



Algoritmos fonéticos en el desarrollo de un sistema de información de marcas y signos distintivos

Celso Gonzales-Cam

Pontificia Universidad Católica del Perú. Especialidad de Bibliotecología (Perú)
celso.gonzales@pucp.edu.pe

Resumen

El presente artículo trata sobre la aplicación de algoritmos fonéticos en el desarrollo de un sistema de información de marcas y signos distintivos, y su optimización para una recuperación más efectiva de la información.

Palabras Clave

Sistemas de información; algoritmos fonéticos; Soundex

1. Algoritmos Fonéticos.

La palabra algoritmo viene del latín, *dixit algorithmus*. Fue el matemático persa [al-Iwarizmi](#), quien es considerado el padre del álgebra introdujo este término en su obra *Hisab al yabr ua al muqabala*, (ةلباقملا و ربحلا باسح). Podemos definir a los algoritmos como un listado definido, ordenado y finito que permite dar solución a un problema determinado. Los algoritmos son deterministas, pues sus resultados deben ser inequívocos e iguales. Existen diferentes tipos de algoritmos, como los matemáticos y fonéticos.

La existencia de los algoritmos fonéticos obedece a la necesidad de recuperar información que tiene una semejanza sonora; y cuya representación a través de la palabra escrita pueda diferir de su pronunciación. Uno de los más antiguos es el Soundex, creado por Robert Russell:

Hay ciertos sonidos que forman el núcleo del idioma inglés, y cuyos sonidos son inadecuadamente representados por letras del alfabeto, como un sonido puede en algunos casos ser presentado por más de una letra o combinación de letras, y una letra o combinación de letras puede ser representado por dos o más sonidos. Es por esto que muchos nombres pueden tener dos o más diferentes pronunciaciones que se encuentren en el índice alfabético, o un índice que separe nombres acorde a la secuencia de las letras contenidos en el alfabeto (Robert Russell, Patent)¹.

Soundex

Este algoritmo fue desarrollado en 1918 por Robert Russell and Margaret Odell y patentado finalmente en 1922. Una variante de este algoritmo fue llamado American Soundex, y utilizado en 1930 para el análisis retrospectivo de los censos de Estados Unidos de 1890-1920. El [National Archives and Records Administration](#), administra el uso oficial por el gobierno norteamericano.

Una de las variantes es el Daitch-Mokotoff Soundex System creado por Randy Daitch y Gary Mokotoff, de la Jewish Genealogical Society. Esta nació ante la necesidad de cubrir nombres eslavos y yiddish, pero que incluía una independencia lingüística y étnica de los términos. Los apellidos LEWINSKY-LEVINSKI (876450) o AUERBACH-OHRBACH (097500), tenían una misma correspondencia fonética.

Double Metaphone

Este algoritmo fue desarrollado por Lawrence Phillips, y publicado en *Computer Language**, Vol. 7, No. 12 (December), 1990, como parte de una clase de algoritmo llamado como *phonetic matching* o *phonetic encoding*. Este algoritmo utiliza reglas de codificación mucho más extensas que sus predecesores,

¹ http://genealogy.about.com/od/census/a/russell_index.htm

There are certain sounds which form the nucleus of the English language, and those sounds are inadequately represented merely by the letters of the alphabet, as one sound may sometimes be represented by more than one letter or combination of letters, and one letter or combination of letters may represent two or more sounds. Because of this, a great many names may have two or more different spellings which in an alphabetic index, or an index which separates names according to the sequence of their contained letters in the alphabet, necessitates their filing in widely separate places

manteniendo componentes de caracteres no latinos, y retorna una codificación primaria y secundaria para diferentes pronunciaciones de una sola palabra en idioma Ingles.

