [24]"; this transformation "includes normalising data [24]." Olson and Delen in [25] indicate that "the main advantage is to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation." It was noted that "normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements [26]". We discover that "a direct application of geometric measures (distances) to attributes with large ranges will implicitly assign bigger contributions to the metrics than the application to attributes with small ranges. The attributes should be dimensionless because the numerical values of the ranges of dimensional attributes depend on the units of measurements and, therefore, the choice of the units of measurements may greatly affect the results of clustering. One should not use distance measures without normalization of data [27]".

III. Related Works

Campbell et al. state that “in the simplest case, the components of [the sparse] vectors are the raw frequency counts of each term in each document [28]”. They also observed that “search engines of the World Wide Web (www) are based on certain information retrieval models like Boolean model, Probabilistic model, and Vector space model [28]”. Our interest is in the vector space model; Campbell et al. indicate that “the main purpose of [information retrieval models] is to retrieve relevant documents specific to a search [28]”. It was observed that “[vector space model] uses a storage matrix where columns represent the documents in a collection and whose rows represent the term frequencies among the documents [28]”. They also stated that “For ad-hoc querying, dynamic queries are compared against a static document database in order to find documents closest to the query [28]”. Simplistically speaking, a search engine has “static database of documents, a query processor, to convert incoming (dynamic) queries into a format compatible with the representation model, and a relevant measure to compare converted queries against documents [28]”. The researchers indicate that “when conducting a query, one method is to search through the storage matrix and match the query terms with row terms producing the document closest to the query [28]”.

Researchers have established that “Shannon’s entropy method is one of the various methods for finding weights [29]”. It has been observed that “multiple attribute decision making (MADM) refers to making preference decisions (e.g., evaluation, prioritization, and selection) over the available alternatives that are characterized by multiple, usually conflicting, attributes [29]”. It was observed that “since each criterion has a different meaning, it cannot be assumed that they all have equal weights, and as a result, finding the appropriate weight for each criterion [29]”. They discovered that “in MADM the greater the value of the entropy corresponding to a special attribute, which imply the smaller attribute’s weight, the less the discriminate power of that attribute in decision making process [29]”.

It is indicated in [29] that “the raw data are normalized to eliminate anomalies with different measurement units and scales. This process transforms different scales and units among various criteria into common measurable units to allow for comparisons of different criteria”. It was showed in [30] that “the entropic-weight method, from Shannon's entropy theory, was applied for the purpose of obtaining a classification”. Vajapeyam, summarizes “Shannon’s entropy [as] a direct measure of the number of bits needed to store the information in a variable, as opposed to its raw data [31]”. He adds that “entropy is a direct measure of the ‘amount of information’ in a variable [31]”.

Inambao et al. came up with a digital identity model that would “supply trusted digital identities [32]”; the model would “identify and extract various forms of identity attributes from various forms (identity tokens) [32]”. The model was established on Euclidean Distance metric based on Euclidean geometry. This model identified attributes that were very key as identifiers of an entity, in other words, these are attributes that can closely identify an entity. This model helps in “quantifying, implementing, and validating of the attributes from application forms (or identity tokens) [32]”.

Chinyemba and Phiri [33] showed “how to secure biometric data whilst at rest and or in motion so as to deter attackers in public organizations”. Biometric identification contributes immensely to a person’s identification and can therefore, contribute to the collection of digital identity attributes for individual identification. Ibou et al. indicated that “attribute-based digital identity modelling [needed] to take into account privacy issues [34]” and “proposed [a] model [that] takes into consideration three fundamental aspects, namely security, privacy and identity theft [34].

The work of Phiri et al. introduced a “multifactor authentication system based on two identity attributes metrics models [35]”. This broadens the scope of digital identification in an Identity Management system; we could have different modes of identification to make the digital identification robust and effective. Strengthening of the security of digital identity would include the developing of multi-modal authentication. This would include a combination of different authentication methods. For instance, like in the case of “when using an ATM bank card, in addition to the PIN number the user may be requested to submit a biometric feature such as a fingerprint in order to withdraw a certain amount of money above a given limit. A combination of biometrics, token based credentials and pseudo metrics will most likely form a very effective defense against imposters [35]”. The researchers where hoping that “an additional fourth category of inputs would take into account identity attributes such as the name, date of birth, address and other acquired identity attributes for consideration [35]”. Our research efforts are building on these past research work.

The work of Phiri et al. introduced a “multifactor authentication system based on two identity attributes metrics models [36]”. They argued that this would “reduce the cases of cybercrime since it becomes difficult to forge all the proposed four authentication factors that include biometrics [36]”. They went on to demonstrate “the performance of the three fuser block technologies namely Artificial Neural Networks (ANN), Fuzzy Inference System (FIS) and Adaptive Neuro-Fuzzy Inference System (ANFIS) using the term weight and entropy identity attributes metrics.

The current research was given birth by this work of Phiri et al. as indicated in the close of their work indicating that they considered the “future works [would] look at other combinations of the authentication factors and metrics modelling methodologies [36]”.

IV. Research Method

Consulting Creswell [37] indicates that this study is quantitative in nature and therefore, a survey to inquire into perceptions of observers was planned to use a questionnaire that would attend to these perceptions. As this study is quantitative in nature, extensive literature in quantitative studies was reviewed. Previous works that have applied the areas that have a bearing on this research with quantitative techniques applied were reviewed. Quantitative data was analyzed with the help of spread-sheets (e.g. Microsoft Excel). The techniques that have been used include data mining techniques and statistical analysis. PDF application forms for identity token requesting for identity attributes of individuals, within the research sample, were extracted from internet. The identity attributes were drawn from a list of internationally identified identity attributes by the International Standard Organization (ISO).

Documents in PDF format from the corpus of the Government of the Republic of Zambia (the researcher's residence) documents were searched and harvested from the internet. To test the proposed model, we got a set of application forms for identity token at random from our selected area. We picked ten (10) documents from the pdf documents out of 32 documents that were extracted from internet. Our model is focused on identifying the set of attributes that would identify a claimant of a digital identity that sufficiently matches the entity to be identified. Matching of a claimant could be done on one claimant or multiple claimants. In simple terms from our documents, if one document represents a token that owns the digital identity which is being claimed by the claimant, we can compare the attributes of digital identity of this entity to those of the claimant. For the purposes of this research, the ten documents will suffice, of which one would be the object and nine others will be the claimants of the digital identity. All the ten documents were tested on the metrics in the proposed mathematical model.

As indicated, ten (10) documents were picked from the Government of Zambia sets of documents. These documents are listed in Table I.

A. *Identity Attribute Text Mining*

Literature for International Standard Organization was consulted to identify attributes that are recognized as standard in the enrolment of diverse online services. Therefore, identified attributes by International world standards, ISO/IEC JTC 1/SC 27, were considered and used in this research. These standards have identified a list of attributes that could be collected from individuals during the time of enrolment for digital services of individuals; "Validation can occur during Identity Proofing, Identity Information Verification and Verification" [38] regarding entities from identity tokens. A list of attributes from ISO/IEC JTC 1/SC 27 indicates elements that would form identifiers to identify an individual, these are shown in Table II.

Tokens of identity are equally identified by this ISO standard. The identity documents and service enrolment application forms are documents that fell in the category of the international standards of ISO/IEC 29003:2013 [38]; These documents, according to the research samples, were searched from the internet and obtained in PDF format. TalkHelper PDF Converter version 2.2.9.0 tool was used to convert documents into PDF, for documents that were in other formats other than PDF. Documents which were already in PDF needed no format conversion. Fig. 5 shows TalkHelper PDF Converter that we used in this study.

