Identity Attributes Mining, Metrics Composition and Information Fusion Implementation Using Fuzzy Inference System

Jackson Phiri
Harbin Institute of Technology, School of Computer Science, Harbin, China
jackson. phiri@gmail.com

[1,a]Tie Jun Zhao and [2,b]Jameson Mbale
[1]Harbin Institute of Technology, School of Computer Science, Harbin, China
[2]University of Namibia, Department of Computer Science, Windhoek, Namibia
[a]tjzhao@mtlab.hit.edu.cn, [b]jmbale@unam.na

Abstract—**Term weight a technique in text mining and entropy from Shannon's information theory are both used to quantify information. In this paper, using term weight and entropy, Sugeno-Style fuzzy inference is envisaged in the implementation of information fusion in a multimode authentication system in an effort to provide a solution to identity theft and fraud. Three corpora are used to mine the identity attributes and generate the statistics required to compose the metrics values from the application forms and questionnaires using term weight and entropy. Triangular and Sigmoidally shaped membership functions are used in the fuzzification of the three inputs categories namely biometrics, pseudo metrics and device based credentials.**

Index Terms—**identity attributes, metrics, information fusion, fuzzy logic, authentication, term weight, entropy**

I. Introduction

The advent of the information technology has led to the computerisation of our day to day business and social life activities. Despite the numerous benefits that come with this development, it has also come with its own challenges including that of security [1] and how to analyse the large quantity of data being generated on a daily basis. Data mining or knowledge discovery has become a very useful branch of computer science in the last decade. It is a process of analyzing data from different sources and angles in order to summarizing it into useful information [2]. Today most services providers from both the public and private sectors do provide online application for their services. For some service providers, they have posted the application forms in portable document file (PDF) or word format. In this paper application forms for the various services offered both in the real and cyberspace are used as the source of the identity attributes. In addition we roll out a questionnaire to obtain the identity attributes based on the user's opinion. Three corpora namely the AntConc, ConcApp and the TextSTAT are then used to mine the identity attributes from these sources to generate the statistical information required to compute the term weight and entropy which are translated as the metrics values of these identity attributes [3][4][5]. These metrics are used in the implementation of an information fusion.

Today most business and government organizations are grappling with identity theft, identity fraud, virus attack, espionage and many other similar challenges [6]. The main aim in this paper is to address an area of security challenge in identity management called authentication through multi-factor authentication [7][8]. We propose a combination of biometrics, pseudo metrics and a device based credential such as a smart card for a user to be effectively authenticated [9]. Each of these identity attributes is assigned a weight which is composed through text mining technique called term weight and Shannon's information theory called entropy. These weights are combined or fused together in a technique of information fusion using Sugeno-Style fuzzy inference system. For the biometrics these are the ratios or weights obtained by comparing the biometric in the database and that submitted by the user during score level authentication [8]. The user needs to meet a specified threshold value set for a given multimode authentication system in order to be successfully authenticated. Multimode authentication will most likely make it very difficult for any imposter to forge the whole range of identity attributes required. The introduction of fuzzy inference system also brings in an element of intelligent behaviour while taking care of the fuzziness of some identity attributes such as biometrics (e.g. face recognition and signature recognition) and soft biometrics such as height, weight and skin colour [10][11].

II. Background Information

Identity management systems come in a wide range of implementations. Today with the cases of identity theft and fraud on the increase, we have seen an increasing usage of biometrics and Radio Frequency identification (RFID) cards. In [12], a survey recently conducted in

doi:10.4304/jsw.6.6.1025-1033

Europe shows how these technologies have entered the public space and being used on a wide scale. The public transport cards, the biometric passport, micro-payment systems, office ID tokens, customer loyalty cards, students' cards and bank cards are some of the applications using biometrics and RFID today [9]. In this paper, we categories the identity attributes into three major areas namely device metrics, pseudo metrics and biometrics [13]. Devices metrics will include all the identity attributes from the device based credential tokens such as the Media Access Control (MAC) address and Internet Protocol (IP) address for a personal computer or laptop [9]. Others include the International Mobile Equipment Identifier (IMEI) or the Subscriber Identifier Module for the mobile phone. Examples of pseudo metrics are the Personal Identity Number (PIN), passwords, randomly generated pass codes and secrete keywords [9][13]. Finally biometrics will include fingerprints, iris scan, face recognition, signatures recognition and many others [14]. This paper will not look at the details of capturing and storage of the identity attributes. Details of how a user submitting identity attributes from all or any of the three categories can be combined through information fusion for effective authentication of a user are considered in this paper. This is made possible by employing the metric values assigned to the identity attributes during the authentication process at the score level [10][15]. The metric values are used in an information fusion technique implementation using Sugeno-Style fuzzy inference system.

We therefore need to find a way of composing the metric values of the identity attributes. This is achieved by term weight and entropy. Information theory introduced by Claude Shannon in 1948 is used to quantify information and is based on probability theory and statistics. The most important quantities of information theory are entropy and mutual information [16]. Entropy is the information contained in a random variable where as mutual information is the amount of information in common between two random variables [17]. In this paper, we use the statistical information generated by the three corpora from a set of questionnaires and application forms to compute the empirical probabilities of the identity attributes [16]. The entropy is then computed using Eq. 1 given by [17];

$$H(p) = \sum_{i=1}^{n} p_i \log_2 \left(1/p_i\right) \qquad (1)$$

Where p_i is the probability of the identity attribute in the i^{th} sample space.

Data mining uses sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases in a quest to discover new meaning in data [18]. It has found a lot of applications areas including a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery [19]. There are different levels of data analysis techniques which apply various strategies. These include artificial neural networks applied in n*on-linear predictive models*, genetic algorithms for *optimization techniques*, decision trees in tree-shaped structures that represent *sets of decisions*, nearest neighbor method a technique that classifies each record in a dataset based on *k-nearest neighbor technique* [18][19]. Others are rule induction which is based on the extraction of useful if-then rules from data based on *statistical significance* and finally data visualization which uses the *visual interpretation of complex relationships in multidimensional data* [18][19].