Levenshtein

Este algoritmo no es un algoritmo fonético sino analiza la distancia entre dos cadenas, pudiendo determinar si hay una semejanza dentro de estos términos. Fue desarrollado en 1965 por el científico ruso Vladimir Levenshtein. Se denomina Distancia Levenshtein al resultado de encontrar el más eficiente camino para transformar una cadena a otra, a través de mecanismos de inserción, borrado y sustitución. A menor distancia, mayor será la correspondencia entre dos cadenas de texto comparados. Este algoritmo es utilizado para corrección ortográfica, reconocimiento de voz, análisis de ADN y detección de plagio.

La distancia Levenshtein se define como el mínimo número de caracteres que se tienen que sustituir, insertar o borrar para transformar *cadena1* en *cadena2*. La complejidad del algoritmo es $O(m*n)$, donde n y m son las longitudes de *cadena1* y *cadena2* (por tanto, el rendimiento es bastante bueno si se la compara con el de la función [similar_text\(\)](#), que es $O(\max(n,m)**3)$, pero aún así se trata de una función que puede penalizar el rendimiento global del script)².

2. Aplicaciones en la Búsqueda Fonética.

Desde que se realizaron los censos de 1890-1920, se requería que la información tuviera confiabilidad en la forma como se registraron, y por cuestiones de recuperación se requería tener una aproximación exacta del término. Por este motivo fue creado el Soundex, pues era una necesidad en la recuperación de la información de los nombres que tuvieran diferentes registros. El uso de algoritmos fonéticos permitían tener una exactitud en la recuperación. La utilización de estos algoritmos se han extendido a diferentes especialidades, desde la corrección ortográfica de los programas de traducción hasta ayudas cuando se ha ingresado erróneamente el término en los buscadores como Yahoo y Google.

New York State Identification and Intelligence Systems

Un algoritmo fonético desarrollado por la New York State Division of Criminal Justice, denominados **New York State Identification and Intelligence Systems** (NYSIIS), se basaba en la reducción de los nombres a un código de 6 letras. Fue propuesto por Taft, y según una comparación, tenía un ratio de 98.72%, en comparación del Soundex con 95.99%. En 1998 the New York State Division of Criminal Justice sustituye el sistema NYSIIS por el producto NameSearch®, por medio del cual no sólo se identifican las variantes fonéticas sino las producidas por errores de transcripción, formas abreviadas, o variantes originadas por la distinta (Galvez, 2003).

² <http://us3.php.net/manual/es/function levenshtein.php>
<http://www.let.rug.nl/~kleiweg/lev/levenshtein.html>

Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual – INDECOPI

En el Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual – INDECOPI, ha desarrollado un sistema de búsqueda fonética. Esta institución tiene el objetivo de proteger todas las formas de propiedad intelectual: desde los signos distintivos y los derechos de autor hasta las patentes y la biotecnología.

The screenshot shows the INDECOPI website interface. At the top, there is a navigation bar with the text 'SISTEMAS DE INFORMACION EN LINEA' and 'RELACION DE MARCAS PUBLICADAS'. A date selection dropdown is set to '2008-03-25'. The main content area is titled 'Consulta de expedientes' and displays details for 'Expediente Nro.: 342717-2008/OSD'. The details are organized into several sections:

Expediente Nro. : 342717-2008/OSD			
Tipo de Expediente	REGISTRO		
Fecha de Presentación	2008-01-30	Hora de Presentación	16:18
Lugar de Presentación	INDECOPI LIMA		
Procedimiento	DE PARTE		
Fecha de Acumulación		Tipo de Acumulación	
Acumulado a			

Datos de la Marca			
Tipo de Solicitud	MARCA DE PRODUCTO		
Fecha de Solicitud	2008-01-30	Fecha de Registro	
Nº de Certificado		Fecha de Publicación	2008-03-25
Fecha de Vencimiento			
Tipo de Presentación	DENOMINATIVA	Ver Grafico	
Nº de Clase	10		
Denominación	LAP-BAND AP		
Producto, Servicio, Actividad	Banda gástrica laparoscópica para uso en el tratamiento de la obesidad mórbida		