Documents in PDF format were then converted into text files (using TalkHelper PDF Converter version 2.2.9.0) in readiness for text mining. AntConc 3.5.8, a corpus analysis toolkit was used for text mining. This tool was used to get the text frequency of the corpus files from different industries and regions, as discussed above, that were imported into the tool from respective folders. Fig. 4 shows the tool that was used for text mining.

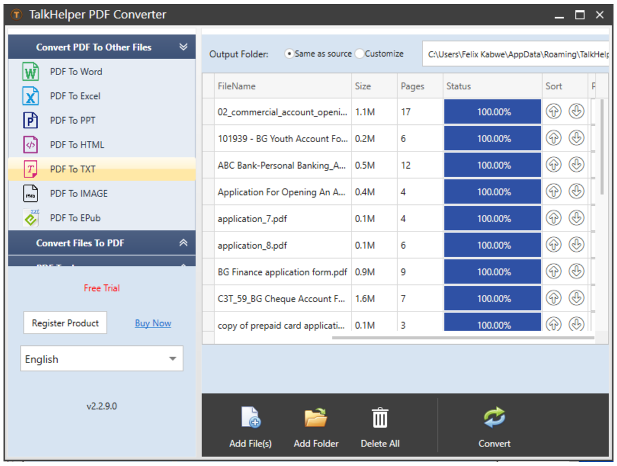
Each identity attribute had its term frequency recorded as indicated, from corpus analysis toolkit. Text mining was done on these documents using the same techniques as discussed above, based on the nineteen (19) existing attributes that we have been using. Table III shows the term frequencies (*Tf*) of each of the respective attributes after text mining.

TABLE I. SAMPLED DOCUMENTS FOR TERM WEIGHTING

Code	**Document name**	**Code**	**Document name**
D1	Airspace application form	D6	Residential Land acquisition application form
D2	Residence Permit application form	D7	Aquaculture Fund application form
D3	Visiting Visa application form	D8	Borehole Form application form
D4	Consent Form application form	D9	Health Professional Council membership application form
D5	Farm small holding application form	D10	Immovable Property application form

TABLE II. A LIST OF STANDARD ATTRIBUTES BASED ON ISO

Attributes (ISO/IEC JTC 1/SC 27)			
First name	Race	ID Number	Work telephone number
Middle name	Gender	Issuing authority	Work email address
Last name	Home address	Expiry date	Bank account details
Date of Birth	Home Unique Property Reference Number (House Number)	Home email address	Height
Place of Birth	Home telephone number	Work address	



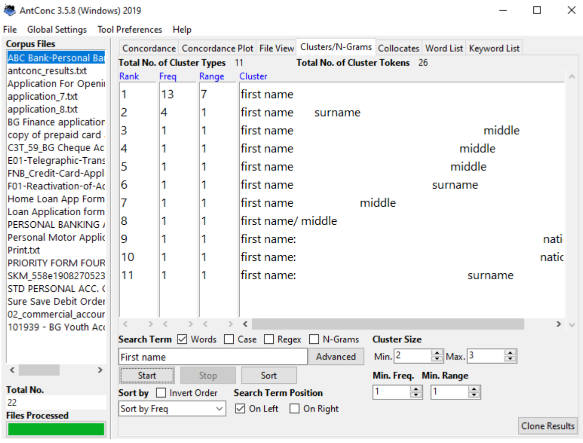
Output Folder: ● Same as source ○ Customize C:\Users\Felix Kabwe\AppData\Roaming\TalkHelp

FileName	Size	Pages	Status
02_commercial_account_openi...	1.1M	17	100.00%
101939 - BG Youth Account Fo...	0.2M	6	100.00%
ABC Bank-Personal Banking_A...	0.5M	12	100.00%
Application For Opening An A...	0.4M	4	100.00%
application_7.pdf	0.1M	4	100.00%
application_8.pdf	0.1M	6	100.00%
BG Finance application form.pdf	0.9M	9	100.00%
C3T_59_BG Cheque Account F...	1.6M	7	100.00%
copy of prepaid card applicati...	0.1M	3	100.00%

Fig. 4. TalkHelper PDF Converter Version 2.2.9.0.

TABLE III. TERM FREQUENCIES OF TEN DOCUMENTS FOR COMPUTING TF*IDF WEIGHTING (ZAMBIAN FIGURES)

ATTRIBUTE	Term (Tf$_i$)									
	D1: Tf_1	D2: Tf_2	D3: Tf_3	D4: Tf_4	D5: Tf_5	D6: Tf_6	D7: Tf_7	D8: Tf_8	D9: Tf_9	D10: Tf_{10}
First name	5	5	4	2	4	4	2	3	1	2
Middle name	5	5	4	2	4	4	2	3	1	2
Last name	5	5	4	2	4	4	2	3	1	2
Date of Birth	0	5	2	0	0	0	0	0	1	0
Place of Birth	0	3	3	0	0	0	0	0	0	0
Race	0	0	0	0	0	0	0	0	0	0
Gender	0	3	3	0	2	2	0	0	1	0
Home address	1	1	2	3	4	1	3	0	1	3
House Number	1	0	0	0	0	0	0	0	0	1
Home telephone number	1	0	0	1	1	1	1	0	0	0
ID Number	1	1	1	0	1	1	1	0	1	0
issuing authority	1	0	0	1	1	1	0	1	1	1
Expiry date	0	1	1	0	0	0	0	0	0	0
Home email address	0	0	0	1	1	1	1	0	1	0
Work address	1	1	0	2	4	1	0	0	1	3
Work telephone number	1	0	0	1	1	1	1	0	1	0
Work email address	0	0	0	1	1	1	1	0	1	0
Bank account details	0	0	0	0	0	0	0	0	1	0
Height	0	0	0	0	0	0	0	0	0	0
Sum	**22**	**30**	**24**	**16**	**28**	**22**	**14**	**10**	**13**	**14**



Corpus Files
ABC Bank-Personal Ba
antconc_results.txt
Application For Openi
application_7.txt
application_8.txt
BG Finance applicatior
copy of prepaid card
C3T_59_BG Cheque Ac
E01-Telegraphic-Trans
FNB_Credit-Card-Appl
F01-Reactivation-of-Ac
Home Loan App Form
Loan Application form
PERSONAL BANKING
Personal Motor Applic
Print.txt
PRIORITY FORM FOUR
SKM_558e1908270523
STD PERSONAL ACC. (
Sure Save Debit Order
02_commercial_accour
101939 - BG Youth Acc

Total No. 22
Files Processed

```
Total No. of Cluster Types   11        Total No. of Cluster Tokens   26
Rank   Freq   Range   Cluster
1      13     7       first name
2      4      1       first name    surname
3      1      1       first name                          middle
4      1      1       first name                      middle
5      1      1       first name                     middle
6      1      1       first name                  surname
7      1      1       first name        middle
8      1      1       first name/ middle
9      1      1       first name:                                  nati
10     1      1       first name:                                  natic
11     1      1       first name:                        surname
```

Search Term ☑ Words ☐ Case ☐ Regex ☐ N-Grams

Search term: First name

Cluster Size Min. 2 Max. 3

Min. Freq. 1 Min. Range 1

Sort by ☐ Invert Order: Sort by Freq

Search Term Position ☑ On Left ☐ On Right

Fig. 5. AntConc 3.5.8, a Corpus Analysis Toolkit for Data Mining.

