Analítika

Hacia un Algoritmo Optimo de Emparejamiento de Nombres

Towards an Optimum Name Matching Algorithm

Juan Carlos Delgado Loyola

www.ecuadorencifras.gob.ec

Hacia un Algoritmo Óptimo de Emparejamiento de Nombres[1]
Towards an Optimum Name Matching Algorithm

Juan Carlos Delgado Loyola[2]

Dirección de Registros Administrativos. Instituto Nacional de Estadística y Censos

Quito - Ecuador

Resumen[3]

La gran variedad de algoritmos de emparejamiento de nombres no son suficientes por si mismos para proveer de una juntura óptima entre bases de datos administrativas pertenecientes a los mismos ciudadanos. De acuerdo con otras comparaciones experimentales de técnicas de emparejamiento, las variantes en la composición de nombres de personas tienen un rol importante en el proceso de emparejamiento, así que se ha hecho primero un análisis previo de lo mismo. El presente estudio realiza una evaluación de un algoritmo de emparejamiento con una muestra representativa de nombres de personas tomada de la población del Ecuador en el año 2010. Esta muestra se empareja con una similar que contiene nombres de ciudadanos de la base de datos del Registro Civil; y se asume que estuvieron presentes durante el día del censo. Se incluye un modelo de procesos para el nuevo algoritmo que combina técnicas fonéticas y de distancia de edición. Finalmente, se obtiene evidencia estadística a través de la diferencia significativa en el factor de exactitud para valores antes y después de la ejecución del algoritmo.

Palabras Clave: emparejamiento de nombres,técnicas fonéticas, distancia de edición.

JCL: 1: Fonética/Fonología; 2: Morfología[4]

[1]Primero quiero agradecer a todas las personas que asistieron a Diálogo Estadístico en INEC y a todos aquellos que tuvieron la oportunidad de conocer en detalle este trabajo, y alentaron a su publicación, especialmente a Boris Espinoza, Cesar Vicuña, Stalyn Flores, Yandré Jaime, Lorena Moreno y Karla Pasquel. En segundo lugar me es grato reconocer a Juan Fernando Galarraga, como colaborador docente en mis estudios relacionados con estas investigaciones y a su aporte profesional en la revisión de este trabajo.

[2]juancarlos_delgado@inec.gob.ec

[3]El presente estudio utiliza datos individuales con fines investigativos y estadísticos, por lo tanto no vulnera el principio de confidencialidad y reserva de la Información establecido en el artículo 21 de la Ley de Estadística, el artículo 6 de la Ley Orgánica de Transparencia y Acceso a la Información Pública y demás normas conexas.

[4]Journal Computacional Linguistic (JCL): Niveles 1: Fonética/Fonología; 2: Morfología

1. Introducción

Las técnicas de emparejamiento de nombres tienen su fundamento teórico en el Procesamiento del Lenguaje Natural (PLN) y el reconocimiento de entidades (RE) a partir de textos escritos en cualquier idioma y generalmente las utilizan los algoritmos de búsqueda. Aquellas técnicas que tienen el propósito de emparejar nombres provenientes de diferentes fuentes se las ha denominado aquí:Técnicas de Reconocimiento de Nombres de Entidad (RNE). Los trabajos de Reynar (1998), y Huang et al. (2007) son útiles en este análisis.

La aplicación sistemática de las técnicas de emparejamiento de nombres en grandes bases de datos ha hecho posible algunas aplicaciones como: el cobro efectivo de impuestos, la ubicación de historias clínicas, la verificación de datos para chequeo de visas, el seguimiento a refugiados y personas sospechosas de terrorismo, la identificación de clientes potenciales, el censo basado en registros, etc. Dichos estudios han sido tratados por Hermansen (2006). Son útiles también las aplicaciones y técnicas RNE tratadas en Schay (2011). Finalmente, hay estudios de estadísticas basadas en registros administrativos que sugieren estos tipos de emparejamiento como los nombrados por Wallgren (2012).

Con la finalidad de obtener un registro único de ciudadanos partiendo de la información del último censo de población y vivienda de Ecuador y los registros de cedulados del Registro Civil, se ha propuesto un algoritmo optimizado de emparejamiento de nombres de personas, pues no se cuenta con cédulas de identidad en el censo. Este artículo trata sobre la evaluación de dicho algoritmo y se ha organizado de la siguiente forma: Una referencia a las técnicas RNE se presenta en el capítulo II. En el capítulo III se presenta los resultados de la evaluación del algoritmo RNE a través de la medición de su factor de exactitud. La confirmación de validez estadística de resultados propone el rechazo de la hipótesis nula, relacionada con la uniformidad de eventos antes y después del experimento, con una certeza de significación del 95 %.

