Propuesta de Algoritmo de Reconocimiento de Nombres de Entidad para Integrar Bases de Datos Gubernamentales

Investigador Egresado/ Maestrante: Ing. Juan C. Delgado
Investigador Tutor/Director: Ing. Juan Fernando Galárraga, Msc
Universidad de las Fuerzas Armadas ESPE, Ecuador
Latacunga, Ecuador

juancarlos_delgado@inec.gob.ec;
jfgalarraga@espe.edu.ec

Noviembre, 2016

Spanish Version, 2016

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Contenido

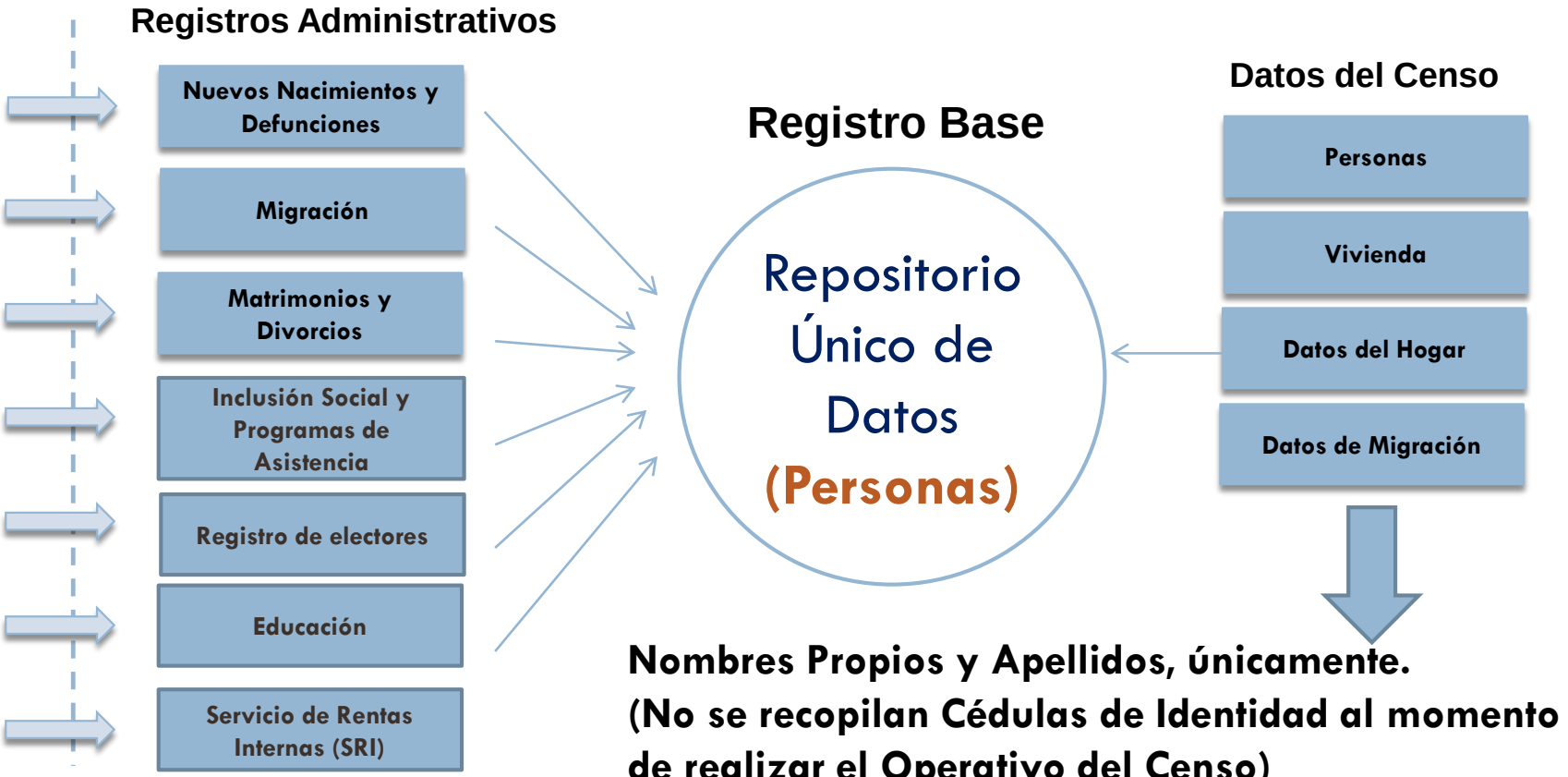
- Motivación y Objetivos
- Antecedentes
- Diseño de la Investigación
- Evaluación de Resultados
- Conclusiones y Trabajo Futuro

Motivación y Objetivos (1)

- ***Reforzar el emparejamiento de datos para el censo Basado en Registros.***
  - *La Mayoría de la información sobre la población ha sido recopilada desde Bases de Datos Gubernamentales no conectadas entre sí.*
  - *No hay un código numérico único para identificar a los ciudadanos.*
  - *No hay acceso a la información completa de un ciudadano.*
  - *No existe un repositorio común de ciudadanos.*