In this paper three corpora software which integrate the above mentioned techniques at various levels are used to mine the text information in form of identity attributes and generate the statistical information of the identity attributes [3][4][5]. The statistical information is used in the composition of identity attributes metrics values. In text mining, the two fundamental tasks are text clustering and text classification. Text clustering include feature extraction, document clustering and post processing where as features extraction includes stop-word removal, steaming, term weighing, key feature extraction and matrix deduction [20]. In text mining, term weight is composed of term frequency and document frequency [20]. The *term frequency* $tf_{t,d}$ of the term *t* and document *d* is defined as the number of times that *t* occurs in *d*. The score for the *document-query pair* where the summation is over the term *t* in both the query *q* and the document *d* is given by [21];

$$Score = \sum\nolimits_{t \in q \cap d} (1 + \log tf_{t,d}) . \qquad (2)$$

Document Frequency (*df*) is the measure of the informativeness of the term *t*. The most useful component of *df* is the Inverse Document Frequency (*idf*). Collection Frequency (*cf*) of the term *t* is the number of occurrences of *t* in the collection, counting multiple occurrences [20][21]. In this paper we use the term weight to compose the identity attributes metrics. It is given by the following equation where *N* is number of the documents in the corpus under consideration [21];

$$w_{t,d} = (1 + \log tf_{t,d}) \times \log_{10} \frac{N}{df_t} . \qquad (3)$$

Most of the time human thinking and reasoning involve inexact information because much of human knowledge is vague and imprecise [22]. The sources and the nature of inexact information usually differ for different problem domains. The following are possible reasonable sources of inexactness of information; lack of adequate data, inconsistency of data, inherent human fuzzy concepts, matching of similar rather than identical situations, differing opinions, ignorance, imprecision in measurements and lack of available theory to describe a particular situation [23]. Most of the information used in user authentication systems such as biometrics, height, colour of eyes are not exactly and precise hence making fuzzy inference the best option. Two commonly used fuzzy inferences are Mamdani and Sugeno-Style inference system [22]. In FIS inference process, there are four major steps, which include fuzzification of the input variables, rule evaluation, aggregation of the rule outputs, and finally defuzzification [23]. Sugeno-Style fuzzy inference is used in this paper because it is computationally effective and works well with optimisation and adaptive techniques. In fuzzy set theory, fuzzy set *A* of the universe *X* is defined by the function $\mu_A(x)$ called the membership function of *x* in set *A*. MatLab toolbox is used in this paper and includes eleven different types of built-in membership function which are in turn, built from several basic functions which include the *Piece-Wise Linear functions*, the *Gaussian Distribution function*, the *Sigmoid Curve* and *Quadratic and Cubic Polynomial Curves* [24]. In this paper we use the *Triangular membership function* (*trimf*) and the *Sigmoidally shaped membership function* (*sigmf*) [23][24]. Sigmoidally function depends on two valuables *a* and *c* and is used in fuzzification of biometrics inputs. It is represented by the following function [22][24];

$$f(x,a,c)=\frac{1}{1+e^{-a(x-c)}} . \qquad (4)$$

Triangular membership function is a function of a vector *x* and depends on three scalar parameters *a*, *b* and *c*. The parameters *a* and *c* locate the feet of the triangle while the parameter *b* locate the peak and is given by [23][24];

$$f(x:a,b,c)=\max\left(\min\left(\frac{x-a}{b-a},\frac{c-x}{c-b}\right),0\right) . \qquad (5)$$

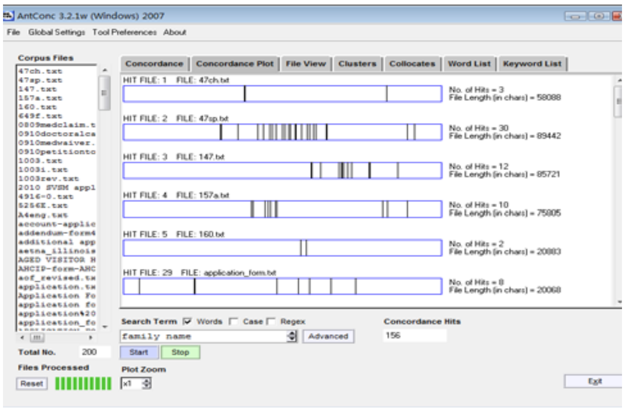
III. Related Works

Information fusion has seen a lot of applications in the areas of robotics, geographical information systems and data mining technologies. For example, [25] uses a combination of artificial neural networks, Dempster-Shafer evidence theory-based information fusion and Shannon entropy to form a weighted and selective information fusion technique to reduce the impact uncertainties on structure damage identification. In [26] the quantitative metrics are proposed to objectively evaluate the quality of fused imagery. Information fusion in biometric verification for three biometrics at the matching score level is provided in [27] while [28] provides a two-level approach to multimodal biometric verification. In this paper we introduce the pseudo metrics and device metrics in addition to the biometrics and use Sugeno-Style fuzzy inference system to implement information fusion during multimode authentication.

IV. Methodology

The methodology begins by looking at the sources of credential identity attributes. In this paper the various services offered by both the public and private sectors are used as the first source of the identity attributes. These categories include the home affairs department services such as the passport and birth certificate, transport department such as car and driver's license registration, insurance companies' services, financial services, health care services and education services. Application forms in PDF format from the service providers' websites are downloaded and then converted into the text file for analysis using three corpora. The online application forms are captured using the Webspider integrated into the TextSTAT corpus. Secondly, we roll out questionnaires targeting international respondents especially those from the G20 countries for the opinion based responses on the identity attributes deemed as important and can uniquely identify the user in their respective countries. Social networks and email addresses are used as the vehicle for capturing the responses from the electronic copy of the questionnaire. A total of 200 application forms and 100 questionnaires are used in this paper and each of the three corpora is then used to generate the word frequency, document frequency and collection frequency for the various identity attributes from the two sources. This creates six sample spaces of the identity attributes. Fig. 1 for example shows a total of 200 application forms in the corpus with *156 hit* (*collection frequency*) for the search of a *family name* using *AntConc corpus*. Also shown is the number of hits (term frequency) for each of the 156 application forms with the family name. For example the file named *47ch.txt* has a term frequency of *3*, meaning in this application form, the family name appears three times. This query process is repeated using the three corpora to generate the required statistics from the two sources. These statistical results are then used to compute the entropy yield and term weights of the identity attributes.

