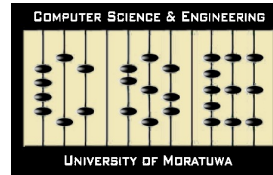
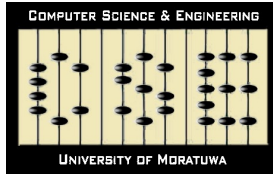


Measuring the Correlation of Personal Identity Documents in Structured Format

Sachithra Dangalla, Chanaka Lakmal, Chamin Wickramarathna, Chandu Herath,
Gihan Dias, Shantha Fernando



Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka



idstack^{one}

The Decentralized Protocol for Document Verification built on Digital Signatures

Our Team



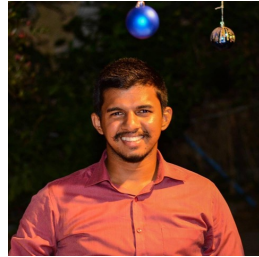
Chanaka Lakmal

BSc Eng. (Hons)
(Moratuwa)



Sachithra Dangalla

BSc Eng. (Hons)
(Moratuwa)



Chamin Wickramarathna

BSc Eng. (Hons)
(Moratuwa)



Chandu Herath

BSc Eng. (Hons)
(Moratuwa)

Supervisors



Prof. Gihan Dias

PhD (UCD), MSc (UCSB),
BSc Eng. (Hons)
(Moratuwa), MIE (SL),
CEng

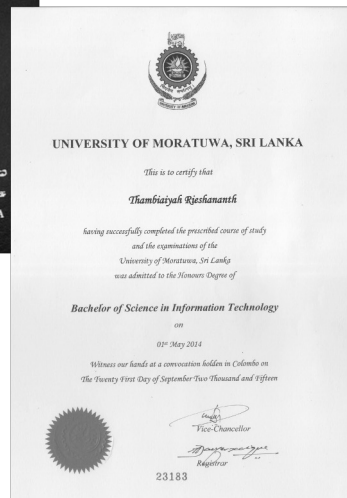
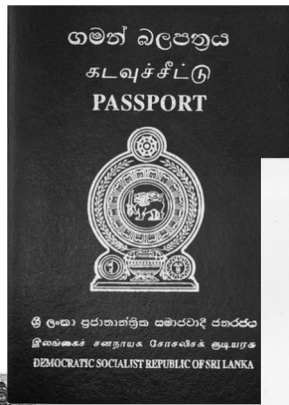
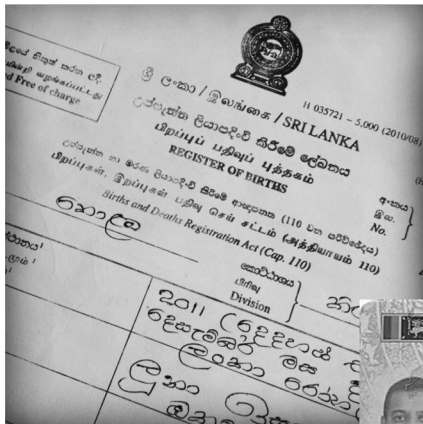


Eng. Dr. Shantha Fernando

PhD (TU Delft), MPhil
(Moratuwa), BSc Eng. (Hons)
(Moratuwa), MIE (SL), MIEE
(UK), CEng

Problem

In countries like Sri Lanka the document verification process is highly based on the printed copies.



Problem

Documents have **unpredictable layouts** that have raw information and there is **no mechanism** in the current research field to **extract them into common format of machine readable**.

The document is subjected to a **process of verification** by an authority by using a **hand written signature**.

Non existence of a mechanism in the current research field that can **calculate trustworthiness** of individual document.

Non existence of a mechanism in the current research field that can **calculate correlation** among set of documents belonging to a person.

Motivation

On average, **3.1 days** are added to most processes in order to collect **physical signatures** and eSignatures **reduce document turnaround time by 80%**

Source: *AIIM White paper study* - <http://www.aiim.org/pdfdocuments/MIWP-DigitalSignatures-2013.pdf>

The Electronic Transactions Act (ETA) No. 19 of 2006 (Section 7) gives **e-signatures the same legal weight as traditional hand-written signatures.**

Source: *Electronic Transaction Act, No.19 of 2006 (Section 7), Sri Lanka*

In budgets for the years 2016 and 2017, **the SL government allocated LKR 15 billion** to implement policy of digitalizing the economy.