Motivación y Objetivos (2)

Integración de Bases de Datos Gubernamentales



Nombres Propios y Apellidos, únicamente.
(No se recopilan Cédulas de Identidad al momento de realizar el Operativo del Censo)

Motivación y Objetivos (3)

- ***Los nombres de personas se escriben de diferentes modos.***
  - *Hay variantes fonéticas en la pronunciación*
  - *El reconocimiento de caracteres a través de un dispositivo de digitalización o scanner es todavía incipiente*
  - *Hay palabras mal deletreadas y falta completitud en los datos.*
  - *Se utilizan abreviaciones.*

Motivación y Objetivos (3)

El objetivo principal de esta investigación es cómo generalizar y reducir el problema de reconocimiento de nombres de personas para llevarlo a su forma más simple, por medio de un algoritmo de Procesamiento de Lenguaje Natural aplicado capaz de generar los casos de emparejamiento más probables en la comparación de datos de nombres de ciudadanos.

- *En este proceso, el algoritmo que se construya debe poseer las estrategias más relevantes en las dos categorías de técnicas de Reconocimiento de Nombres de Entidades que existen, para lograr exactitud y fiabilidad de resultados al realizar los emparejamientos de nombres. : 1) los algoritmos fonéticos y 2) los algoritmos de edición de distancia a través de la comparación de palabras.*

Antecedentes (1)

- Trabajos Relacionados

- El aspecto más relevante en los trabajos previos de investigación es el hecho de que:

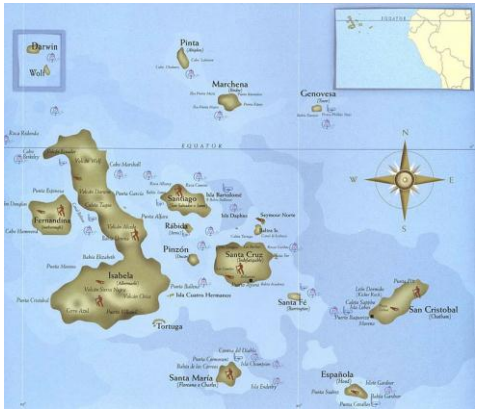
- (…) no hay una técnica única y exclusiva que pueda resolver por si sola todas las tareas de emparejamiento de nombres (…) especialmente cuando se trata de reconocer o buscar nombres de personas.

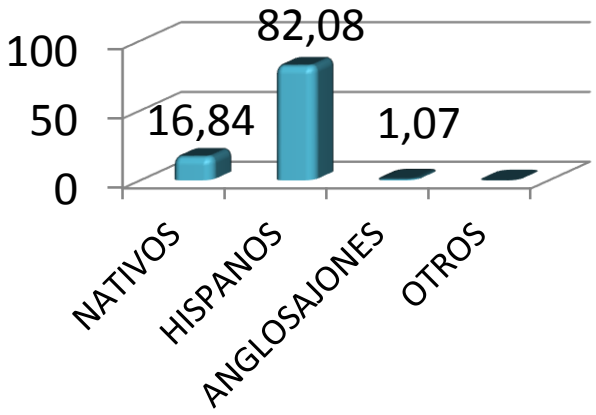
Antecedentes (2)

- Evaluación Experimental
- Técnicas RNE para Emparejamiento
  - Evaluación
  - Nombres de las Muestras de Datos
  - Emparejamiento por Comparación Fonética
  - Emparejamiento por Comparación de Deletreo y Distancia

Antecedentes (3)

Origen de nombres de Ciudadanos de Galápagos



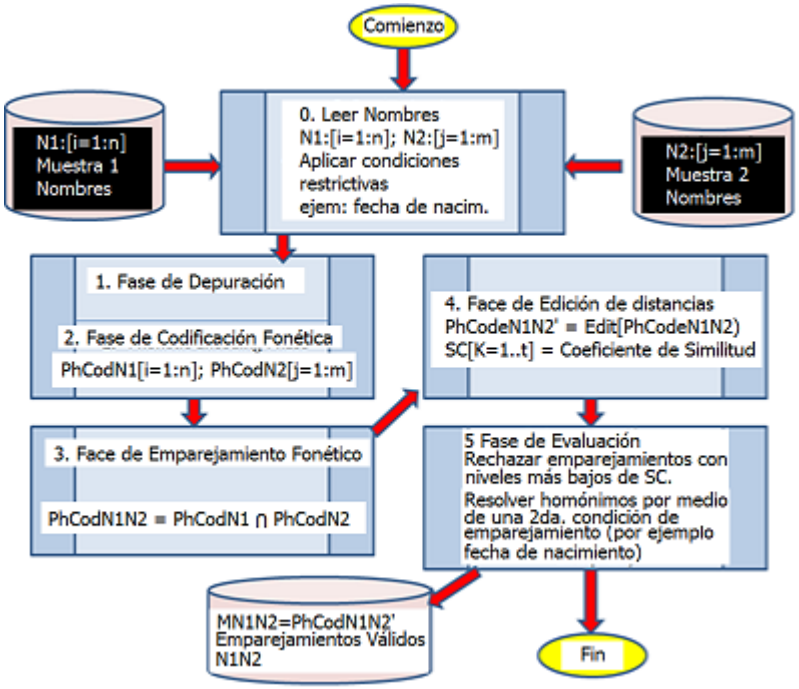


Población por origen de apellidos.