We begin by sampling fifty most frequent identity attributes from the statistical results generated by our three corpora. These are used to form the sample space of identity attributes used to generate probabilities of randomly picking any of the fifty identity attributes from



47ch.txt
47sp.txt
147.txt
157a.txt
160.txt
649f.txt
0809medclaim.t
0910doctoralca
0910medwaiver.
0910petitionto
1003.txt
1003i.txt
1003rev.txt
2010 SVSM appl
4916-0.txt
5256E.txt
A4eng.txt
account-applic
addendum-form4
additional app
aetna_illinois
AGED VISITOR H
AHCIP-form-AHC
aof_revised.tx
application.tx
Application Fo
application fo
application%20
application_fo

HIT FILE: 1 FILE: 47ch.txt
No. of Hits = 3
File Length (in chars) = 58088

HIT FILE: 2 FILE: 47sp.txt
No. of Hits = 30
File Length (in chars) = 89442

HIT FILE: 3 FILE: 147.txt
No. of Hits = 12
File Length (in chars) = 85721

HIT FILE: 4 FILE: 157a.txt
No. of Hits = 10
File Length (in chars) = 75805

HIT FILE: 5 FILE: 160.txt
No. of Hits = 2
File Length (in chars) = 20883

HIT FILE: 29 FILE: application_form.txt
No. of Hits = 8
File Length (in chars) = 20068

Search Term ☑ Words ☐ Case ☐ Regex
family name
Concordance Hits 156
Total No. 200
Files Processed
x1

Figure 1. AntConc corpus search results for family name showing term and document frequencies.

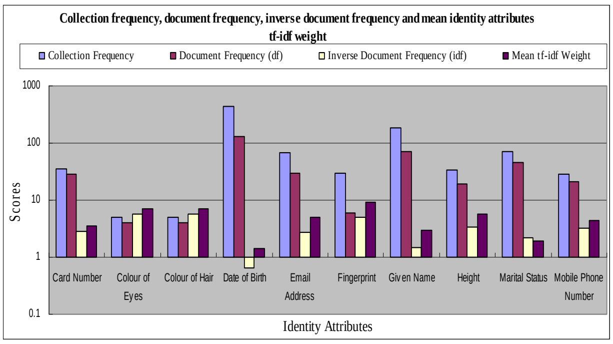


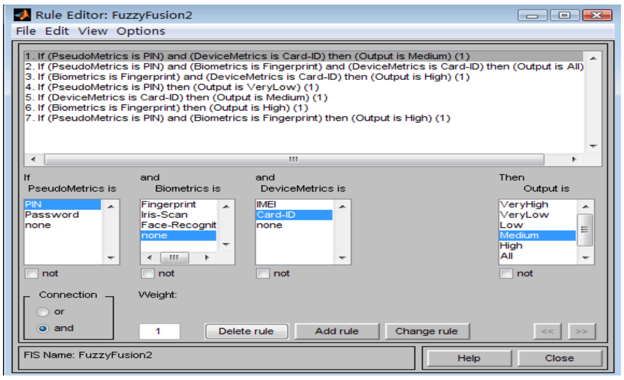
Figure 2. Collection frequency, document frequency, inverse document frequency, and the mean computed tf-idf weight

the two sources. Using (2), the identity attributes inverse document frequency and then the term weights are computed for each of the fifty identity attribute. Fig. 2 shows the results of 10 selected identity attributes showing the collection frequency, document frequency, inverse document frequency and the computed mean tf-idf weight. In the second method, we begin by using the generated collection frequency by the three corpora to compute the empirical probabilities of each of the fifty selected identity attribute in the sample space. Using (1), the entropy yield of each identity attributes from the six sample spaces is then computed, Where p_i is the probability of the identity attribute in each of the generated six sample spaces (two sources and using three corpora). Table I shows the computed term weight and entropy yield for a selected set of ten identity attributes. Using Sugeno-Style fuzzy inference system, an information fusion technique is then designed and implemented with *three inputs*, *seven rules* and *one output*. The three categories namely pseudo metrics, biometrics and device metrics are used to design an information fusion technique. For each of the categories a single identity attributes is used in the example implementation as follows; a fingerprint is used to represent the biometrics, a PIN number is used for the pseudo metrics and finally the card number is used for the device metrics.

We compare the results from the two implementations which use the term weight metric values and the entropy metric values, and finally draw the conclusions.

TABLE I.
COMPUTED ENTROPY AND TERM WEIGHT METRICS VALUES.