2. Marco conceptual

2.1. Técnicas de emparejamiento

Establecer una comparación aproximada entre dos textos diferentes que tienen diferentes orígenes de datos es una tarea compleja que requiere más de una técnica. Para el proceso de emparejamiento de nombres se consideran tres grupos de técnicas: 1) fonéticas; 2) de deletreo y distancia; y 3) combinadas. Las técnicas fonéticas establecen comparaciones de palabras por similitud en la percepción de sonidos, cuando hay más de una forma escrita para representar el mismo nombre. Estas técnicas asignan códigos a cada secuencia de caracteres basados en el sonido que estos producen. El emparejamiento se hace entre las formas canónicas de los nombres. Las técnicas de deletreo y distancia, por su parte, generan un valor máximo de similitud equivalente al valor mínimo de distancia que resulta de operaciones de inserción,

borrado o sustitución de caracteres hechas para equiparar dos palabras; esta técnica se conoce como Distancia de Leveinshtein-Damerau. Otras técnicas de deletreo obtienen el valor de similitud mediante otros métodos como el relacionado con el reconocimiento de patrones de texto y la división de palabras en sub-unidades de N caracteres (N-grams). Estos métodos no necesitan de ninguna transformación fonética. Finalmente, todo lo que se fusione a través de métodos fonéticos con métodos de distancia de edición se conoce como técnicas combinadas. La más conocida es Editex. Esta última introduce verificación de sonidos iguales en las operaciones de distancia de la Técnica de Levenshtein y Damerau para descartar errores en la medición debido a la presencia de similitud fonética en los nombres.

Lo relevante de las técnicas de emparejamiento fonético son las adaptaciones que de ellas se han hecho a lo largo de los años para obtener un mayor número de similitudes aproximadas de nombres. Lo que comenzó con un simple algoritmo de indexación de apellidos en lenguas anglosajonas, mediante reducción de su representación escrita a 6 dígitos (Soundex), ha dado origen a una serie de algoritmos similares con reglas adicionales de representación adaptados a nombres en lenguas europeas (Metaphone, Phonex, NYSIIS), hindúes y ahora también, asiáticas. De interés experimental para el presente estudio son las dos nuevas adaptaciones al código Soundex para la lengua castellana. La primera es propuesta por Fernandez L. (2010) conocida como Soundex-SP y contempla reglas de indexación para las letras Y, LL, y CH. La segunda es de Mazariegos O. (2012) y va más allá al introducir reglas adaptadas a la pronunciación centroamericana como la asignación del dígito 7 a las letras Q y J y la reducción del sonido de la 'CH' a 'V' y el de la 'LL' a 'J'.

La similitud de dos cadenas de caracteres es determinada por el valor de retorno de la función que calcula la distancia mínima de edición entre las cadenas de caracteres s y t, denominada distancia de Leveinshtein (distld). La ecuación 1 muestra el cálculo de esta distancia. El valor mínimo se obtiene de la sumatoria de transformaciones entre todas las combinaciones posibles entre las posiciones de la cadena s y la cadena t. En la sumatoria; x, y son los valores absolutos para las operaciones de inserción, borrado y sustitución de caracteres. W_i es un valor de peso aplicado a cada operación i.

$$distld(s,t) = \text{mín} \sum_{i=1}^{N} W_i(|x|,|y|) \tag{1}$$

Par comprender mejor este cálculo, a cada transformación de $(|x|,|y|)$ se la registra en una matriz $d(i = 1..s, j = 1..t)$, donde la posición i es de la primera cadena de caracteres (s) y la posición j es de la segunda cadena (t). Las operaciones de una transformación se expresan en la ecuación 2. El vector $c(i,j)$ tiene los valores asignados durante el proceso. El vector W_i es el peso asignado a cada operación. Damerau introdujo en la ecuación original una nueva operación: la de transposición. Esta identifica cuando un carácter ha ocupado el lugar que le correspondía al siguiente o anterior en el nombre. Por ejemplo, la transposición en los nombres 'Gabriel' y 'Grabiel' son muy comunes.

$$\begin{aligned} d(i,j) = \text{mín}\{&d(i-1,j)+1, insersion \\ &d(i,j-1)+1, borrado \\ &d(i-1,j-1)+c(i,j), sustitucion \\ &d(i-2,j-2)+c(i,j-1)+c(i-1,j)+1)\}transposicion \end{aligned} \tag{2}$$

Cabe mencionar otras técnicas de deletreo como la de Guth, en la cual se obtiene un valor de similitud por acumulación de resultados en las variables dicotómicas de cumplimiento $(1 = Si; 0 = No)$ de hasta doce reglas de similitud entre posiciones anteriores o posteriores a cada caracter en dos palabras supuestamente similares. Sin embargo sus resultados no son muy convincentes para nombres cortos.