"As the time passes, a lot of information and new challenges related to information acquisition and data mining are emerging very rapidly [39]". Efforts of curbing online risks ought to match the rapid growth of technology and online services.

V. PROPOSED MODEL

The proposed model Identity Attribute Metric Model based on the Distance Metrics in this research is the Cosine Similarity measure.

A. Model Quantification

A cluster is a collection of data objects that are similar to objects within the same cluster and dissimilar to those in other clusters. Similarity between two objects is calculated using a distance measure [40]. Charulatha et.al indicate that "Clustering is the grouping of similar instances/objects some sort of measure that can determine whether two objects are similar [26]". As pointed out by Backer and Jain, "in cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (i.e., chosen subjectively based on its ability to create 'interesting' clusters) [34]". "From the scientific and mathematical point of view distance is defined as a quantitative degree of how far apart two objects are [41]."

Researchers note that "it is natural to ask what kind of standards we should use to determine the closeness, or how to measure the distance (dissimilarity) or similarity between a pair of objects, an object and a cluster, or a pair of clusters [34]". "In order for the distance metrics to make sense, good data transformation or normalization is required. In data normalization methods, the objective is usually to ensure that the computed distance metric or similarity measure will reflect the inherent distance or similarity of the data [42]".

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications and clustering too [42]. An important property of the cosine similarity is its independence of document length. For example, combining two identical copies of a document d1 to get a new pseudo document d2, the cosine similarity between d1 and d2 is 1, which means that these two documents are regarded to be identical. Given another document d3, d1 and d2 will have the same similarity value to d3 [42] as shown in equation (1).

$$\text{Sim}\left(Tf_{d_1}, Tf_{d_3}\right) = \text{Sim}\left(Tf_{d_2}, Tf_{d_3}\right) \qquad (1)$$

Documents with the same composition but different totals will be treated identically. When the term vectors are normalized to a unit length such as 1, and in this case the representation of d1 and d2 is the same [42].

Cosine similarity measure has a high positive correlation than the Euclidean Distance [43]. The cosine of 0^0 is 1 and it is <1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90^0 have a similarity of 0 and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space where the outcome is nearly bounded in [0,1]. Cosine similarity is particularly used in positive space where the outcome is nearly bounded in [0,1] [43]. Cosine similarity gives a useful measure of how similar two documents are likely to be in terms of their subject matter [43]. This distance metric will give us a number from the closed interval [0, 1], 0 denoting that the two vectors are overlapping and 1 denoting that there is an angle of 90 ° which is the highest difference between the vectors [44].

Cosine Similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of an angle between them [44]. It can be derived using the Euclidean dot product. Given two non-zero vectors, "x" and "y", the dot product of the two vectors would be represented by

$$\boldsymbol{x.y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\|\, Cos\,\theta \qquad (2)$$

This will translate to

$$Cos\,\theta = \frac{\boldsymbol{x.y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \qquad (3)$$

This also agrees with trigonometry and complex numbers; given two vectors, x and y in a vector space, the Cosine of the angle (θ) between these two vectors would be represented by the equation above.

Given two vectors X and Y, the Cosine Similarity, Cos (θ) is expressed as a dot product and magnitude as

$$\text{Similarity} = S(X,Y) = Cos\,(\theta) = Cos\,(X,Y)$$

$$= \frac{X.Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \qquad (4)$$

These two vectors could be that "X" is a set of attributes that of an applicant who claims ownership of the identity attributes while "Y" could be verifier identity attributes. The Cosine function in equation (3) can be represented as a Similarity distance measure in equation (4) as is also indicated in [45].

B. Identity Verification

The choice of this model was based on two considerations that could be applied in this study:

1) For verification of ownership

a) When we are specifically interested in attending to one applicant for verification of ownership claim of a particular digital identity with known identity attributes.

b) When multiple applicants make claims of ownership claims of a particular digital identity with known identity attributes and we need to verify.

2) The principle of orientation of two similar vectors in a metric space that is inherent with the cosine Similarity distance.

Cosine Similarity measure is used in data mining as a technique for documents that are similar based on the text that these documents contain. For instance, this metric is used in considering those who share same tags on a blog, persons who viewed same documents, customers who bought similar items online.

Verifying online identity for claimants could help establish who the legitimate owner would be from a multiple of identity claimants. We could use the metrics and mathematical computations to achieve this. Therefore, this model can be used in the verification process of an applicant or applicants in the Digital Identity Management System.

C. Testing the Model

For us to identify the hierarchy of importance of attributes in the corpus, we need to consider the term weight of each attribute within the corpus of the ten (10) documents. We have represented the ten (10) documents in our functions as d_1, d_2, d_3,…d_{10}. The general expression of d_i, represents the same ten documents ranging from d1 to d_{10}.

Chen and Chang indicate that "TF and TF-IDF are widely applied to count the weight of a term" [43]. They further add that "TF represents the number of times a term occurs in a document, and TF-IDF is the combining of TF and IDF weights. IDF indicates the general importance of a term in overall documents" [43].

Researchers indicate that Term frequency (*Tf*) factor is represented by the "logarithm of the term frequency to scale the effect of unfavorably high term frequency [44]". This is expressed as.

$$\text{TF} = 1 + \log tf \qquad (5)$$

D. Indeterminate Considerations

It is important to recognize that the function $\mathbf{log}\,\boldsymbol{tf}$ runs into indeterminate when $\boldsymbol{tf}$ becomes zero (0) since

$\log 0 = \infty$ and

$1 + \infty$ are indeterminate

We therefore, evaluate this part of the function; we have a logarithmic property that for any n = 1, 2, 3, … we have

$$\frac{x-1}{x} \leq f_n(x) \leq x-1 \qquad (6)$$

Therefore,

$$\frac{x-1}{x} \leq \log x \leq x-1 \qquad (7)$$

It follows that the upper bound of log x is x – 1

Therefore, replacing ***tf*** in the function *TF*= 1 + $\log tf$ we have

TF= 1 + (x-1) (8)

For x = 0, we have

TF= 1 + (0-1) = 0 (9)

The Inverse Document Frequency component (IDF) of the function is expressed when we "multiply original *tf* factor by an inverse collection frequency factor (N is the total number of documents in a collection, and n_i is the number of documents to which a term is assigned) [43]".

It was indicated in [46] IDF can be calculated by

$$\text{idf} = \frac{The\ number\ of\ total\ docments}{The\ number\ of\ documents\ include\ term} \qquad (10)$$

This is represented by the expression:

$$\text{IDF} = \log\frac{N}{n_i} \qquad (11)$$

This function will be indeterminate when n_i = 0. We observe that

$$\log\frac{N}{n_i} = \log N - \log n_i \qquad (12)$$

In our corpus, N = 10. We could have situations when n_i = 0; at that point then our function would become indeterminate.

That is,

IDF = log 10 – log 0 = log 10 (13)

From our established statement above, in (7), it therefore follows that

IDF = log 10 – ($x-1$) (14)

When x = 0, then we have

IDF = log 10 – ($0-1$) = log 10 +1 (15)

Table IV represents the term frequencies (*TF*) of the corpus. The functions in Table IV, Table V, and Table VI for the term frequencies Tf_i and idf_i, have their indeterminate logarithmic functions resolved and therefore, present the outcomes of the functions.

In 1993 Buckley stated that "over the past 25 years, one class of term weights has proven itself to be useful over a wide variety of collections. This is the class of *tf***idf* (term frequency times inverse document frequency) weights [47]". "TF-IDF is also one of the most popular term-weighting schemes for user modeling and recommender systems [48]".