#	Identity Attribute	Mean Identity Attributes Entropy	Mean Identity Attributes Term Weight
1	Card ID	0.048644	3.50358
2	Date of Birth	0.263487	1.44031
3	Email	0.085400	5.06057
4	Fingerprint	0.042741	8.97436
5	Given Name	0.217056	2.94966
6	Height	0.056102	5.57110
7	Mobile Phone ID	0.047691	2.76955
8	PIN	0.053285	1.543401
9	Sex	0.215595	1.82216
10	Surname	0.379362	1.34112


Rule Editor: FuzzyFusion2

1. If (PseudoMetrics is PIN) and (DeviceMetrics is Card-ID) then (Output is Medium) (1)
2. If (PseudoMetrics is PIN) and (Biometrics is Fingerprint) and (DeviceMetrics is Card-ID) then (Output is All)
3. If (Biometrics is Fingerprint) and (DeviceMetrics is Card-ID) then (Output is High) (1)
4. If (PseudoMetrics is PIN) then (Output is VeryLow) (1)
5. If (DeviceMetrics is Card-ID) then (Output is Medium) (1)
6. If (Biometrics is Fingerprint) then (Output is High) (1)
7. If (PseudoMetrics is PIN) and (Biometrics is Fingerprint) then (Output is High) (1)

If PseudoMetrics is: PIN (selected), Password, none — □ not
and Biometrics is: Fingerprint, Iris-Scan, Face-Recognit, none (selected) — □ not
and DeviceMetrics is: IMEI, Card-ID (selected), none — □ not
Then Output is: VeryHigh, VeryLow, Low, Medium (selected), High, All — □ not

Connection: ○ or ● and

Weight: 1

FIS Name: FuzzyFusion2

Figure 3. Five fuzzy rules and linguistic variables

A. *Case I: Using Term Weight Metric Values*

The computed metric values obtained by using the technique of term weighing are used to assign scores for the input variables to the Sugeno-Stlye fuzzy inference system as shown in Table II. The biometric fingerprint has a maximum score of 8.97, the card number 3.50 and finally the PIN number 1.54. Equation 4 which is a Sigmoidally (sigmf) shaped is used as the membership function for the fingerprint. The membership function is labelled as *fingerprint* and the values of *a* and *b* are assigned 0.8 and 5.24 respectively as shown in Table II and Fig. 3. Triangular (trimf) shaped represented by (3) is used as a membership function for both the pseudo metrics and device metrics. For the pseudo metrics, the function is labelled as *PIN* and the values of *a*, *b* and *c* are 0.54, 1.54 and 2.54 respectively. For the device metrics, the membership function is labelled as *Card-ID* and the respective values of *a*, *b* and *c* are 2.5, 3.5 and 4.5.

B..*Case II: Using Entropy Metric Values*

Sigmoidally function is used in the fuzzification of biometrics while Triangular (trimf) shaped function is used in the fuzzification of both the pseudo metrics and device metrics. The functional labelling is similar to case I but differ in the values of the function variables. For biometrics the values of the variables *a* and *b* are 50 and 0.05 respectively. For the pseudo metrics, the values of *a*, *b* and *c* are 0.043, 0.053 and 0.063 respectively, while for the device metrics, the membership function variables *a*, *b* and *c* are assigned the values 0.038, 0.048 and 0.058 respectively as shown in Table III.

The fuzzification functions both in case I and case II are mapped to similar output variables. Since we are

TABLE II.
TERM WEIGHT INPUT METRIC VALUES AND FUNCTION VARIABLE VALUES

#	Grouping	Inputs	Term Weight	Function valuable values
1	Device Metrics	Card Number	3.5035	a=2.5, b=3.5, c=4.5
2	Pseudo Metrics	PIN Number	1.5434	a=0.54, b=1.54 c=2.54
3	Biometrics	Fingerprint	8.9744	a=0.8, b=5.24

TABLE III.
ENTROPY INPUT METRIC VALUES AND FUNCTION VARIABLE VALUES

#	Grouping	Inputs	Entropy	Function valuable values
1	Device Metrics	Card Number	0.048644	a=0.038, b=0.048, c=0.058
2	Pseudo Metrics	PIN Number	0.053285	a=0.043, b=0.053 c=0.063
3	Biometrics	Fingerprint	0.0427	a=0.5, b=0.05

TABLE IV.
INPUTS AND CORRESPONDING OUTPUTS TO THE SUGENO-STYLE FIS INFORMATION FUSION TECHNIQUE USING TERM WEIGHT METRICS

#	Pseudo	Biometrics	Device Metrics	Output
1	0.000	0.000	0.000	0.009
2	1.500	0.000	0.000	0.118
3	0.000	8.970	0.000	0.571
4	0.000	0.000	3.500	0.318
5	1.500	0.000	3.500	0.841
6	1.500	8.970	0.000	1.240
7	0.000	8.970	3.500	1.440
8	1.500	8.970	3.500	3.460

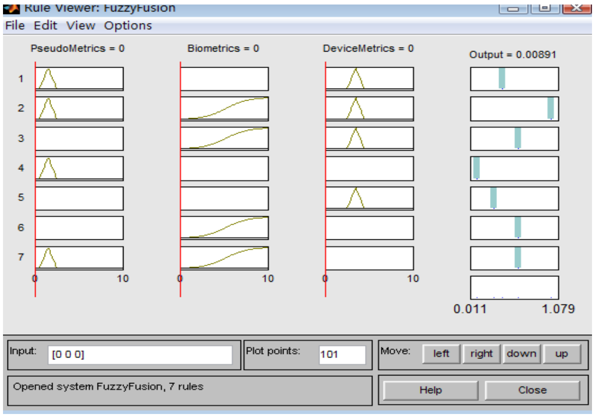
using the Segeno-Style, the output linguistic variables are assigned discrete output values between the range of zero and one as follows; *All* (with a score of 0.9) which is the output when all the three inputs identity attributes matched the copies in the databases and are assigned the maximum possible values from term weights or entropy metrics. The others are *VeryHigh* with a score of 0.7, *High* with a score of 0.5, *Medium* with a score of 0.4, Low with a score of 0.2 and *VeryLow* with a score of 0.1. Using the three inputs, the fuzzification functions and corresponding values of the variables, output linguistic variables, seven rules were then formulated as shown in Fig. 3. The weighted average is used to come up with the defizzification crisp output value which is then used to implement multimode authentication [23][25].