De las técnicas de deletreo que analizan caracteres comunes en los nombres para obtener un valor máximo de similitud, las más exhaustivas en la búsqueda y, por lo general; más precisas son las de Jaro y Jaro-Winkler, analizadas por Christen (2006). El algoritmo de Jaro calcula un valor de similitud entre dos cadenas, aceptando los caracteres que están dentro de la mitad de la longitud de la cadena más larga. La similitud de Winkleres una medida mejorada a la de Jaro. La ecuación 3 muestra este valor de similitud.

$$sim_{jaro}(s_1, s_2) = \frac{1}{3}\left(\frac{c}{|s_1|} + \frac{c}{|s_2|} + \frac{c-t}{c}\right) \tag{3}$$

Dónde $s1$ y $s2$ son las dos cadenas de caracteres a comparar, c es el número de caracteres comunes y t es el número de transposiciones:

2.2. Trabajos relacionados

Al igual que Kumar et al. (2010), este estudio propone utilizar el potencial de las técnicas fonéticas en la indexación de nombres, con el fin de reducir el gran volumen de datos iniciales a un conjunto de registros similares en pronunciación en poco tiempo de proceso. Ante la pregunta ¿Puede desarrollarse una codificación fonética adaptada al origen etimológico de los nombres?, se encontró que existen ciertos estudios de lenguajes latinos, los cuales proveen una descripción de procesos fonológicos en la creación de nombres personales y cómo los cambios lingüísticos y culturales afectan a dichos nombres. En trabajos futuros pueden ser de utilidad los estudios de Fall and Giraud-Carrier (2005), pues construyen chequeadores de deletreo fonético para mejorar en la precisión del emparejamiento fonético. También se puede encontrar algo similar en los experimentos de Mendoza and Zamudio (2005), Christen (2006) y Tibón (2005).

El emparejamiento de nombres a través de los algoritmos de distancia de edición ha tenido una amplia aceptación en la comunidad científica que se dedica a la fusión probabilística de registros. Tomando en cuenta el trabajo iniciado por Cohen et al. (2001), dónde las

en Registro Civil y que se presume estuvieron presentes en el día del censo, es decir el 28 de noviembre de 2010. Las muestras se prepararon para emparejar los nombres de la base de empadronados con los nombres de la base de cedulados utilizando el algoritmo RNE, sin que intervenga otro campo adicional ni tampoco las cédulas, únicamente los nombres de ciudadanos.

3.2. Construcción del algoritmo RNE

El algoritmo RNE se construyó para aplicarlo en un caso de estudio destinado a evaluar similitudes de a) nombres propios y b) apellidos a la vez entre los registros N1 (1...n) de la base de empadronados (CPV2010) y los registros de la muestra N2 (1...m) de cedulados (RCIVIL). En este caso n = m. El objetivo de la implementación fue fusionar registros por nombres de las personas, por lo cual el algoritmo realizaría un promedio general de los valores probabilísticos de similitud obtenidos en (a) y en (b). Antes de la utilización de los resultados en el análisis se descartaron casos de homónimos, es decir, cuando para un mismo registro de N1 (1...n) le corresponde más de un registro similar en N2 (1...m).

Para la construcción del algoritmo fue necesario realizar un experimento preliminar para comparar e identificar las técnicas de emparejamiento a aplicar acordes con el caso de estudio citado. Una muestra de 2,375 nombres propios, y otra de 2,253 apellidos principales, cada uno con al menos diez variantes en su escritura y pronunciación en el país, se seleccionaron entre los más frecuentes para comparar dichas técnicas.