Source: *Budget Speech 2017, Ministry of Finance, Sri Lanka*

ICTA implement e-Document Attesting System at Ministry of Foreign Affairs, Sri Lanka

The Information and Communication Technology Agency (ICTA) has taken another step towards creating a digitally-empowered nation by implementing an electronic Document Attesting System (eDAS) at the Ministry of Foreign Affairs (MFA).

Source: <http://www.ft.lk/article/596399/ICTA-implements-e-Document-Attesting-System-at-Ministry-of-Foreign-Affairs> (February 2017)

Motivation



Processing time

12 hours



15 minutes

Error Rates

40%



5%

Source: *Accepting E-Documents with E-Signatures*,
VERITE Research (February 2017)

TABLE 1: TIME TAKEN (IN HOURS) TO COMPLY WITH DOCUMENTATION REQUIREMENTS FOR INTERNATIONAL TRADE

	To Export	To Import
Singapore	4	1
UAE	6	37
Malaysia	10	10
Oman	31	24
India	61	67
Pakistan	62	153
Sri Lanka	76	58

Source: *World Bank, Doing Business Index 2016*

Source: *The Global Enabling Trade Report 2016*,
World Economic Forum



The Decentralized Protocol for Document Verification built on Digital Signatures

<http://www.idstack.one>

Literature Review

Identity :	Something you have	– The availability of a physical object in possession of the person. (Eg. Key)
	Something you know	– A predefined fact or knowledge that is known by the person. (Eg. Password)
	Something you are	– Biometrics or measurable personal traits. (Eg. Fingerprint)

An identity document is a piece of documentation that is specifically designed to prove the identity of an individual. It belongs to the first of the three categories introduced by Miller.

Name is an important identity attribute. **Person name disambiguation** has been an interesting research topic throughout history.

- *Lisbach* : Linguistic identity matching
- *Fleischman and Hovy* : Maximum entropy model, probabilistic measures and agglomerative clustering techniques
- *Niu, Li and Srihari* : Supervised technique to analyze frequency of name in multiple documents
- *Torvik, Weeber, Swanson and Smalheiser* : Person name disambiguation in medical domain
- *Euzenat and Valtchev* : Web ontology language and pair-wise similarity matching
- *Nagaraj and Thiagarasu* : Ridge regression and Eigen values to measure similarity

- *B. Miller, "Vital signs of identity [biometrics],"*
- *B. Lisbach and V. Meyer, Linguistic identity matching. Springer, 2013.*
- *M. Ben Fleischman and E. Hovy, "Multi-Document Person Name Resolution,"*
- *C. Niu, W. Li, and R. K. Srihari, "Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction,"*
- *V. I. Torvik, M. Weeber, D. R. Swanson, and N. R. Smalheiser, "A probabilistic similarity metric for medline records: A model for author name disambiguation,"*
- *J. Euzenat and P. Valtchev, "Similarity-based ontology alignment in OWL-Lite,"*
- *R. Nagaraj and V. Thiagarasu, "Correlation similarity measure based document clustering with directed ridge regression," Indian J. Sci. Technol.*

Literature Review Contd.

When measuring the name similarity, the translation between languages and the differences in the alphabets of the languages cause variations in representations of the names.

Similarity in names is measured by:

- **Phonetic similarity:** Traditional Soundex implementations
Improved Soundex implementations
Phonetic distance measuring techniques (acoustic vowel distances)
- **Order of name segments:** Names are represented in different ways (Eg. <FirstName> <LastName>, <FamilyName><FirstName>)
- **Textual similarity:** Algorithms to calculate textual similarities (Eg. Levenshtein distance, Jaccard similarity, Sorensen Dice)

In addition to the name, other attributes too contribute to the representation of the identity of an individual with regard to identity documents.