V. Results and Discussion

Fig. 1 above shows an example search for the statistical results generated from 200 application forms and 100 questionnaires by using the AntConc, ConcApp and TextSTAT corpora. These statistical results generated by the three corpora are first used to compute the identity attributes term weights and then the identity attributes entropy yield. Fig. 2 shows an extract of 10 identity attributes from the 50 identity attributes analyzed in this paper showing the collection frequency, document frequency, inverse document frequency, and the computed mean weight. Table I on the other hand shows the computed term weights and corresponding entropy yield for a selected set of ten identity attributes. Three of these identity attributes namely the fingerprint, card unique identity number (Card-ID) and PIN number are then used to implement an information fusion technique using Sugeno-Style fuzzy inference system. Using these three identity attributes as described in the methodology, seven rules are formulated as shown in Fig. 3, Table VI shows a set of results generated using term weight metric values. For the pseudo metrics and device metrics, the values assigned to these identity attributes are precise. It is either you have the correct PIN or card number or the wrong value. There is no middle ground. This means that, these two identity attributes can either be assigned a zero when the wrong pseudo or device metric is submitted for authentication such that it does not much the copy in the database or is assigned a computed metric value if the correct identity attribute is submitted such that it matched the copy in the database during score level authentication. For the biometrics, the output is not precise, but a ratio usually between zero and one depending on the implementation [29]. In our case, for both the biometrics and the two identity attributes, we use two different sets of the metric values. The term weight metrics inputs values are between zero and ten while the entropy metric input values are between zero and one. We now look at the results from the two input metrics as case I and case II.

A. Case I: Results When Using Term Weight Metrics Values

With three possible inputs, we have a minimum of eight different possible outputs. If none of the submitted identity attribute is correct such that each identity attribute is assigned a zero at the score level authentication, the Sugeno-Style fuzzy information inference information fusion technique then gives an output of 0.00891 as shown in Table IV and Fig. 4. On the other hand, when all the submitted copies matched the copies in the databases and are assigned the maximum possible scores as shown in Table II (term weights), then the maximum output of the information fusion system is *3.46* as shown in Fig. 5. Table IV shows the other six combinations and the respective outputs which are within the range between 0. 00891 and 3.46. For example, when biometric is the only correct identity attribute of all the submitted identity attributes, the system will give an output of 0.571 as shown in the fourth row of Table IV. On the other hand, when the pseudo metric is the only

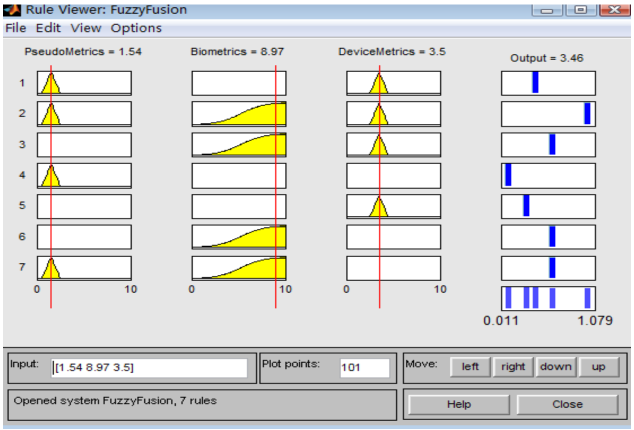


Input: [0 0 0]

Plot points: 101

Opened system FuzzyFusion, 7 rules

Figure 4. Fuzzy information fusion output when none of the three identity attributes is correct using term weight metrics values



PseudoMetrics = 1.54
Biometrics = 8.97
DeviceMetrics = 3.5
Output = 3.46

Input: [1.54 8.97 3.5]
Plot points: 101

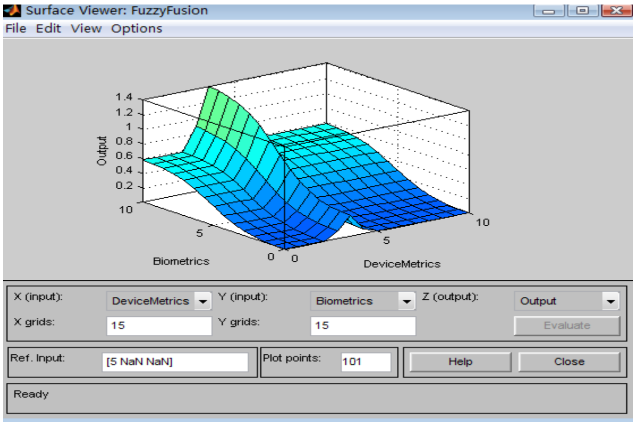
Opened system FuzzyFusion, 7 rules

Figure 5. Fuzzy information fusion output when all of the three identity attributes submitted are correct using term weight metrics values

correct identity attribute, the system gives an output of 0.118. However, when the device metric (card identity) is the only correct value the system gives 0.318. More useful outputs are the combination of the card identity and the PIN number which we use in our day to day lives especially with the bank credit and debit cards. This combination gives 0.841 which is slightly higher than the value obtained by a biometric alone. Combining the three then gives even a more interesting value of 3.45 which is an optimum value for the system. While the other two metrics (pseudo and device metrics) are precise, biometrics are not and will vary from time to time. For example when the biometric value drops from the optimum of 8.97 to 5.0 while keeping the other two values constant at their optimum, the system gives an output of 2.06 giving a difference of 1.4. This shows how biometrics may influence the operation of the system which is an essential feature as the usage of biomerics continues to rise. Fig. 6 shows the surface view of the biometric input, device metric input and the output.

B. *Case II: Results When Using Entropy Metric Values*

Case I and case II use similar functions for the fuzzification of the biometrics, pseudo metrics and device metrics. The difference comes in the values of the variables of the fuzzification functions. When none of the submitted identity attributes matched the copies in the database such that each was assigned a zero or a value close to zero for a biometric identity attribute, then the system gives an output of 0.0367 as shown in Fig. 7. However, when all the submitted identity attributes matched the copies in the database and each was assigned the maximum possible value as computed in Table III, then the Sugeno-Style fuzzy system information fusion technique gives an output of 2.62 as shown in Fig. 8. Fig. 9 shows the surface view of the device metrics, pseudo metrics and the corresponding output value. The rest of the values for the other six combinations are spread in between this range as shown in Table V. For example, when one is using the correct card device as the only correct identity attribute the system gives 0.373. On the other hand, when the biometric is the only correct identity attributes the system gives 0.391 and finally when the PIN number is the only correct identity attribute, the system gives a value of 0.173. Using this range it is possible to implement a multimode authentication system.