En cuanto a técnicas de emparejamiento fonético, los resultados de este primer experimento indicaron que un alto porcentaje (82.08 %) de ciudadanos con apellidos de origen hispano en el país, y un porcentaje similar (70.03 %) de estas personas con nombres propios también de origen hispano, influían significativamente en la cantidad de coincidencias detectadas mediante una u otra técnica fonética. Por consiguiente, se tomó como pivote a la técnica Soundex-SP para comparar el número de coincidencias aproximadas que eran capaces de reconocer las otras técnicas. Se obtuvo que la técnica Soundex (71.97 % de casos) era la que mayor se acercaba al número de casos de nombres propios detectados por la técnica Soundex-SP, seguida de las otras técnicas: Phonex (78.38 %), Metaphone (41.77 %) y NYSIIS (21.83 %). En forma similar se compararon los números de casos de similitud aproximada de apellidos entre la técnica Soundex-SP y las otras técnicas, encontrando la misma distribución con similares porcentajes: Soundex (77.36 %), seguida de Phonex (71.42 %), Metaphone (28.82 %) y NSIIS (18,09 %).

En cuanto a técnicas de emparejamiento de deletreo y distancia, para el mismo experimento se estableció un límite porcentual de similitud aproximada de >= 95 % tanto para nombres propios como apellidos. Dicho porcentaje corresponde a valores mínimos de distancia de edición entre 1 y 2 puntos entre las palabras comparadas y para técnicas de deletreo corresponde alos valores máximos de su coeficiente de proximidad superior al dicho límite porcentual. De todas estas técnicas, con la de Editexse obtuvo un rango mayor de casos de apellidos que su-

peraron el límite (45.10 %) seguida de Levenshtein (43.40), Levenshtein-Damerau (43.16 %), Guth (13.71 %), de N-gram (3.70 %), de Jaro (4.65 %) y la de Jaro-Winkler (5.31 %). Para casos de nombrespropios, se obtuvo una distribución de resultados similar: Editex (69.87 %), Levenshtein (66.67 %), Levenshtein-Damerau (67.25 %), Guth (24.82 %), de N-gram (2.66 %), de Jaro (6.90 %) y la de Jaro-Winkler (8.71 %).

De la comparación de tiempos de procesamiento entre técnicas se encontró que las técnicas fonéticas son relativamente mucho más rápidas que las técnicas de deletreo y distancia.Tomando en cuenta que la indexación de nombres mediante códigos fonéticos se la hizo una sola vez en cuestión de pocos minutos, la operación de juntura entre nombres similares entre la muestra M1 y la muestra M2 se la hizo a su vez en cuestión de segundos, con un máximo de 80 segundos en apellidos y de 159.5 segundos en nombres propios para la técnica Metaphone, seguido de las otras técnicas Soundex, Soundex-SP, Phonex y NYSIIS, respectivamente. Sin embargo el tiempo empleado en la ejecución de técnicas de deletreo y distancia, aplicadas para emparejar las mismas muestras, fue del orden de horas de procesamiento, siendo la de Levenshtein-Damerau (6.15 horas) la que mayor tiempo obtuvo. En cuanto a la técnica combinada Editex, esta superó a todas en tiempo de ejecución: (10.38 horas).

Con los resultados de la comparación de técnicas se analizó la factibilidad de construir el algoritmo RNE en 5 fases. El flujo de dicho algoritmo se muestra en la Figura 1.