- Análisis Exploratorio de Datos
  - La Comparación de errores fonéticos en los códigos de emparejamiento muestran que, en su gran mayoría, estos dependen de las características intrínsecas, tanto en los nombres propios como en los apellidos, tales como la pronunciación local de las palabras.
  - La Comparación fonética en el emparejamiento muestra que el método Soundex tiene un nivel más alto de emparejamientos que otras técnicas fonéticas tanto para nombres propios como para apellidos de la muestra.
  - La Comparación por deletreo y cálculo de distancia de edición en el emparejamiento muestra que la distancia de Levenshtein tiene un nivel más alto de emparejamientos cuando hay muchas variantes en los nombres a reconocer en los datos de la muestra. En estos algoritmos no interesa el origen fonético de los nombres.

Diseño de la Investigación

- Construcción del Algoritmo propuesto
  - La fase de depuración de nombres
  - La fase de codificación fonética
  - La fase de emparejamiento fonético
  - La fase de edición de distancias
  - La fase de evaluación



Evaluación de Resultados(1)

- Muestras de datos experimentales.
  - M1: turistas extraídos de visitas anuales a Galápagos
  - M2: ciudadanos residentes extruidos del Censo de Pob.
- Evaluación del Algoritmo RNE.

Factor F1 (factor de exactitud): $$F1 = \frac{(2*P*R)}{P+R}$$ Dónde, P= Precision y R = Relevancia	Factor C obtenido por la Ecuación $$C = \frac{(M*0.5+PF*0.5)}{100}$$ Dónde, M = madurez, PF = Prevención de Fallos

Evaluación de Resultados(2)

- Muestras de Datos experimentales.
  - M1: turistas extraídos de visitas anuales a Galápagos
  - M2: ciudadanos residentes tomados de Población
- Evaluación del Algoritmo RNE

PRE – POST Prueba de Muestras Relacionadas	**Correlación**
F1: Factor de Exactitud	Sig < 0.05
C: Factor de Confiabilidad	Sig < 0.05

Conclusiones

- La variación fonética en los nombres depende del origen etimológico de estos hasta cierto punto, considerando la pronunciación local de las palabras y sus variantes al deletrearlas.
- La integración de las bases de datos gubernamentales requiere de un algoritmo de emparejamiento óptimo. Tanto la precisión como la velocidad tienen que ser reforzadas por la aplicación de una distancia de edición y un código fonético.
- El valor de Sig en la prueba T para el factor (F1) muestra que el cambio del factor después de aplicar el algoritmo es suficiente para estimar que no se debe al azar. También, un ligero incremento que se ha notado en el factor de Confiabilidad (C) refuerza esta estimación.

Trabajo Futuro

- Evaluar el rendimiento del algoritmo con muestras tomadas de otra población representativa en el mundo.
- Reforzar la fase de codificación fonética del algoritmo, considerando similitudes de sonidos en los apellidos nativos para obtener mejores resultados cundo se trate de integrar registros de población indígena.
- Reforzar el proceso para obtener códigos fonéticos tomando en consideración la pronunciación local de los nombres en otras lenguas.
- Reforzar la exactitud de la fase de edición de distancias con estrategias de otros algoritmos similares.

Referencias Relevantes

- [4] Division, U. N. (2015, January). United Nations Fundamental Principles of Offficial Statistics. Implementation guidelines. New York, United States of America (2015): Revised on January 22, 2015
- [12] Peng T., Lin L, Kennedy J., A Comparison of Techniques for Name Matching. Pages 1-7., Edinburg, UK 2010.
- [15] Mohd Anuar, Fatahiyah, Rossitza Setchi, and Yu-Kun Lai. "Trademark retrieval based on phonetic similarity." Systems, Man and Cybernetics (SMC), IEEE International Conference on. IEEE, 2014.
- [20] Nelson, P. (2015). The Unique Challenge of Searching for Names. How to turn a Generic Search Engine into a World Class Directory Search. San Diego, CA, USA: Search Technologies Corp. (2015).

Propuesta de Algoritmo de Reconocimiento de Nombres de Entidad para Integrar Bases de Datos Gubernamentales

Investigador Egresado/ Maestrante: Ing. Juan C. Delgado
Investigador Tutor/Director: Ing. Juan Fernando Galárraga, Msc
Universidad de las Fuerzas Armadas ESPE, Ecuador
Latacunga, Ecuador

Universidad de las Fuerzas Armadas ESPE, Ecuador,
Departamento de Ciencias de la Computación, Sangolquí, Ecuador
juancarlos_delgado@inec.gob.ec;
jfgalarraga @espe.edu.ec

Noviembre, 2016

Spanish Version, 2016