- *"Soundex System", National Archives, 2017. [Online]. Available: <https://www.archives.gov/research/census/soundex.html>*
- *D. Holmes and M. C. McCabe, "Improving precision and recall for Soundex retrieval,"*
- *H. Raghavan and J. Allan, "Using soundex codes for indexing names in ASR documents,"*
- *M. Wieling, E. Margaretha, and J. Nerbonne, "Inducing a measure of phonetic similarity from pronunciation variation,"*
- *G. Kondrak, "Phonetic alignment and similarity,"*
- *W. Heeringa, "Measuring dialect pronunciation differences using Levenshtein distance,"*
- *S. Banerjee and T. Pedersen, "The design, implementation, and use of the Ngram statistics package,"*
- *P. Achananuparp, X. Hu, and X. Shen, "The evaluation of sentence similarity measures,"*
- *W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string metrics for matching names and records,"*
- *A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity,"*

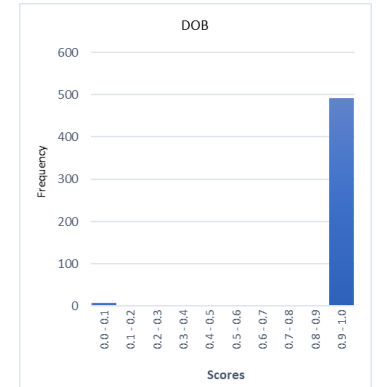
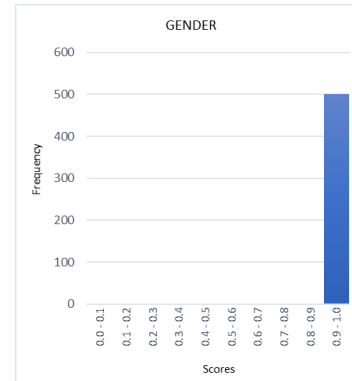
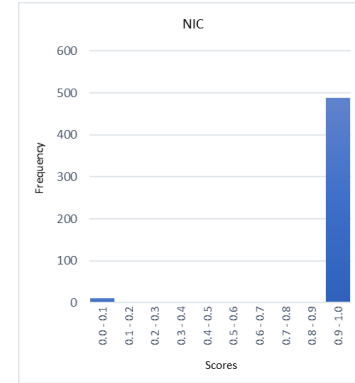
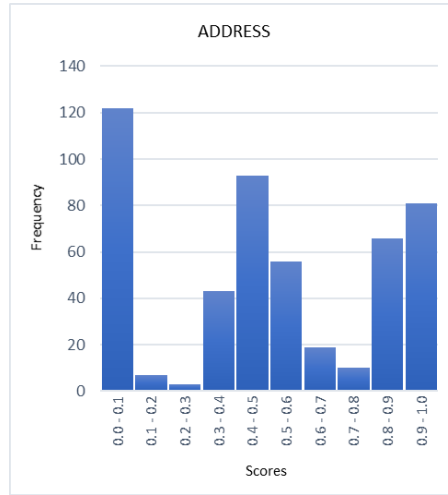
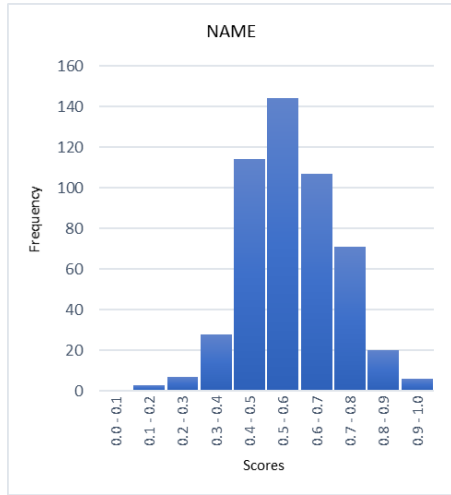
Correlation Score

Person-Identity based attributes can be divided into 4 categories				
	Space-related attributes	Time-related attributes:	Classifying attributes:	Identification codes:
Example	Address, Country, Region	Date of Birth, Date of Issue, Exp. Date	Gender, Marital Status	Passport No, NIC No, Social Security No
Possibility to be updated	High	Very Low	Low	Very Low
For research scope covering Sri Lankan context we selected 4 attributes	Address	Date of Birth	Gender	NIC

- *B. Lisbach and V. Meyer, Linguistic identity matching. Springer, 2013.*
- *J. W. M. Campbell, "The role of biometrics in ID document issuance," Keesing's Journal of Documents & Identity, no. 4, pp. 6–8, 2004.*

Correlation Score : Results

Distribution attribute scores of data collected from the survey



Correlation Score Contd.