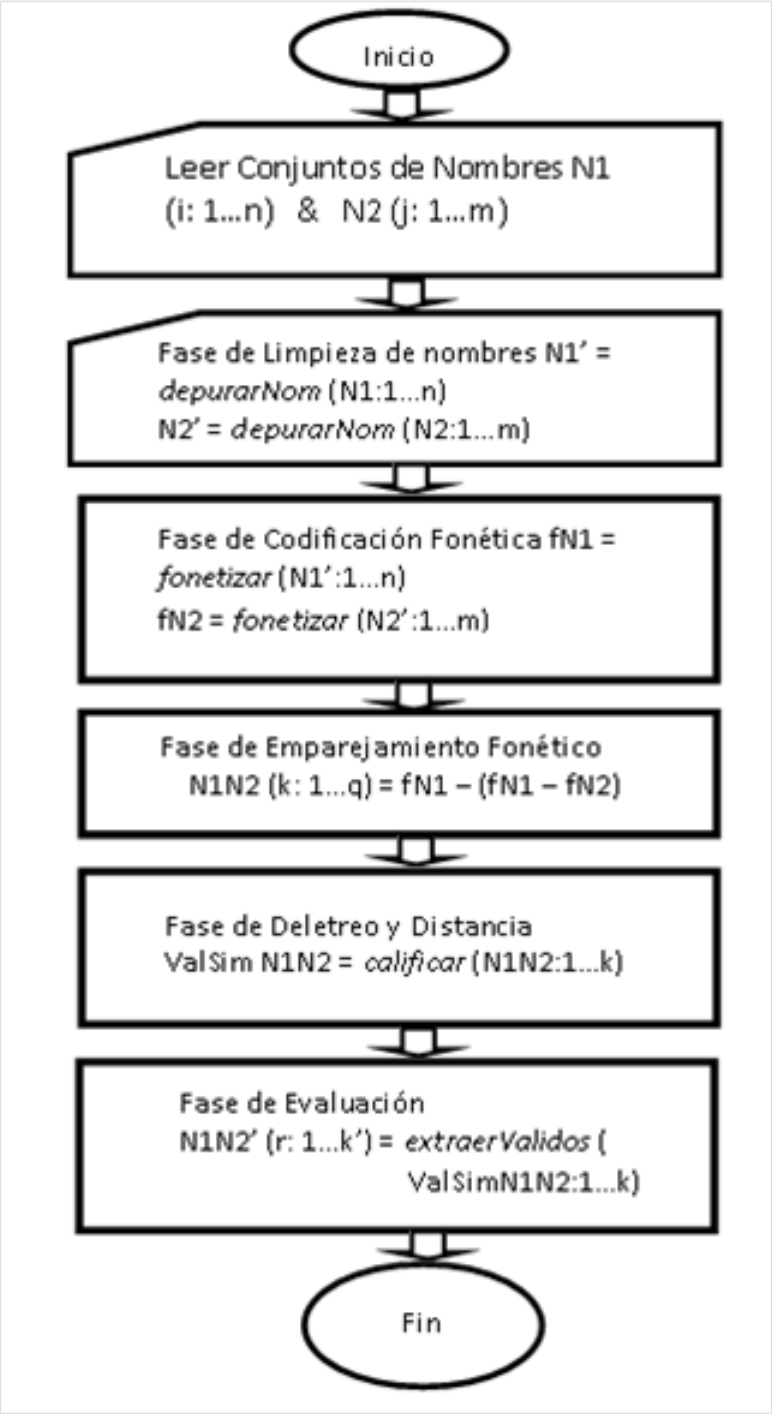

Figura 1: Algoritmo RNE

- **Fase de limpieza de nombres.**

En esta fase, se trató de que los nombres se simplificaran hacia un alfabeto estándar. En efecto se hizo una traducción de ciertos fonemas a su equivalente fonético único utilizado en la mayoría de lenguajes. Por ejemplo, para las letras que representan el sonido 'S' cuando tienen la misma pronunciación que la C o la Z se las convirtió al carácter 'S'. Además, tuvieron que ser retirados los espacios dentro del texto y se trató de evitar al máximo todos los caracteres especiales tales como "-", ".","(",")". El resultado de esta fase fue la transformación a datos estandarizados y limpios.

- **Fase de codificación fonética.**

Tanto los nombres propios como los apellidos en cada conjunto de datos debieron ser traducidos a su equivalencia de código fonético. Por consiguiente, se utilizó el algoritmo Soundex-SP por el hecho se producían más emparejamientos que con otros algoritmos. Lo que se consiguió en esta fase es un resultado preliminar de registros por similitud fonética.

- **Fase de emparejamiento fonético.**

Se aplicó una simple operación de juntura para emparejar nombres propios y apellidos codificados con la técnica fonética. Para tal efecto se dividió la muestra grande de 15,746 registros en 17 pequeñas muestras. A través de esto consiguió mayor rapidez en la ejecución del algoritmo. Para reducir el número de coincidencias entre los conjuntos de la muestra N1 ($i = 1 \ldots n$) y la muestra N2 ($j = 1 \ldots m$), donde $n = m$ se compararon los códigos fonéticos en lugar de los nombres originales y no se introdujo ninguna condición adicional con otros campos de la base de datos. Una operación de diferencia entre conjuntos (N1 - (N1 - N2)) permitió simular la juntura de dos bases de datos. Lo que se consiguió en esta fase fue una gran cantidad emparejamientos aproximados por similitud fonética con un total de k registros, donde $k < (m * n)$.