Considering the TD-IDF term weight scheme, from our findings above, we would have the weighting computational outcomes to be as indicated in Table IV. The metric would be represented by:

$$W_i = TF*IDF = Tf_i*idf_i = (1+\log Tf_{i,d})*\log\frac{N}{df_i} \qquad (16)$$

The terms of the functions have been explained above. We obtain the weighting of the attributes (terms) by considering the function (16) above of which the outcomes are indicated in Table IV.

E. Term Importance

Jiao et al. established that "a classic way to assess the importance of a term is the so-called *tf-idf* (term frequency - inverse document frequency) term weighting scheme [49]". They further indicated that the term importance "is based on two assumption:

a) idf assumption: rare terms are more informative than frequent terms,

b) tf assumption: multiple occurrences of a term in a query document are more relevant than single occurrence [49].

After sorting the outcomes of the computations of the weighting in Tf*idf we are able to arrange in order of which attribute is more important than the others.

F. Euclidean Distance based Similarity

Past efforts [32] have showed that Euclidean Distance Geometry could "improve the authentication in digital identity management system and particularly improve the security in digital financial services".

The Euclidean distance between two points or terms (t_1 and t_2), from a corpus, in a two dimensional space is represented by the function.

$$d_{t_1,t_2} = \sqrt{\sum_{i=1}^{n}(t_{1i}-t_{2i})^2} \qquad (17)$$

VI. Results

A. Term Frequencies (TF*IDF) in Proposed Metrics

Table V shows the index document frequencies (IDF) in our term weighting function.

B. Weighted Identity Attributes

Table VII shows the rating of the terms on which weighting has been applied. This rating indicates which identity attributes are most important in identifying a digital identity claimant against online interests in this corpus. We are interested to see which identity attributes are key in identifying an identity claimant.

C. Model Based on Cosine Similarity Measure

In order to demonstrate the effectiveness of our proposed model, we would need to apply our model on the dataset which considered the weighting of the attributes. The results of these metrics have been recorded in Tables VIII, IX, and X.

TABLE IV. TERM FREQUENCIES ON TEN DOCUMENTS FOR THE METRICS (ZAMBIAN FIGURES)

ATTRIBUTE	$Tf_i = 1 + \log Tf_{i,d}$									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
First name	1.6990	1.6990	1.6021	1.3010	1.6021	1.6021	1.3010	1.4771	1.0000	1.3010
Middle name	1.6990	1.6990	1.6021	1.3010	1.6021	1.6021	1.3010	1.4771	1.0000	1.3010
Last name	1.6990	1.6990	1.6021	1.3010	1.6021	1.6021	1.3010	1.4771	1.0000	1.3010
Date of Birth	0.0000	1.6990	1.3010	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
Place of Birth	0.0000	1.4771	1.4771	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Race	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gender	0.0000	1.4771	1.4771	0.0000	1.3010	1.3010	0.0000	0.0000	1.0000	0.0000
Home address	1.0000	1.0000	1.3010	1.4771	1.6021	1.0000	1.4771	0.0000	1.0000	1.4771
House Number	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Home telephone number	1.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
ID Number	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
issuing authority	1.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000
Expiry date	0.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Home email address	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
Work address	1.0000	1.0000	0.0000	1.3010	1.6021	1.0000	0.0000	0.0000	1.0000	1.4771
Work telephone number	1.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
Work email address	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000
Bank account details	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
Height	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLE V. INVERSE FUNCTION FOR THE TF*IDF WEIGHTING (ZAMBIAN FIGURES)

ATTRIBUTE	Total No. of docs	$idf_i = \log\frac{N}{df_i}$ = log N - log df_i									
	N	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
First name	10	0.30103	0.30103	0.39794	0.69897	0.39794	0.39794	0.69897	0.52288	1.00000	0.69897
Middle name	10	0.30103	0.30103	0.39794	0.69897	0.39794	0.39794	0.69897	0.52288	1.00000	0.69897
Last name	10	0.30103	0.30103	0.39794	0.69897	0.39794	0.39794	0.69897	0.52288	1.00000	0.69897
Date of Birth	10	0.00000	0.30103	0.69897	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
Place of Birth	10	0.00000	0.52288	0.52288	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Race	10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Gender	10	0.00000	0.52288	0.52288	0.00000	0.69897	0.00000	0.00000	0.00000	1.00000	0.00000
Home address	10	1.00000	1.00000	0.69897	0.52288	0.39794	1.00000	0.52288	0.00000	1.00000	0.52288
House Number	10	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000
Home telephone number	10	1.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	0.00000	0.00000
ID Number	10	1.00000	1.00000	1.00000	0.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
issuing authority	10	1.00000	0.00000	0.00000	1.00000	1.00000	1.00000	0.00000	1.00000	1.00000	1.00000
Expiry date	10	0.00000	1.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Home email address	10	0.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
Work address	10	1.00000	1.00000	0.00000	0.00000	0.39794	1.00000	0.00000	0.00000	1.00000	0.52288
Work telephone number	10	1.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
Work email address	10	0.00000	0.00000	0.00000	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.00000
Bank account details	10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
Height	10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

TABLE VI. TF*IDF WEIGHTING OF THE IDENTITY ATTRIBUTES ON TEN DOCUMENTS (ZAMBIAN FIGURES)

ATTRIBUTE	W_{i,d}$=TF_i * IDFi = Tf_i \, X \, \log\frac{N}{idf_i}$									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
First name	0.511441	0.511441	0.637524	0.909381	0.637524	0.637524	0.909381	0.772355	1.000000	0.909381
Middle name	0.511441	0.511441	0.637524	0.909381	0.637524	0.637524	0.909381	0.772355	1.000000	0.909381
Last name	0.511441	0.511441	0.637524	0.909381	0.637524	0.637524	0.909381	0.772355	1.000000	0.909381
Date of Birth	0.000000	0.511441	0.909381	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
Place of Birth	0.000000	0.772355	0.772355	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Race	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Gender	0.000000	0.772355	0.772355	0.000000	0.909381	0.000000	0.000000	0.000000	1.000000	0.000000
Home address	1.000000	1.000000	0.909381	0.772355	0.637524	1.000000	0.772355	0.000000	1.000000	0.772355
House Number	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
Home telephone number	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000
ID Number	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
issuing authority	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000
Expiry date	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Home email address	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
Work address	1.000000	1.000000	0.000000	0.000000	0.637524	1.000000	0.000000	0.000000	1.000000	0.772355
Work telephone number	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
Work email address	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000
Bank account details	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
Height	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

TABLE VII. LISTING OF IMPORTANCE OF THE IDENTITY ATTRIBUTES (ZAMBIAN FIGURES)