Super Attributes:

1. Name : **Partial match**
2. Address : **Partial match**
3. DOB : Exact match
4. Gender : Exact match
5. NIC : Exact match

- *D. Dessimoz and P. C. Champod, "Multimodal Biometrics for Identity Documents 1 State-of-the-Art Research Report," Most, no. September, 2005.*
- *R. Clarke, "Roger Clarke's 'Id and Authentication Basics'", Rogerclarke.com, 2017. [Online]. Available: <http://www.rogerclarke.com/DV/IdAuthFundas.html>. [Accessed: 20- Aug- 2017].*
- *B. Miller, "Vital signs of identity [biometrics]," IEEE Spectr., vol. 31, no. 2, pp. 22–30, 1994.*

Correlation Score : Name

Name is an important identity measure.

There are many researches conducted for person name disambiguation.

Supervised techniques



Data collection

Real time analysis



Correlation of given documents

- M. Ben Fleischman and E. Hovy, "Multi-Document Person Name Resolution," *ACL 2004 Work. Ref. Resolut. its Appl.*, pp. 1–8, 2004.
- C. Niu, W. Li, and R. K. Srihari, "Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction," *Proc. 42nd Meet. Assoc. Comput. Linguist. (ACL'04), Main Vol.*, pp. 597–604, 2004.
- V. I. Torvik, M. Weeber, D. R. Swanson, and N. R. Smalheiser, "A probabilistic similarity metric for medline records: A model for author name disambiguation," *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, no. 2, pp. 140–158, 2005.

Correlation Score : Name Contd.

Problems:

1. Sri Lankan names are long and repetitive:
 - existing name-similarity measuring methodologies can give conflicting results
2. Mostly in Sinhala or Tamil:
 - translation to English can have variations
3. Different documents represent the name with **different attributes** and **attribute order**
4. Non-linguistic typing mistakes
5. People can change their names

Solution:

An algorithm that calculates:

- phonetic similarity
- order of name segments
- string similarity

- *H. Raghavan and J. Allan, "Using soundex codes for indexing names in ASR documents," Proc. Work. Interdiscip. Approaches to Speech Index. Retr. HLT-NAACL 2004., pp. 22–27, 2004.*
- *M. Wieling, E. Margaretha, and J. Nerbonne, "Inducing a measure of phonetic similarity from pronunciation variation," J. Phon., vol. 40, no. 2, pp. 307–314, 2012.*
- *G. Kondrak, "Phonetic alignment and similarity," Comput. Hum., vol. 37, no. 3, pp. 273–291, 2003.*

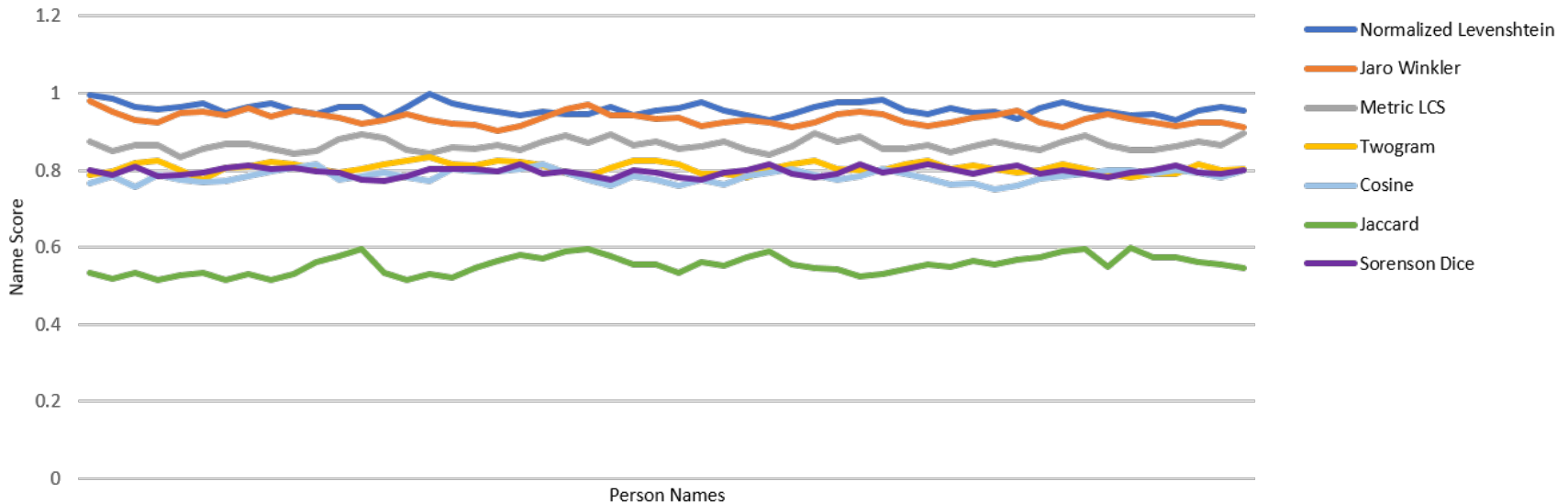
Correlation Score : Name Contd.