- **Fase de deletreo y distancia.**

Una vez que el conjunto de datos de emparejamiento fonético fue generado por la operación de juntura en la fase previa, la similitud de cada par de nombres emparejados fonéticamente (N1 - (N1 - N2) fue calificada por la operación de distancia de Damerau-Levenshtein que bien pudo haber sido también un algoritmo de deletreo más complejo tal como Editex o Ngrams. Un coeficiente de similitud expresado en porcentaje se mostró con los resultados de distancia calculada para cada componente de nombres de persona: apellidos y nombres propios. Luego, un coeficiente promedio se calculó para todo el nombre. Si se hubiera requerido una mayor precisión con menos emparejamientos, Jaro y Jaro Winkler pudo haberse usado. Lo que se consiguió en esta fase fue una alta completitud, exactitud y precisión. Además, se

obtuvo un conjunto de registros menor del que se obtendría con el producto cartesiano entre N1 y N2 equivalente a la operación N1 + N2 de (m * n) si es que solamente se aplicara una técnica de deletreo con todos los registros N1 (i = 1... n) y N2 (j = 1... m). En lugar de esto se aplicó la distancia solo a los registros fonéticamente coincidentes.

- **Fase de evaluación.**

Los registros obtenidos en la fase previa tuvieron que ser clasificados por coeficiente de similitud en orden descendiente. Aquellos emparejamientos con el más alto coeficiente se separaron en un conjunto de datos final. El resto de valores bajos de emparejamientos fueron rechazados. Los emparejamientos duplicados con valores altos o bajos tuvieron que ser analizados para detectar la presencia de homónimos. No se obtuvieron casos de homónimos pero en caso de que se hubieren presentado, se pudo haber utilizado una condición de restricción (ej. lugar de nacimiento). Lo que se consiguió en esta fase fue un conjunto emparejamientos altamente aproximado. Al final se encontró, por cada persona, el promedio de los valores de similitud alcanzados para las cadenas de caracteres de nombres propios y de apellidos. Luego se ordenaron los registros finales en forma descendente por el valor promedio de similitud para cada persona y se descartaron los registros que tenían valores de similitud inferiores al 95 %. El criterio para elegir este valor fue experimental y se basó en una revisión histórica de casos correctos en muestras de prueba mientras se probaba el algoritmo. Los casos de sinonimia entre un registro de persona en N1 con varios idénticos de los similares en N2 pudieron haberse comparado con relación a otra variable común como el lugar de nacimiento, sin embargo para efectos de este caso de estudio, solo fue indispensable evaluar las similitudes a través de las técnicas sin introducir otro tipo de comparaciones que involucraran otros campos, es decir, se eliminó el ruido en el algoritmo.

4. Resultados de ejecución del algoritmo RNE

Para aplicar el algoritmo RNE a la búsqueda de coincidencias de personas, se dividió la muestra aleatoria N1 de 15,746 empadronados en 32 sub-muestras homogéneas de alrededor de 524 casos. A cada una de las sub-muestras se la emparejó con la muestra N2 de 15,746 cedulados. Como resultado se obtuvo un total de 9,092 registros emparejados de 4,097 coincidencias exactas y 4,995 coincidencias aproximadas. Así, la completitud de casos emparejados fue del 57.7 %.del total de N1, con un 45.0 % coincidencias exactas antes de aplicar el algoritmo y un 54 % de coincidencias aproximadas después de aplicar el algoritmo.

5. Evaluación del algoritmo RNE

Se estableció como unidad de análisis el algoritmo RNE medido a través de su factor de exactitud (F1). Se hicieron varias corridas sucesivas con las 32 sub-muestras tomadas de la

base de datos de empadronados CPV2010 emparejados con la muestra de 15,746 cedulados de Registro Civil a nivel Nacional. La hipótesis nula Ho y la hipótesis alternativa H1, antes y después de aplicar el algoritmo se describen en las ecuaciones 1 y 2, con un nivel de significación del 95 %.

$$\text{Ho: } p > 0{,}05 \rightarrow \text{F1 antes} = \text{F1 después}$$
$$\text{H1: } p \leqq 0{,}05 \rightarrow \text{F1 antes} < \text{F1 después}$$
$$\text{F1 antes} < \text{F1 después}$$

, siendo p el valor significativo de la prueba t para muestras relacionadas.
El Factor F1 se obtuvo mediante la ecuación 3:

$F1 = \frac{(2*P*R)}{P+R}$, Dónde P= precisión y R = Relevancia
Los valores alcanzados antes y después de la ejecución del algoritmo para el factor F1 en las sub muestras se presentan en la tabla 1. Además se describen los falsos positivos y falsos negativos detectados luego de una revisión manual realizada sobre los resultados.