Personas Jurídicas / Naturales			
Titular	RUC / LE / DNE	Domicilio Procesal	Nacionalidad
ALLERGAN, INC.			ESTADOS UNIDOS DE AMERICA
Representante Legal	RUC / LE / DNI	Domicilio Procesal	Pais, Dpto, Prov, Dist.
ESTUDIO COLMENARES S.R.L.	20258609969		PERU / LIMA / LIMA / MIRAFLORES (LIMA 18)

Figura 1. Sitio Web del INDECOPI

3. Desarrollo del Sistema de Marcas y Signos Distintivos.

Dentro de las Interfaces de búsqueda, se creyó conveniente realizar una tipo de búsqueda general y otra búsqueda fonética. Dentro de esta búsqueda se realizan a través de *Palabra Exacta*, *Prefijo de Palabra*, *Parte de la Palabra* y *Sufijo de la Palabra*. Diversas interfaces utilizan como el gobierno australiano esta variedad de casos de búsqueda, en el sistema Australian Trade Mark On-line Search System (ATMOSS)³, añadiendo la posibilidad de diferencia en imágenes. No se integró en una sola interfaz porque la búsqueda general utilizaba un esquema avanzado, y las búsquedas fonéticas no estarían integradas al sistema, en caso de una o más combinación de palabras, en una búsqueda compleja.

El sistema Webmarks fue desarrollado con la finalidad de ser un buscador de marcas y signos distintivos de un estudio de abogados que mantiene información de sus clientes, y requiere un procedimiento de detección de nombres de empresas de pronunciación similar. Los algoritmos fonéticos son de gran utilidad para alcanzar un acercamiento a los probables infractores que utilizan la semejanza fonética como competencia desleal para que el consumidor pueda confundirse con el producto verdadero.

³ http://pericles.ipaustralia.gov.au/atmoss/falcon.application_start

El sistema de búsqueda fonética realiza dos procedimientos, utilizando las funciones nativas de Soundex de lenguaje script PHP, y dentro del proceso se realiza la búsqueda fonética en cada una de las palabras del texto y en la frase completa. Dentro del sistema, el registro de marcas conocidas como Macdonald y Sony Erickson, entre otras, fue analizado utilizando términos muy similares, aunque con diferente grafía. En la prueba se plantearon el nivel de exactitud en cada uno de los términos relacionados. Dentro de la evaluación, el sistema analizaba un espectro mayor de alcance, a diferencia de sistemas de corrección ortográfica, pues las palabras asociadas, no procede de una falta de escritura, sino la intención encubierta de engañar o confundir con una marca muy semejante con la verdadera.

Conclusiones

Como concluye (Galvez, 2003) la utilización de la recuperación fonética se ve incrementado por la combinación de los métodos n-grams y edit-distances, que llevan a un aumento significativo de palabras probables.

La combinación del algoritmo Soundex y Double Metaphone incluidos dentro desarrollado para PHP muestran una mayor recuperación.

Bibliografía

Algoritmos fonéticos: Soundex . (2008). La Tecla de Escape. Disponible en:

<<http://www.latecladeescape.com/w0/content/view/77/49/>>

Gálvez, Carmen (2006) Identificación de nombres personales por medio de sistemas de codificación fonética. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação* 2 semestre 2006(22):pp. 105-116. Disponible en:

<<http://eprints.rclis.org/archive/00010804/>>

Galvez, Carmen and Moya-Anegón, Félix. (2007). Approximate Personal Name-Matching Through Finite-State Graphs. *Journal of the American Society for Information Science and Technology* 53(13). Disponible en:

<<http://eprints.rclis.org/archive/00011688/>>

Giménez Tudurí, Carmen. (1994). La oficina de marcas y diseños de Alicante. *Métodos de Información (MEI)* 1(1):pp. 38-39. Disponible en:

<<http://eprints.rclis.org/archive/00003831/>>

Levenshtein distance (2008). Algorithm implementation. Wiki Books. Disponible en:

<http://en.wikibooks.org/wiki/Algorithm_implementation/Strings/Levenshtein_distance>

Lloret Romero, Núria. (1994) La información sobre patentes y marcas a nivel nacional. *Métodos de Información (MEI)* 1(1):pp. 30-31. Diponible en:

<<http://eprints.rclis.org/archive/00003829/01/1994-01-30.pdf>>

Mokotoff, Gary. (2008). Soundexing and Genealogy. Disponible en:

<<http://www.avotaynu.com/soundex.html>>

Nelson, Adam.(2008). Implement Phonetic ("Sounds-like") Name Searches with Double Metaphone Part III: VBScript and ASP & Database Solutions. Disponible en:
<<http://www.codeproject.com/KB/asp/dmetaphone3.aspx>>

Shanker Singh, Brijesh. (2003). Search Algorithms. DRTC Workshop on Digital Libraries: Theory and Practic. March 2003. DRTC, Bangalore. Disponible en:
<https://drtc.isibang.ac.in/bitstream/1849/32/2/E_Searchalgo_brijesh.pdf>

Datos del autor

Celso Gozales-Cam

Docente de la Especialidad de Bibliotecología y Ciencias de la Información de la Pontificia Universidad Católica del Perú.

Actualmente se desempeña como Administrador Web de la Universidad del Pacífico, Perú.

celso.gonzales@pucp.edu.pe

ANEXO

The Daitch-Mokotoff Soundex Coding Chart

Letter	Alternate Spelling	Start of a name	Before a vowel	Any other situation
NC = not coded				
AI	AJ, AY	0	1	NC
AU		0	7	NC
Ą	(Polish a-ogonek)	NC	NC	6 or NC
A		0	NC	NC
B		7	7	7
CHS		5	54	54
CH	Try KH (5) and TCH (4)			
CK	Try K (5) and TSK (45)			
CZ	CS, CSZ, CZS	4	4	4
C	Try K (5) and TZ (4)			
DRZ	DRS	4	4	4
DS	DSH, DSZ	4	4	4
DZ	DZH, DZS	4	4	4
D	DT	3	3	3
EI	EJ, EY	0	1	NC
EU		1	1	NC
Ę	(Polish e-ogonek)	NC	NC	6 or NC
E		0	NC	NC
FB		7	7	7
F		7	7	7
G		5	5	5
H		5	5	NC
IA	IE, IO, IU	1	NC	NC
I		0	NC	NC
J	Try Y (1) and DZH (4)			
KS		5	54	54
KH		5	5	5
K		5	5	5
L		8	8	8
MN			66	66
M		6	6	6
NM			66	66
N		6	6	6
OI	OJ, OY	0	1	NC
O		0	NC	NC

P	PF, PH	7	7	7
Q		5	5	5
RZ, RS	Try RTZ (94) and ZH (4)			
R		9	9	9
SCHTSCH	SCHTSH, SCHTCH	2	4	4
SCH		4	4	4
SHTCH	SHCH, SHTSH	2	4	4
SHT	SCHT, SCHD	2	43	43
SH		4	4	4
STCH	STSCH, SC	2	4	4
STRZ	STRS, STSH	2	4	4
ST		2	43	43
SZCZ	SZCS	2	4	4
SZT	SHD, SZD, SD	2	43	43
SZ		4	4	4
S		4	4	4
TCH	TTCH, TTSCCH	4	4	4
TH		3	3	3
TRZ	TRS	4	4	4
TSCH	TSH	4	4	4
TS	TTS, TTSZ, TC	4	4	4
TZ	TTZ, TZS, TSZ	4	4	4
Ț	(Romanian t-cedilla)	3 or 4	3 or 4	3 or 4
T		3	3	3
UI	UJ, UY	0	1	NC
U	UE	0	NC	NC
V		7	7	7
W		7	7	7
X		5	54	54
Y		1	NC	NC
ZDZ	ZDZH, ZHDZH	2	4	4
ZD	ZHD	2	43	43
ZH	ZS, ZSCH, ZSH	4	4	4
Z		4	4	4
Letter	Alternate Spelling	Start of a name	Before a vowel	Any other situation

Fuente: <http://www.jewishgen.org/infofiles/soundex.html#DM>