X (input): DeviceMetrics
Y (input): Biometrics
Z (output): Output
X grids: 15
Y grids: 15
Ref. Input: [5 NaN NaN]
Plot points: 101

Ready

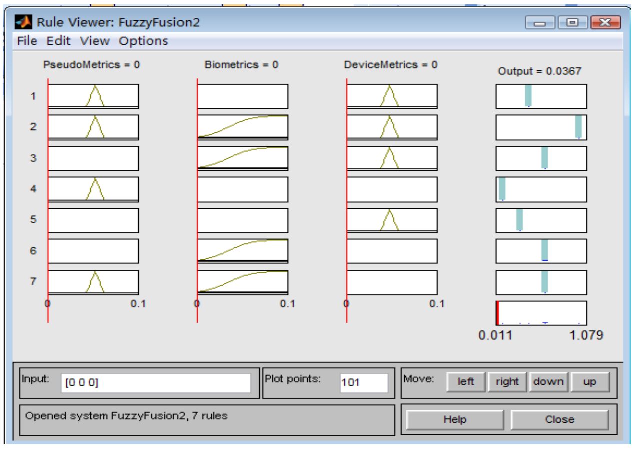
Figure 6. Surface viewer of the biometrics, device metrics and the respective output of the when using term weight metrics values

TABLE V.
INPUTS AND CORRESPONDING OUTPUTS TO THE SUGENO-STYLE FIS INFORMATION FUSION TECHNIQUE USING ENTROPY METRICS

#	Pseudo	Biometrics	Device Metrics	Output
1	0.000	0.000	0.000	0.0367
2	0.053	0.000	0.000	0.173
3	0.000	0.043	0.000	0.391
4	0.000	0.000	0.048	0.373
5	0.053	0.000	0.048	0.971
6	0.053	0.043	0.000	0.881
7	0.000	0.043	0.048	1.080
8	1.500	0.043	0.048	2.620

C. *Comparison of the results*

The inputs from term weight metric values are in the range between one and ten, while those from entropy



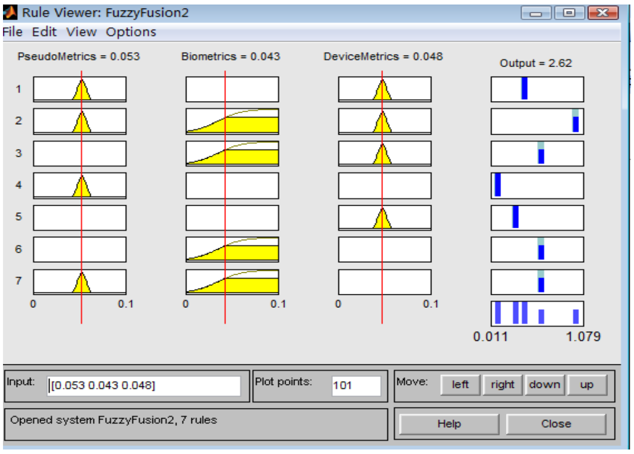
PseudoMetrics = 0
Biometrics = 0
DeviceMetrics = 0
Output = 0.0367

Input: [0 0 0]
Plot points: 101

Opened system FuzzyFusion2, 7 rules

Figure 7. Fuzzy information fusion output when none of the three identity attributes is correct using entropy metrics values

metrics are in the range between zero and one. Using P to represent a PIN number, F a fingerprint input and C a card based credential token, Fig. 10 shows a comparison of the output values from the two sets of the metric values. To one decimal place Table VI shows the output values reflected in Fig. 10. It can be seen that by using the entropy metric values, the range between the lowest possible output (0.037) and the largest possible output (2.62) is lower (2.58) as compared to that when using term weight metrics whose range is 3.45. In both cases, the lowest possible output when none of the submitted identity attribute is correct is zero to one decimal place as shown in Table VI. The maximum possible output in both cases is more than 2.5 to one decimal place. Using this range, it is possible to implement a multimode authentication system that requires a user to meet a certain pre-defined threshold value in order to be authenticated. Intelligent systems like this one using fuzzy logic will most likely make it very difficult for imposters to guess or have access to all the three categories of identity attributes.



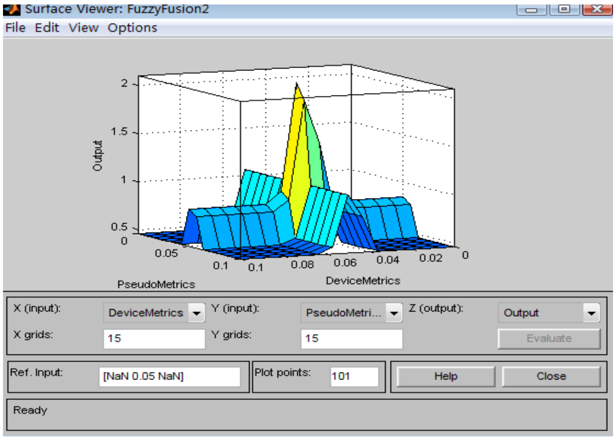
Input: [0.053 0.043 0.048]

Plot points: 101

Opened system FuzzyFusion2, 7 rules

Figure 8. Fuzzy information fusion outputs when all of the three identity attributes submitted are correct using entropy metrics values

A similar example using artificial neural networks with four inputs, four neurons in the hidden layer, one neuron in the output layer, sigmoid function as the transfer function and Back propagation algorithm for learning, the network successfully yielded values between 0.131854 and 0.912294 to six decimal places. While neural networks have the ability to learn, they do not have the ability to take into account the fuzziness or imprecision of the input identity attributes such as biometrics which are becoming central to most authentication systems.