Tabla 1: Resultados de cálculo de los Factores F1

MUESTRA	EXACTOS ANTES	APROX ANTES	APROX DESPUÉS	FACTOR DE EXACTITUD PRE	FACTOR DE EXACTITUD POST	FALSO POSITIVO PRE	FALSO NEGATIVO PRE	FALSO POSITIVO POST	FALSO NEGATIVO POST
MCN01RC	174	0	255	0.770889488	1	0	255	0	0
MCN02RC	208	0	104	0.858725762	0.9967846	0	102	2	0
MCN03RC	120	0	28	0.918238994	0.9931973	0	26	2	0
MCN04RC	121	0	166	0.775675676	1	0	166	0	0
MCN05RC	173	0	156	0.808353808	1	0	156	0	0
MCN06RC	124	0	162	0.781163435	0.9929577	0	158	4	0
MCN07RC	116	0	141	0.788643533	0.9861933	0	134	7	0
MCN08RC	81	0	145	0.759930915	0.9865471	0	139	6	0
MCN09RC	120	0	170	0.775510204	0.9913043	0	165	5	0
MCN10RC	117	0	172	0.783258595	0.9509982	0	145	27	0
MCN11RC	124	0	153	0.786647315	0.9890511	0	147	6	0
MCN12RC	114	0	2	0.991452991	1	0	2	0	0
MCN13RC	137	0	0	1	1	0	0	0	0
MCN14RC	165	0	337	0.765765766	0.9169364	0	260	77	0
MCN15RC	115	0	174	0.780923994	0.9509982	0	147	27	0
MCN16RC	120	0	166	0.779661017	0.9822064	0	156	10	0
MCN17RC	97	0	148	0.768025078	1	0	148	0	0
MCN18RC	107	0	142	0.789022298	0.960334	0	123	19	0
MCN19RC	121	0	155	0.781740371	0.9963636	0	153	2	0
MCN20RC	115	0	170	0.775568182	0.9784946	0	158	12	0
MCN21RC	113	0	163	0.777614138	0.9777778	0	151	12	0
MCN22RC	181	0	219	0.816936488	0.9010989	0	147	72	0
MCN23RC	131	0	166	0.790896159	0.9669565	0	147	19	0
MCN24RC	124	0	151	0.796850394	0.9583333	0	129	22	0
MCN25RC	118	0	147	0.787692308	0.9827255	0	138	9	0
MCN26RC	117	0	201	0.762254902	0.9888712	0	194	7	0
MCN27RC	130	0	156	0.785714286	1	0	156	0	0
MCN28RC	127	0	216	0.776623377	0.9314642	0	172	44	0
MCN29RC	141	0	154	0.79776848	0.9845095	0	145	9	0
MCN30RC	120	0	313	0.786248132	0.7557471	0	143	170	0
MCN31RC	117	0	142	0.786482335	0.9941748	0	139	3	0
2MCN32RC	109	0	21	0.930909091	0.9922481	0	19	2	0
TOTALES	4097		4995				4420	575	

Fuente: Caso de estudio emparejamiento N1 y N2

La prueba de Kolmogorov Smirnov (KS) para una muestra dio como resultado valores de significancia menores a 0.05 para las series de valores de F1 correspondiente a los emparejamientos N1 vs. N2 de la Tabla 1. De esta manera, ambas series para F1 antes y F1 después no se aproximan a la curva Normal, por lo tanto no justifica aplicar la prueba normal t de muestras relacionadas. En su lugar se aplicó la prueba no paramétrica de rango de Wisconsin para muestras relacionadas.

De acuerdo con los resultados de la Tabla 2, en los 32 casos de ejecución del algoritmo en las sub-muestras, el estadístico significativo (2 colas) es inferior a 0.05.