ATTRIBUTES	Term (Tf_i)										T F_r*IDF_r
	D1: Tf_1	D2: Tf_2	D3: Tf_3	D4: Tf_4	D5: Tf_5	D6: Tf_6	D7: Tf_7	D8: Tf_8	D9: Tf_9	D10: Tf_{10}	Total	$\sum_{i=1}^{n} Tf_i \, X \, \log\frac{N}{idf_i}$
Home address	1	1	2	3	4	1	3	0	1	3	19	7.86397063
First name	5	5	4	2	4	4	2	3	1	2	32	7.43595130
Middle name	5	5	4	2	4	4	2	3	1	2	32	7.43595130
Last name	5	5	4	2	4	4	2	3	1	2	32	7.43595130
ID Number	1	1	1	0	1	1	1	0	1	0	7	7.00000000
issuing authority	1	0	0	1	1	1	0	1	1	1	7	7.00000000
Work telephone number	1	0	0	1	1	1	1	0	1	0	6	6.00000000
Work address	1	1	0	2	4	1	0	0	1	3	13	5.40987908
Home telephone number	1	0	0	1	1	1	1	0	0	0	5	5.00000000
Home email address	0	0	0	1	1	1	1	0	1	0	5	5.00000000
Work email address	0	0	0	1	1	1	1	0	1	0	5	5.00000000
Gender	0	3	3	0	2	2	0	0	1	0	11	3.45409156
Date of Birth	0	5	2	0	0	0	0	0	1	0	8	2.42082187
House Number	1	0	0	0	0	0	0	0	0	1	2	2.00000000
Expiry date	0	1	1	0	0	0	0	0	0	0	2	2.00000000
Place of Birth	0	3	3	0	0	0	0	0	0	0	6	1.54471062
Bank account details	0	0	0	0	0	0	0	0	1	0	1	1.00000000
Race	0	0	0	0	0	0	0	0	0	0	0	0.00000000
Height	0	0	0	0	0	0	0	0	0	0	0	0.00000000
Sum	**22**	**30**	**24**	**16**	**28**	**22**	**14**	**10**	**13**	**14**	**193**	

D. Verification of Ownership

For the purposes of verification of ownership of the attributes by an online user, we will assume that the object of ownership is the user of document 2 from our corpus of ten documents. Document 2 was purposed to capture attributes of a people who would apply for residence permit. It is only an individual who has entered responses that match the attributes of the specific individual that would be said to be said to be uniquely similar. For the sake of assessment of key attributes, we would consider the attributes involved in identifying the digital identity of our object and compare with the other attributes from the other nine (9) documents. We are going to look at the attributes of the second document and compare them to each of the documents of the nine other documents, respectively. Using our proposed model of the Cosine Similarity measure we would then observe the performance on similarity of the attributes of the second document to those of the other nine.

E. Verification Based on Term Frequencies

We have the following vectors from the Term Frequencies of the attributes of the 10 documents of the corpus:

i. Airspace (D1): Tf1 = D2

= (5, 5, 5, 5, 3, 0, 3, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0)

ii. Residence Permit (D2): Tf2 = D1

= (5, 5, 5, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0)

iii. Visiting Visa (D3): Tf3 = D3

= (4, 4, 4, 2, 3, 0, 3, 2, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0)

iv. Consent Form (D4): Tf4 = D4

= (2, 2, 2, 0, 0, 0, 0, 3, 0, 1, 0, 1, 0, 1, 2, 1, 1, 0, 0)

v. Farm Smallholding (D5): Tf5 = D5

= (4, 4, 4, 0, 0, 0, 2, 4, 0, 1, 1, 1, 0, 1, 4, 1, 1, 0, 0)

vi. Residential Land (D6): Tf6

= (4, 4, 4, 0, 0, 0, 2, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0)

vii. Aquaculture Fund (D7): Tf7

= (2, 2, 2, 0, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0)

viii. Borehole Form (D8): Tf8

= (3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)

ix. Health Prof Council (D9): Tf9

= (1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0)

x. Immovable Property (D10): Tf10

= (2, 2, 2, 0, 0, 0, 0, 3, 1, 0, 0, 1, 0, 0, 3, 0, 0, 0, 0)

Replacing the variables of the documents in our model, we let the documents to be identified by d1, d2, …, d10. We then apply our model:

Similarity = $S(X,Y) = Cos\ (\theta)\ = Cos\ (X,Y)$

$$= \frac{X * Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$$

1. For $S(D_2, D_1)$:

d2.d1 = (5, 5, 5, 5, 3, 0, 3, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0)*

(5, 5, 5, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0) = ((5x5) +(5x5)+(5x5)+(5x0)+(3x0)+(0x0)+(3x0)+(1x1)+(0x1)+ (0x1)+(1x1)+(0x1)+(1x0)+(0x0)+(1x1)+(0x1)+(0x0)+(0x0)+(0 x0)) = 78

This follows that

$D_2{*}D_i$	$d_{2*}d_1$	$d_{2*}d_2$	$d_{2*}d_3$	$d_{2*}d_4$	$d_{2*}d_5$	$d_{2*}d_6$	$d_{2*}d_7$	$d_{2*}d_8$	d_{2*9}	$d_{2*}d_{10}$
Outcome	78	122	92	35	75	69	34	45	26	36

$\|d_2\| = \sqrt{5^2 + 5^2 + 5^2 + 5^2 + 3^2 + 0^2 + 3^2 + 1^2 + 0^2 + 0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2}$ = 122

$\|d_1\| = \sqrt{5^2 + 5^2 + 5^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2}$ = 82

Therefore,

Similarity = $S(d_2, d_1)\ =\ Cos(d_2, d_1) = \frac{d_2 * d_1}{\|d_2\|\|d_1\|} = \frac{78}{122\ x\ 82}$ = <u>0.007797</u>

It follows that for the rest of the computations we have

$\|d_i\|$	$\|d_1\|$	$\|d_2\|$	$\|d_3\|$	$\|d_4\|$	$\|d_5\|$	$\|d_6\|$	$\|d_7\|$	$\|d_8\|$	$\|d_9\|$	$\|d_{10}\|$
Outcome	82	122	76	30	90	60	26	28	13	32

2. For $S(D_2, D_2)$:

$S(d_2, d_2)\ =\ Cos(d_2, d_2) = \frac{d_2 * d_2}{\|d_2\|\|d_2\|} = \frac{122}{122\ x\ 122}$ = 0.008197

3. For $S(D_2, D_3)$:

$S(d_2, d_3)\ =\ Cos(d_2, d_3) = \frac{d_2 * d_3}{\|d_2\|\|d_3\|} = \frac{92}{122\ x\ 76}$ = 0.009922

4. For $S(D_2, D_4)$:

$S(d_2, d_4)\ =\ Cos(d_2, d_4) = \frac{d_2 * d_4}{\|d_2\|\|d_4\|} = \frac{92}{122\ x\ 76}$ = 0.009563

5. For $S(D_2, D_5)$:

$S(d_2, d_5)\ =\ Cos(d_2, d_5) = \frac{d_2 * d_5}{\|d_2\|\|d_5\|} = \frac{75}{122\ x\ 90}$ = 0.006831

6. For $S(D_2, D_6)$:

$S(d_2, d_6)\ =\ Cos(d_2, d_6) = \frac{d_2 * d_6}{\|d_2\|\|d_6\|} = \frac{69}{122\ x\ 60}$ = 0.009426

7. For $S(D_2, D_7)$:

$$S(d_2, d_7) = Cos(d_2, d_7) = \frac{d_2 * d_7}{\|d_2\|\|d_7\|} = \frac{34}{122\ x\ 26} = 0.010719$$

8. For $S(D_2, D_8)$:

$$S(d_2, d_8) = Cos(d_2, d_8) = \frac{d_2 * d_8}{\|d_2\|\|d_8\|} = \frac{45}{122\ x\ 28} = 0.013173$$

9. For $S(D_2, D_9)$:

$$S(d_2, d_9) = Cos(d_2, d_9) = \frac{d_2 * d_9}{\|d_2\|\|d_9\|} = \frac{26}{122\ x\ 13} = 0.016393$$

10. For $S(D_2, D_{10})$:

$$S(d_2, d_{10}) = Cos(d_2, d_{10}) = \frac{d_2 * d_{10}}{\|d_2\|\|d_{10}\|} = \frac{36}{122\ x\ 32} = 0.009221$$

Sorting the Cosine measure of the outcome that was calculated based on the Term frequencies of the documents di, d2, d3,…,d10 will give us the following:

It was observed that using term frequencies in our computations yields a result where the metrics using Cosine similarity measure gives a notable result. Comparing a document to itself in the computations yields a result third on the table as shown in Table VIII; this implies that the document is far from being identical to itself. This is clear indication that using term frequencies includes errors from the documents, which would include noise and other errors. Using standardized data helps in improving accuracy of results.