X (input): DeviceMetrics

Y (input): PseudoMetri...

Z (output): Output

X grids: 15

Y grids: 15

Ref. Input: [NaN 0.05 NaN]

Plot points: 101

Ready

Figure 9. Surface viewer of the pseudo metrics, device metrics and the respective output of the using entropy metric values

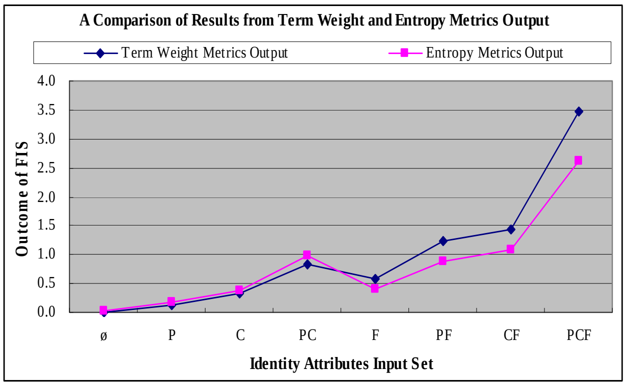


Figure 10. A comparison of the outputs from term weight metrics values and entropy metric values, where ø = none of the identity attribute is correct, P for PIN number, C for card number and F for fingerprint

Identity fraud and theft have been on the increase in the last decade and have become a major concern for the public and private sectors especially as they relate to problems like terrorism, money laundering and financial crime, drug trafficking and weapons smuggling [1][6]. With the emerging security concerns in the areas of immigration, border crossings, airline passengers and driver's licenses [8][14], introducing intelligent technologies to online authentication systems will play a very big role to help reduce cases of identity theft and fraud seen on most online services today.

VII. Conclusion

In this paper, we demonstrated data mining for a novel application. Identity attributes metrics were successful composed using the identity attributes mined using three corpora. Using the composed metric model, an artificial intelligent technology called fuzzy inference system was used to combine three categories of identity attributes through information fusion for optimum recognition of the user during multi-modal authentication. In this implementation, a user is required to submit a given range of identity attributes to enable him/her meet the

TABLE VI.
Comparison of Inputs and Corresponding Outputs to the Sugeno-Style FIS Using Entropy and Term Weight

Input	ø	P	C	PC	F	PF	CF	PCF
Term Weight Metrics Output	0.0	0.1	0.3	0.8	0.6	1.2	1.4	3.5
Entropy Metrics Output	0.0	0.2	0.4	1.0	0.4	0.9	1.1	2.6

threshold value set in a given authentication system. Different applications will require different input values and hence will have different threshold values. For example, 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 defence against imposters. This will consequently help to reduce the cases of identity theft and fraud.

An example of areas where this work is directly applicable include electronic doors leading to secure areas, transport cards, micro-payment systems cards, national identity cards, ATM applications system, online banking, student record systems, e-governments, border security programs that include travel security systems with passport, ticket, and baggage verification systems. In the future, an additional fourth category of inputs which would take into account identity attributes such as the name, date of birth, address and other acquired identity attributes will be considered. Also information fusion technique such as neural-fuzzy inference, Dempster-Shafer, Hidden Markov Model will be considered in the implementation.

REFERENCES

[1] R. Dhamija and L. Dusseault, “The Seven Flaws of Identity Management: Usability and Security Challenges,” *Security & Privacy, IEEE*. Vol. 6, 2008, doi = dx.doi.org/10.1109/MSP.2008.49.

[2] S. Liao, et al, “Ontology-Based Data Mining Approach Implemented on Exploring Product and Brand Spectrum,” *Expert Systems with Applications*, Volume 36, Issue 9, pp. 11730 – 11744, November 2009, doi: 10.1016/j.eswa.2009.04.030.

[3] L. Anthony. AntConc, “Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom,” *Professional Communication Conference Proceedings*, pp. 729, 2005, doi = 10.1109/IPCC.2005.1494244

[4] ConcApp Corpus, Available online (September 2010) at: http://www.edict.com.hk/PUB/concapp/

[5] TextSTAT Corpus, Available online (October 2010) at: http://neon.niederlandistik.fu-berlin.de/en/textstat/

[6] S. Mike, “Unify and Simplify: Re-Thinking Identity Management,” Network Security, Vol. 2006, 2006, doi = dx.doi.org/10.1016/S1353-4858(06)70411-1

[7] M. Hansen, A. Schwartz and A. Cooper, “Privacy and Identity Management,” *Security & Privacy, IEEE* Vol. 6, 2008, doi: 10.1109/MSP.2008.41.

[8] A. Bhargav-Spantzel, A. C. Squicciarini, E. Bertino, S. Modi, M. Young, and S. J. Elliott, “Privacy Preserving Multi-Factor Authentication with Biometric,” *Journal of Computer Security*, 2007

[9] B. Geoff, “The Use of Hardware Tokens for Identity Management,” *Information Security Technical Repor*t, Elsevier, Vol. 9 2004, doi = dx.doi.org/10.1016/S1363-4127(04)00012-3.

[10] M. He, et al, “Performance Evaluation of Score Level Fusion in Multimodal Biometric Systems,” *Pattern Recognition*, Volume 43, Issue 5, pp. 1789-1800, May 2010, doi: 10.1016/j.patcog.2009.11.018.

[11] L. Nanni, A. Lumini, S. Brahnam, “Likelihood Ratio Based Features for a Trained Biometric Score Fusion,” *Expert Systems with Applications*, Volume 38, Issue 1, January 2011, Pages 58-63, doi: 10.1016/j.eswa.2010.06.006.