Tabla 2: Prueba t del signo rango de Wilcoxon

		N	Rango Medio	Suma de Rangos
POST_Factor de Exactitud − PRE_Factor de Exactitud	Rangos Negativos	1[a]	2.00	2.00
	Rangos Positivos	30[b]	16.47	494.00
	Lazos	1[c]		
	Total	32		

Test Estadísticas		
	POST_Factor de Exactitud	PRE_Factor de Exactitud
Z	-4.821[a]	
Asymp. Sig. (2-colas)	.000	

a. Basado en rangos negativos
b.Test del Signo Rango de Wilcoxon
Fuente: Resultados SPSS de Prueba t para muestras Relacionadas

6. Conclusiones

- Las técnicas fonéticas permiten reducir el número de casos de nombres coincidentes para un volumen de datos muy grande en un tiempo relativamente corto, del orden de unos cuantos segundos. Además funcionan eficientemente como métodos de indexación para búsquedas de nombres similares. Sin embargo, carecen de precisión en la valoración de la similitud entre nombres y solamente detectan semejanzas de escritura debido a una pronunciación equivocada utilizando a reglas conocidas del lenguaje.

- Las técnicas de deletreo y distancia son independientes del lenguaje en la que están escritas las palabras. Sin embargo necesitan del máximo de comparaciones posibles entre los caracteres presentes en los nombres y requieren de mucho más tiempo de procesamiento. Un caso de excepción de mejora a estas técnicas es la técnica combinada Editex, sin embargo, al introducir un peso de similitud fonética en las operaciones de la distancia de edición, multiplica considerablemente el tiempo de procesamiento y se vuelve dependiente de las reglas del lenguaje en que están escritas las palabras.

- Un algoritmo combinado de técnicas de emparejamiento como el propuesto aquí tiene un efecto óptimo para encontrar similitudes entre nombres de personas en las bases institucionales del Estado, sumando las ventajas de ambos tipos de técnicas. El análisis

del factor de exactitud de dicho algoritmo (F) demuestra que efectivamente se presenta un cambio de optimización al utilizarlo y que este no es debido al azar, ya que se tienen evidencias suficientes para rechazar la hipótesis nula Ho: Si F antes es igual a F después, con un nivel de significación del 95 %. Esto se explica por el p valor <= 0.05 ó sigma obtenido de la prueba del signo rango de Wisconsin para muestras relacionadas aplicada a los valores de F antes y después de aplicar el algoritmo. El p valor y los datos de la Tabla 2 permiten concluir que hay una diferencia significativa entre los valores antes y después del factor de exactitud F. Lo cual permite confiar en la hipótesis alternativa H1: F antes < F después en el nivel de significación del 98.5 %.

Referencias

Christen, P. (2006). A Comparison of Personal Name Matching: Techniques and Practical Issues. *Department of Computer Science, The Australian National University.*

Cohen, W., Ravikumar, P., and Fienberg, S. (2001). A Comparison of String Distance Metrics for Name-Matching Tasks. *Carnegie Mellon University, Carnegie Mellon University, Carnegie Mellon University*, pages 1– 6.

Fall, C. and Giraud-Carrier, C. (2005). Searching trademark databases for verbal similarities. *World Patent Information*, pages 135–143.

Hermansen, J. (2006). Advanced Global Name Recognition Technology - Entity Analytics Solutions. *IBM Corporation*, pages 1–13.

Huang, C.-R., Ŝimon, P., Hsieh, S.-K., and Prévot, L. (2007). Rethinking Chineseword Segmentation: Tokenization, Character Classification, or Wordbreak Identification.

Kumar, A., Rawat, S., and Garg, S. (2010). Based Search of Indian Names in Databases. *ITT Kampur, India*, pages 1–14.

Mendoza, A. and Zamudio, R. (2005). Nombres propios de procedencia latina. *AÑO VIII, No.17*, pages 153–182.

Navarro, G. (2001). Guide Tour to Approximate String Matching. *Dept. of Computer Science, University of Chile, Blanco Encalada 2120*, pages 1–68.

Nayan, A., Kiran, R., and P, S. (2002). Named Entity Recognition for Indian Languages. *Institute of Information Technology*, pages 1–103.

Peng, T., Li, L., and Kennedy, J. (2001). A Comparison of Techniques for Name Matching. *Edinburg, UK*, pages 1–7.

Reynar, J. (1998). Topic segmentation: Algorithms and applications.

Schay, W. (2011). A Generic Framework for the Matching of Similar Names. *Faculty of Engineering and the Built Environment. University of Witwatersrand*, pages 11–203.

Tibón, G. (2005). Diccionario Etimológico Comparado de Nombres Propios de Persona. *3ed.Fondo de Cultura Económica*, pages 1–248.

Wallgren, A. (2012). Estadísticas basadas en Registros Administrativos: Aprovechamiento estadístico de datos administrativos. *México: INEGI, 2012. INEC, Biblioteca, Administración Central*.