We therefore standardize the data using term weights and repeat the computations as above and record the results. The results are reflected in Table IX.

TABLE VIII. RESULTS ON UN-WEIGHTED DATA ON THE COSINE MEASURE

Rating	Function	Item	How close is the document to the object (D_2)?
1	S(d2,d5)	Document 2 compared to Document 5	0.006830601
2	S(d2,d1)	Document 2 compared to Document 1	0.007796881
3	S(d2,d2)	Document 2 *compared to itself*	0.008196721
4	S(d2,d10)	Document 2 compared to Document 10	0.009221311
5	S(d2,d6)	Document 2 compared to Document 6	0.00942623
6	S(d2,d4)	Document 2 compared to Document 4	0.009562842
7	S(d2,d3)	Document 2 compared to Document 3	0.009922347
8	S(d2,d7)	Document 2 compared to Document 7	0.010718789
9	S(d2,d8)	Document 2 compared to Document 8	0.013173302
10	S(d2,d9)	Document 2 compared to Document 9	0.016393443

F. Verification based on Term Weights

We have the following weights of the ten documents:

D_2*D_i	d_2*d_1	d_2*d_2	d_2*d_3	d_2*d_4	d_2*d_5	d_2*d_6	d_2*d_7	d_2*d_8	d_2*d_9	d_2*d_{10}
Outcome	3.784715	6.239353	5.545708	2.167639	3.955580	3.978167	3.167639	1.185042	5.818119	2.939995

We also have

$\|d_i\|$	$\|d_1\|$	$\|d_2\|$	$\|d_3\|$	$\|d_4\|$	$\|d_5\|$	$\|d_6\|$	$\|d_7\|$	$\|d_8\|$	$\|d_9\|$	$\|d_{10}\|$
Outcome	7.784715	6.239353	6.066322	8.077454	8.859156	9.219310	8.077454	2.789598	13.000000	5.673987

We therefore, have the following Cosine similarity measures from the data we have above:

1. For $S(D_2, D_1)$: $S(d_2, d_1) = Cos(d_2, d_1) = \frac{d_2 * d_1}{\|d_2\|\|d_1\|} = \frac{3.784715}{6.239353\ x\ 7.784715} = 0.077920$

2. For $S(D_2, D_2)$: $S(d_2, d_2) = Cos(d_2, d_2) = \frac{d_2 * d_2}{\|d_2\|\|d_2\|} = \frac{6.239353}{6.239353x\ 6.239353} = 0.160273$

3. For $S(D_2, D_3)$: $S(d_2, d_3) = Cos(d_2, d_3) = \frac{d_2 * d_3}{\|d_2\|\|d_3\|} = \frac{5.545708}{6.239353\ x\ 6.066322} = 0.146518$

4. For $S(D_2, D_4)$: $S(d_2, d_4) = Cos(d_2, d_4) = \frac{d_2 * d_4}{\|d_2\|\|d_4\|} = \frac{2.167639}{6.239353\ x\ 8.077454} = 0.146518$

5. For $S(D_2, D_5)$: $S(d_2, d_5) = Cos(d_2, d_5) = \frac{d_2 * d_5}{\|d_2\|\|d_5\|} = \frac{3.955580}{6.239353\ x\ 8.859156} = 0.071561$

6. For $S(D_2, D_6)$: $S(d_2, d_6) = Cos(d_2, d_6) = \frac{d_2 * d_6}{\|d_2\|\|d_6\|} = \frac{3.978167}{6.239353\ x\ 9.219310} = 0.069158$

7. For $S(D_2, D_7)$: $S(d_2, d_7) = Cos(d_2, d_7) = \frac{d_2 * d_7}{\|d_2\|\|d_7\|} = \frac{3.167639}{6.239353\ x\ 8.077454} = 0.062852$

8. For $S(D_2, D_8)$: $S(d_2, d_8) = Cos(d_2, d_8) = \frac{d_2 * d_8}{\|d_2\|\|d_8\|} = \frac{1.185042}{6.239353\ x\ 2.789598} = 0.068085$

9. For $S(D_2, D_9)$: $S(d_2, d_9) = Cos(d_2, d_9) = \frac{d_2 * d_9}{\|d_2\|\|d_9\|} = \frac{5.818119}{6.239353\ x\ 13.000000} = 0.071730$

10. For $S(D_2, D_{10})$: $S(d_2, d_{10}) = Cos(d_2, d_{10}) = \frac{d_2 * d_{10}}{\|d_2\|\|d_{10}\|} = \frac{2.939995}{6.239353\ x\ 5.673987} = 0.083046$

Our main interest is to identify the text from the documents that would be the best identifier of the online user. The details of the digital object of an applicant of identity and verification, which in our case is represented by the identifying attributes, would need to accurately match attributes of verification. We therefore, consider the importance of attributes that is in the corpus of ten documents. Table IX shows the documents that are sorted in the order of importance; in this case, the documents would represent the applicants that are being subjected for verification by the process of authentication.

From Table IX, we see that it was important to normalize the Term frequencies from the documents so as to remove the errors from data. Without normalizing the data, we have the rating of the document affected to a point that the document compared to itself shows deficit in the content of terms. Removing the errors through normalization done by term weighting of the data from the corpus of the ten documents gives the rating where document 2 is compared to itself becomes first in rating. This is the natural expectation of the outcome of this process.

We have just established that when an online application or applications from multiple users for authentication, Cosine Similarity measure could help us to accurately identify who the true owner of the digital identity would be. This indicates that Cosine Similarity measure could be a very strong tool in information security to add another level in authentication. Coupled with other techniques, we could build a robust system in information security for Digital Identity management.

G. *Results on Metrics Model*

Table X shows the top ten identity attributes from the ten documents where TF*IDF term weighting was applied. Picking identity attributes that have been found to be higher in terms of weighting would help us identify the owner of the identity attributes for online identity claimant. Applying developed Identity Attribute Metrics, which was developed using the Cosine Similarity measure we obtain the following results:

TABLE IX. RESULTS ON USING WEIGHTED DATA ON THE PROPOSED MODEL

Rating	Function	Documents compared	How close is the document to the object (D_2) ?
1	d_2*d_2	Document 2 *compared to itself*	0.160273
2	d_2*d_3	Document 2 compared to Document 3	0.146518
3	d_2*d_{10}	Document 2 compared to Document 10	0.083046
4	d_2*d_1	Document 2 compared to Document 1	0.077920
5	d_2*d_9	Document 2 compared to Document 9	0.071730
6	d_2*d_5	Document 2 compared to Document 5	0.071561
7	d_2*d_6	Document 2 compared to Document 6	0.069158
8	d_2*d_8	Document 2 compared to Document 8	0.068085
9	d_2*d_7	Document 2 compared to Document 7	0.062852
10	d_2*d_4	Document 2 compared to Document 4	0.043010