[12] European Technology Assessment Group, “RFID and identity management in everyday life,” Available online (October 2010) at: http://www.europarl.europa.eu/stoa/publications/studies/stoa182_en.pdf

[13] J. I. Agbinya, R. Islam and C. Kwok, “Development of Digital Environment Identity (DEITY) System for Online Access,” *Broadband Communications, Information Technology & Biomedical Applications, Third International Conference*, Australia pp. 23 – 26, November 2008, doi: 10.1109/BROADCOM.2008.52.

[14] J. L. Wayman, “Biometrics in Identity Management Systems,” Security & Privacy, IEEE Vol. 6, 2008.

[15] J. Phiri, J. I. Agbinya, J.I, Modelling and Information Fusion in Digital Identity Management Systems, Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, 2006. ICN/ICONS/MCL 2006. International Conference on , vol., no., pp. 181- 181, 23-29 April 2006, doi: 10.1109/ICNICONSMCL.2006.152

[16] B. A. Strange, A. Dugginsa, W. Pennya, R. J. Dolana and K. J. Friston, “Information Theory, Novelty and Hippocampal Responses: Unpredicted or Unpredictable?” Elsevier Science Direct Journal, Neural Networks, Vol. 18, 2005, doi:10.1016/j.neunet.2004.12.004.

[17] R. Togneri and S. J. C. DeSilva, “Fundamentals of Information Theory and Coding Design,” Chapman & Hall Press, Florida, 2002.

[18] E.W.T. Ngai, Li Xiu, D.C.K. Chau, “Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification,” *Expert Systems with Applications*, Vol. 36, Issue 2, pp. 2592-2602, March 2009, doi: 10.1016/j.eswa.2008.02.021.

[19] J. H. Suh, C. H. Park, S. H. Jeon, “Applying Text And Data Mining Techniques to Forecasting the Trend of Petitions Filed to e-People,” *Expert Systems with Applications*, Vol. 37, Issue 10, pp. 7255-7268, October 2010, doi: 10.1016/j.eswa.2010.04.002.

[20] L. Chung-Hong and Y. Hsin-Chang, “A Multilingual Text Mining Approach Based on Self-Organizing Maps,” Springer Netherlands, Vol 8, p.3, 2003, doi = 10.1023/A:1023250105036

[21] V. Avram, “Defining Metrics to Automate the Quantitative Analysis of Textual Information Within a Web Page,” Application of Information and Communication Technologies, AICT 2009 International Conference, pp.1-5, October 2009, doi: dx.doi.org/10.1109/ICAICT.2009.5372575.

[22] C. D. Manning, P. Raghavan and H. Schutze, “Introduction to Information Retrieval,” Cambridge Press, 2008.

[23] L. Mazlack, S. Coppock, “Granulating Data on Non-Scalar Attribute Values,” Fuzzy Systems, Proceedings of the 2002 IEEE International Conference, pp.944-949, 2002, doi:10.1109/FUZZ.2002.1006631.

[24] M. Negnevitsky. Artificial Intelligence, “A Guide to Intelligent Systems,” Second Edition, Addison Wesley, Tasmania, 2004.

[25] Fuzzy Inference System Toolbox Manual, Available (October 2010) at: http://www.mathworks.com/access/helpdesk/help/pdf_doc/fuzzy/rn.pdf.

[26] R. M. Hassan, B. Nath and M. Kirley, “A Fusion Model of HMM, ANN and GA for Stock Market Forecasting,” Expert Systems with Applications, Elsevier, Vol. 33-1, 2007, doi:10.1016/j.eswa.2006.04.007.

[27] Y. Zheng, A. E. Essock, C. B. Hansen and M. A. Haun, “A New Metric Based on Extended Spatial Frequency and its Application to DWT Based Fusion Algorithms,” Elsevier, Information Fusion. Vol. 8, 2007, doi:10.1016/j.inffus.2005.04.003.

[28] M. Cheung, M. Mak and S. Kungi, “A Two-Level Fusion Approach to Multimodal Biometric Verification,” ICASSIP IEEE Conference, pp. 485, 2005, doi:10.1109/ICASSP.2005.1416346.

[29] A. Ross, A. K. Jain, and J. Qian, “Information Fusion in Biometrics,” *In Proceedings of the Third international Conference on Audio- and Video-Based Biometric Person Authentication*, J. Bigün and F. Smeraldi, Eds. Lecture Notes In Computer Science, vol. 2091. Springer-Verlag, London, pp. 354-359, June 2001.

Jackson Phiri received his Bachelor of Computer Science at the University of Zambia (Zambia) in 2004 and his Master of Science (MSc) in computer science at the University of the Western Cape (South Africa) in 2007. He is currently a PhD student at Harbin Institute of Technology (China).

He works as a Lecturer at the University of Zambia (Zambia) and his research interest include data mining, digital identity management systems and applied artificial intelligent technologies.

Tie-Jun Zhao received his PhD degree at Harbin Institute of Technology (HIT) in 1997 and now is a professor (PhD supervisor) at HIT.

He is currently vice director of Minister of Education-Microsoft (MOE-MS)Key Laboratory on Natural Language Processing and Speech, in the School of Computer Science at HIT, China. His research interests include Applied Artificial Intelligent, Human Language Technology, and Signal Processing.

.

Jameson Mbale received his PhD degree in Computer Science from Harbin Institute of Technology, China, in 2003. He obtained M.Sc. degree in Computer Science from Shanghai University in 1996 (China) and B.A. in Mathematics and Computer Science at University of Zambia in 1993 in Zambia.

He is a Senior Lecturer and Head of the Department of Computer Science at the University of Namibia. He lectures/teaches the following courses: Telecommunications, Foundations of Data Communications, Networking and Emerging Technologies, Internet Technologies and Applications, Introduction to Network Security, Computer Networks, Network Systems Security, Wireless and Mobile Computing, Network Administration, Advanced Databases, and Software Engineering. He is CISCO and IT-Essentials instructor. His research interests include network security, wireless networking and telecommunications.