	Document1	Document2
Example	Ranasinghe Arachchilage Kasun Dhanushka Gayanath Ranasinghe	Kevin Gayanath Ranasinha
Phonetic representation	R525 A622 K420 D522 G530 R525	K870 G530 R525
Identify overlapping name segments	R525 A622 K420 D522 G530 R525	K870 G530 R525

$$\text{Pair-wise order Score: } OS_{d_i, d_j, \text{name}} = \frac{\text{Overlapping segments in the same order}}{\text{Unique name segments in } d_i \text{ and } d_j}$$

- G. Kondrak, "N -Gram Similarity and Distance," *Lect. Notes Comput. Sci.*, vol. 3772, pp. 115–126, 2005.
- S. Banerjee and T. Pedersen, "The design, implementation, and use of the Ngram statistics package," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2588, pp. 370–381, 2003.

Correlation Score : Name Contd.



- *W. Heeringa, "Measuring dialect pronunciation differences using Levenshtein distance," Dissertations.Ub.Rug.NI, 2004.*

Correlation Score : Name Contd.

Unique phonetic IDs	Document 1	Document 2	String similarity (Normalized Levenshtein)
R525	Ranasinghe	Ranasinha	80%
A622	Arachchilage	-	
K420	Kasun	-	
D522	Dhanushka	-	
G530	Gayanath	Gayanath	100%
K870	-	Kevin	

Pair-wise String similarity score:
$$SS_{d_i, d_j, name} = \frac{\sum \text{Levenshtein similarity}}{\text{Unique name segments in } d_i \text{ and } d_j}$$

Correlation Score : Name Contd.

	D1	D2	D3	D4
D1	-	a	b	d
D2	a	-	c	e
D3	b	c	-	f
D4	d	e	f	-

$$\text{Total pair-wise score: } CS_{d_i,d_j,name} = \frac{OS_{d_i,d_j,name} + SS_{d_i,d_j,name}}{2}$$

a = Score between D1 & D2

b = Score between D1 & D3

c = Score between D2 & D3 etc.

$$\text{Total document score: } CS_{d_k,name} = \left(\frac{1}{n-1}\right) \sum_{\substack{i=0 \\ i \neq k}}^n CS_{d_k,d_i,name}$$

Final name correlation score of D1 = (a+b+d) / 3

Correlation Score : Algorithms

For partially matching attribute, A (Name, Address)

$$\text{Total document score for attribute } A: \quad CS_{d_k, A} = \left(\frac{1}{n-1} \right) \sum_{\substack{i=0 \\ i \neq k}}^n CS_{d_k, d_i, A}$$

For exactly matching attribute, A (DOB, Gender, NIC)





















$$\text{Total document score for attribute } A: \quad CS_{d_k, A} = \begin{cases} 1, & A_{d_k} = \text{Candidate}_A \\ 0, & \text{otherwise} \end{cases}$$

Final correlation score of document d_k

$$\text{Total score of document } d_k: \quad CS_{d_k} = \frac{1}{n_A} \sum_i^{n_A} CS_{d_k, A}$$

Correlation Score : Demo

~ CORRELATION DOCUMENT SCORE Correlation Score for multiple Documents

Attribute	Avg score	Document 1	Document 2	Document 3
Name	 27.22%	Dangalla Appuhamilage Dona Jenevi Sachithra Dangalla 	DANGALLA 	JENEVI SACHITHRA DANGALLA DANGALLA APPUHAMILAGE DONA 
Address	 0%	- 	KOSWATTA, NAWALA 	- 
Date of Birth	 66.67%	1993-09-30 	- 	1993-09-30 
Gender	 66.67%	Female 	- 	F 
NIC	 0%	- 	937741170V 	932741170V 

REPLY



The Decentralized Protocol for Document Verification built on Digital Signatures

<http://www.idstack.one>

Thank You !