TABLE X. LIST OF TOP TEN IDENTITY ATTRIBUTE FROM THE PROPOSED MODEL (ZAMBIAN FIGURES)

ATTRIBUTE	Term (Tf_i)											
	D1: Tf_1	D2: Tf_2	D3: Tf_3	D4: Tf_4	D5: Tf_5	D6: Tf_6	D7: Tf_7	D8: Tf_8	D9: Tf_9	D10: Tf_{10}	Total	T F_r*IDF_r $\sum_{i=1}^{n} Tf_i X \log \frac{N}{idf_i}$
Home address	**1**	**1**	**2**	**3**	**4**	**1**	**3**	**0**	**1**	**3**	**19**	**7.86397063**
First name	**5**	**5**	**4**	**2**	**4**	**4**	**2**	**3**	**1**	**2**	**32**	**7.43595130**
Middle name	**5**	**5**	**4**	**2**	**4**	**4**	**2**	**3**	**1**	**2**	**32**	**7.43595130**
Last name	**5**	**5**	**4**	**2**	**4**	**4**	**2**	**3**	**1**	**2**	**32**	**7.43595130**
ID Number	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**7**	**7.00000000**
issuing authority	**1**	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**7**	**7.00000000**
Work telephone number	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**6**	**6.00000000**
Work address	**1**	**1**	**0**	**2**	**4**	**1**	**0**	**0**	**1**	**3**	**13**	**5.40987908**
Home telephone number	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**5**	**5.00000000**
Home email address	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**5**	**5.00000000**
Work email address	0	0	0	1	1	1	1	0	1	0	5	5.00000000
Gender	0	3	3	0	2	2	0	0	1	0	11	3.45409156
Date of Birth	0	5	2	0	0	0	0	0	1	0	8	2.42082187
House Number	1	0	0	0	0	0	0	0	0	1	2	2.00000000
Expiry date	0	1	1	0	0	0	0	0	0	0	2	2.00000000
Place of Birth	0	3	3	0	0	0	0	0	0	0	6	1.54471062
Bank account details	0	0	0	0	0	0	0	0	1	0	1	1.00000000
Race	0	0	0	0	0	0	0	0	0	0	0	0.00000000
Height	0	0	0	0	0	0	0	0	0	0	0	0.00000000
Sum	**22**	**30**	**24**	**16**	**28**	**22**	**14**	**10**	**13**	**14**	**193**	

VII. Discussion

Testing the proposed Cosine Similarity measure as an Identity Attribute Metric Modeling is able to identify the document that uniquely has its identity attributes similar to itself as the highest rated and hence identify a claimant of the digital identity as the legitimate owner. This model would be able to identify the true owner claimant from one to multiple claimants of the digital identity. This would help in improving security on identifying the legitimate digital identity owner of a specific identity. Only such an owner should the given access to online assets, services, or attention.

It was observed that the identity attributes from the ISO list were based on physical identification of an individual claimant. The study showed that using Cosine Similarity measure, the legitimate owner of the digital identity would be uniquely identified with the top most score in the computations. To achieve this, mined text of digital identity would need to be normalized, in this case we used the term weight to normalize the data. The calculations with the model give best results on term weighted text. It was also observed that there was a set of digital attributes would score higher than others when we apply our model. After sorting the results of the model on the weighted text mined identity attributes, it was observed that the identity attributes that locate residence of a claimant was of paramount importance. It was equally observed that the identifying names of the claimant, national identification, economic activity, and contacts of the claimant were ranked high in the results of our computations.

VIII. Conclusion

The proposed model was able to identify the legitimate owner of the digital identity attributes and therefore, able to show who the false-identity online claimants were. The model was also able to identify the attributes that were key in identifying the legitimate owner of the claimed identity, in other words, the most important attributes to distinguish the legitimate owner from the false ones could be identified using this model. The identity attributes can be extracted from identity tokens by mining identity attribute text using data mining tools. The study has been able to develop an identity attribute metrics model using the Cosine Similarity distance measure and show that Cosine similarity measure can be used to quantify the identity attributes. The model has been tested on data that was mined and standardized using term weights; the outcome showed that the Cosine Similarity model can identify the unique owner of the digital identity attributes. The model also showed that it could identify a legitimate identity claimant from multiple claims. This model could add value to enhancing security in online activities by validating the true owner a digital identity. This model could also be used in multi modal tools for a robust online digital solution to arrest the challenges of online information security.

This research has developed an Identity Attribute Model that can be used to quantify the identity attributes from real space and cyber space. The model can identify the real owner of digital identity identify such a claimant from a number of identity claimants. This could therefore identify the bogus claimant of digital identity from the real ones. The outcome of this research could be augmented to already established techniques to form a robust multi modal tool of digital identity. The model would help to address the security challenge in identity management systems.

For future research interests, there is need to develop and implement the outcome of this research and build a multimodal solution. That solution would consolidate previous works in this area and come up with a single robust solution. Such a solution should recognize how much threat would be rid of in the online services and activities.

References

[1] J. I. Agbinya, N. Mastali, R. Islam and J. Phiri, "Design and implementation of multimodal digital identity management system using fingerprint matching and face recognition," 7th International Conference on Broadband Communications and Biomedical Applications, Melbourne, VIC, 2011, pp. 272-278, doi: 10.1109/IB2Com.2011.6217932.

[2] J. Phiri, T. J. Zhao, C. H. Zhu and J. Mbale, “Using Artificial Intelligence Techniques to Implement a Multifactor Authentication System”, International Journal of Computational Intelligence Systems, Volume 4, Issue Number 4, pp. 420-430, 2011. DOI: 10.1080/18756891.2011.9727801.

[3] M. S. Gaigole and M. A. Kalyankar, "The Study of Network Security with Its Penetrating Attacks and Possible Security Mechanisms," International Journal of Computer Science and Mobile Computing, vol. 4, no. 5, p. 729, 2015.

[4] C. Kwok, “Digital Identity Management System”, Faculty of Engineering, University of Technology, Sydney, Australia, 2007.

[5] Insight Report, “Identity in a Digital World A new chapter in the social contract”: World Economic Forum, p. 10, September 2018.

[6] K. I-Lung, "Securing mobile devices in the bMarchusiness environment", IBM Global Technology Services, Thought Leadership White Paper, pp. 2-10, October 2011.

[7] A. Pfitzmann, & M. Hansen Anonymity, “Unlinkability, Unobservability, Pseudonymity, and Identity Management-A Consolidated Proposal for Terminology”, 3 ed., p. 28, 2006.

[8] R. Bhasker and B. Kapoor, "Information Technology Security Management", Computer and Information Security Handbook, Burlingtone, Morgan Kaufmann Publishers - an imprint of Elsevier, pp. 259 -267, 2009.

[9] P. J. Windley, “Digital Identity”, Unmasking Identity Management Architecture (IMA), O’Reilly Media, pp. 13, 16, 17, 31, and 32, 2005

[10] G. Roussos, D. Peterson, and U. Patel, “Mobile Identity Management”: An Enacted View, International Journal of Electronic Commerce, Vol 8, Issue 1, pp 81-100, 2003.

[11] C. Satchell, G. Shanks, S. Howard, and J. Murphy, “Identity crisis: User perspectives on multiplicity and control in federated identity management” (2011).

[12] P. Curry (UK) and A. Nadalin (US), “International Standard ISO/IEC WD2 29003”: Text for ISO/IEC 2nd WD 29003 - Information technology – Security techniques – Identity proofing, pp43-44, 2013.

[13] J. I. Agbinya, R. Islam, and C. Kwok, “Development of Digital Environment IdentiTY (DEITY) System for Online Access”: Faculty of Engineering, University of Technology, Sydney, Australia French South African Technical Institute in Electronics (F’SATIE), Tshwane University of Technology, Pretoria, South Africa.

[14] I. Memon and Q. Arain, “Dynamic Path Privacy Protection Framework for Continuous Query Service over Road Networks”: Springer, 30th August 2016.

[15] M. Nieles, K. Dempsey, and V. Y. Pillitteri, “An Introduction to Information Security”: NIST Special Publication, National Institute of Standards and Technology, U.S. Department of Commerce, 800-12, Revision 1, June 2017.

[16] N. Luhmann, “Trust and Power”, Chichester, Wiley, 1979.

[17] R. C. Solomon and F. Flores, "Building Trust in Business, Politics, Relationships, and Life", Oxford, Oxford University Press, 2001.

[18] C. D. Schultz, "A Trust Framework Model for Situational Contexts**":** International Conference on Privacy, Security and Trust - Bridge the Gap Between PST Technologies and Business Services, PST 2006, Markham, Ontario, Canada, October 30 - November 1, 2006.

[19] I. Memon and Q. and Arain, "Map services based on multiple mix-zones with location privacy protection over road network." Wireless Personal Communications 97.2 (2017): 2617-2632.

[20] I. Memon1, H. Mirza2, Q. Arain3, and H. Memon, "Multiple mix zones de-correlation trajectory privacy model for road network": Springer, P. 1, 21 February 2019.

[21] C. Haines & R. Crouch, "Mathematical modeling and applications, Ability and competence frameworks": Springer, Modelling and applications in mathematics education, New York, NY, The 14th ICMI study, pp. 417-424, 2007.

[22] R. Lehrer & L. Schauble, "Origins and evaluation of model-based reasoning in mathematics and science, Beyond constructivism": Models and modeling perspectives on mathematics problem solving, learning, and teaching, pp. 59-70, 2003.

[23] E. Korkmaz and G. Üçoluk, "Choosing A Distance Metric for Automatic Word Categorization". In D.M.W. Powers (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, ACL, pp 111-120, 1998.

[24] H. Abdi, "Normalizing Data", In Neil Salkind (Ed.), Encyclopedia of Research Design. Thousand Oaks, CA: Sage, pp.1 and 3, 2010.

[25] D. L. Olson and D. Delen, "Advanced Data Mining Techniques": Springer, p. 119, 2008.

[26] B. S. Charulatha, P. Rodrigues, T. Chitralekha, and A. Rajaraman, "A Comparative Study of different metrics that can be used in Fuzzy Clustering Algorithms": National Conference on Architecture, Software systems and Green computing-(NCASG2013), Special Issue ISSN 2278-6856, p. 1, 2013.

[27] A. Singhal, "Data Warehousing and Data Mining Techniques for Cyber Security", Advances in Information Security,: Springer, New York, USA, p. 9, 2007.

[28] M. P. Campbell, G. E Cho, S. Nelson, C. Orum, J. V. Reynolds-Fleming, and I. Zavorine, "Term Weighting Schemes in Information Retrieval", National Security Agency, 1997.

[29] F. H. Lotfi and R. Fallahnejad, "Imprecise Shannon's Entropy and Multi Attribute Decision Making": Entropy, 2010, Vol 12, pp. 53-62.

[30] A. Delgado and B. Ayala, "A Computational Model Based on Shannon Entropy to Analyze Social Development in South America": International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol 8, Issue 11, September 2019.

[31] S. Vajapeyam, "Understanding Shannon's Entropy metric for Information" v.sriram.bir@gmail.com (March 24, 2014).

[32] W. Inambao, J. Phiri, and D. Kunda, " Digital Identity Modelling for Digital Financial Services in Zambia": Ictact Journal on Communication Technology, Vol 9, Issue 3, September 2018,

[33] M. K. Chinyemba and J. Phiri, "Gaps in the Management and Use of Biometric Data: A Case of Zambian Public and Private Institutions": Zambia Information Communication Technology (ICT) Journal, Vol. 2, Issue 1, pp. 35-43, 2018.

[34] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms": IEEE Transactions on Neural Networks, Vol. 16, Issue 3, May 2005.

[35] J. Phiri, T. J. Zhao, J. Mbale," Identity Attributes Mining, Metrics Composition and Information Fusion Implementation Using Fuzzy Inference System", Journal of Software, Vol 6, No. 6, pp.1025-1033, Jun 2011.

[36] J. Phiri, D. M. Zulu, J. I. Agbinya and T. Zhao, "Fuser block technologies performance based on identity attributes metrics models," 2013 Pan African International Conference on Information Science, Computing and Telecommunications (PACT), Lusaka, 2013, pp. 188-193, doi: 10.1109/SCAT.2013.7055112.

[37] J. W. Creswell, "Research Design: Qualitative, Quantitative, and mixed methods, Approaches", 4th ed. SAGE Publications, Thousand Oaks, 2013.

[38] P. Curry (UK), A. Nadalin (US), "International Standard ISO/IEC WD2 29003": Information technology - Security techniques - Identity Proofing International Organization for Standardization, International Electrotechnical Commission, 15th July 2013.

[39] Q. Arain, H. Memon, I. Memon, M. Memon, R. Shaikh, and Farman Ali Mangi, "Intelligent travel information platform based on location base services to predict user travel behavior from user-generated GPS traces": International Journal of Computers and Applications, Issue No . 3, Vol. 39, 155–168, 2017.

[40] S. Pandit and S. Gupta, "A Comparative Study on Distance Measuring Approaches for Clustering": International Journal of Research in Computer Science, White Globe Publications, eISSN 2249-8265, Vol. 2, Issue 1, p 1, 2011.

[41] S. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions": International Journal of Mathematics and methods in Applied Sciences, Vol. 1, Issue 4, p. 1, 2007.

[42] K. Maher1 and M. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering": International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7, Issue No. 4, pp. 1715-1720, 2016.

[43] O. E. Oduntan, I. A. Adeyanju, A. S. Falohun, and O. O. Obe, "A Comparative Analysis of Eucledian Distance and Cosine Similarity Measure for Automated Essay-Type Grading": Journal of Engineering and Applied Sciences, Vol. 13, Issue 11, pp. 4198-4204, 2018.

[44] E. Jyoti, Dr. Sanjeev Dhawan, Dr. Kulvinder Singh, "Comparison of Various Similarity Measure Techniques for Generating Recommendations for E-commerce Sites and Social Websites": American International Journal of Research in Science, Technology, Engineering & Mathematics, Vol 2, Issue 11, June-August, 2015, pp. 219-221.

[45] L. Zahrotun, " Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method": Computer Engineering and Applications, Vol. 5, Issue 1, February 2016.

[46] L. Chen and C. Chang, "A New Term Weighting Method by Introducing Class Information for Sentiment Classification of Textual Data": Proceeding of the International MultiConference of Engineers and Computer Scientists, Vol. 1, pp. 16-18, Hong Kong, 2011.

[47] C. Buckley, "The Importance of Proper Weighting Methods" pp. 349-352, August 1993.

[48] J. Beel, S. Langer, and B. Gipp, "TF-ID$_{ij}$F: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections": iConference, 2017.

[49] Y. Jiao, M. Cornec, and J. Jakubowicz, "An entropy-based term weighting scheme and its application in e-commerce search engine": International Symposium on Web Algorithms, Deauville, France, Jun 2015.