Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação
E-ISSN: 1518-2924
bibli@ced.ufsc.br
Universidade Federal de Santa Catarina
Brasil

Gálvez, Carmen
Identificación de nombres personales por medio de sistemas de codificación fonética
Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, núm. 22, segundo semestre, 2006, pp. 105-116
Universidade Federal de Santa Catarina
Florianopolis, Brasil

IDENTIFICACIÓN DE NOMBRES PERSONALES POR MEDIO DE SISTEMAS DE CODIFICACIÓN FONÉTICA

PERSONAL NAME IDENTIFICATION THROUGH PHONETIC CODIFICATION SYSTEMS

Carmen Gálvez, PhD. - cgalvez@ugr.es
Departamento de Biblioteconomía y Documentación
Universidad de Granada

Comente este artigo no blog Ebibli = http://encontros-bibli-blog.blogspot.com/

Resumen
La necesidad de identificar las variantes de los nombres personales es un problema muy conocido en diversas aplicaciones, tales como los sistemas de recuperación de información (SRI), las bibliotecas digitales, las bases de datos de pacientes en un hospital, los sistemas de reservas aéreas, o los sistemas de censo. Los métodos de codificación fonética constituyen uno de los procedimientos para la solución de este problema, permitiendo obtener cadenas canónicas o normalizadas. Estos sistemas se engloban dentro de las técnicas generales de equiparación aproximada de cadenas. En este trabajo se realiza una revisión de los procesos que utilizan los sistemas *Soundex*, *Daitch-Mokotoff Soundex, Phonix*, *Metaphone* y *NYSIIS* para la asignación de claves fonéticas. La codificación fonética permite reducir a una forma común aquellos nombres personales que son similares en cuanto a su pronunciación, haciendo más sencilla la comparación de una cadena con otra, debido a que se almacena el código generado en lugar del nombre completo. Sin embargo, la principal limitación de estos sistemas es que son dependientes del lenguaje utilizado, lo que hace necesario la realización de modificaciones de acuerdo al idioma que se va a emplear.

Palabras-clave: Codificación fonética. Equiparación de nombres personales. Algoritmos de equiparación de nombre.

1 INTRODUCCIÓN

Un problema habitual de la recuperación de información (RI) en base de datos bibliográficas es la determinación de todas las formas variantes de los nombres personales tanto en el momento en el que estas cadenas se introducen en la base de datos, como en el momento de la búsqueda (en la que se establece una correspondencia entre los términos de la consulta y las formas de dichos nombres prealmacenadas). Una variante de nombre propio se podría definir como una cadena, que está conceptualmente relacionada con la forma correcta, o normalizada, de ese nombre. Las variantes se producen por distintas causas como son errores ortográficos, fonéticos o tipográficos (que darían lugar a omisiones, inserciones o sustituciones de caracteres en las cadenas) uso incorrecto de mayúsculas, errores de acentuación, o distinta distribución de los componentes del nombre propio.

Para solucionar los problemas anteriores se aplican técnicas de *equiparación aproximada*, encargadas de establecer una correspondencia entre las variantes y los nombres correctos almacenados en un diccionario. Son muchos los métodos empleados para realizar esa equiparación, entre ellos se encuentran los sistemas de codificación fonética que son los que vamos a tratar aquí.

2 EL PROBLEMA DE LAS VARIANTES DE LOS NOMBRES PERSONALES

Para la identificación de las variantes de los nombres personales se aplican técnicas generales de identificación y búsqueda de cadenas. Thompson y Dozier (1999) distinguen tres procesos: (*i*) *reconocimiento de nombres*; (*ii*) *equiparación de nombres*; y (*iii*) *búsqueda de nombres*. Las técnicas de reconocimiento se han tratado ampliamente en *Message Understanding Conferences* (MUC-4, 1992; MUC-6, 1995) dentro de las tareas específicas de los sistemas de Extracción de Información. En MUC-6, el reconocimiento de entidades, denominadas *named entity* (NE), se presenta como una parte clave de los sistemas de Extracción de Información. En MUC-7 el reconocimiento de NE se define como una tarea consistente en la identificación y categorización de tres subtareas (Chinchor, 1997) que se etiquetan con marcas SGML (*Standard Generalized Markup Language*): ENAMEX (para el etiquetado de nombres de entidad, personas, organizaciones y localizaciones), TIMEX (para el etiquetado de expresiones temporales), y NUMEX (para el etiquetado de expresiones numéricas, valores monetarios y porcentajes). A su vez, existen múltiples trabajos que están dedicados a la especificación de las reglas de formación de nombres personales y a la descripción de su estructura (Gaizauskas *et al.,* 1995; Ravin & Wacholder, 1996; Bikel *et al*., 1997; Baluja *et al.,* 2000).

Las técnicas de equiparación de nombres incluyen los métodos por medio de los cuales se comparan dos cadenas de caracteres, que se han reconocido como nombres, y se determina si las dos cadenas designan de hecho a la misma entidad. Dentro de este tratamiento se pueden producir dos situaciones. Primera, la equiparación es exacta y en este caso no se produce ningún problema. Segundo, la equiparación no es *exacta* haciendo necesario la aplicación de técnicas de *equiparación aproximada* de cadenas (Hall & Dowling, 1980).

Las técnicas de búsqueda de nombres incluyen los procesos a partir de los cuales se usa un nombre como parte de una consulta para recuperar información asociada con ese nombre en una base de datos. En este proceso también se pueden presentar dos situaciones.

Primera, la equiparación es exacta, y en este caso no se produce ningún problema. Segunda, puede surgir el problema de que no se recupere la información relevante porque, debido a las variantes, el sistema no es capaz de establecer una equiparación exacta entre los nombres utilizados en la construcción de la consulta y los nombres incluidos en los registros de la base de datos.

3 PROCEDIMIENTOS PARA LA IDENTIFICACIÓN DE NOMBRES PERSONALES

Algunos de los problemas anteriores se podrían solucionar por medio de la aplicación de programas de comprobación ortográfica, '*spelling ckeckers*', encargados de verificar los errores y corregir las variantes valiéndose de diccionarios en los que se almacenarían las formas correctas (Blair, 1960; Riseman & Elrich, 1971; Ullmann, 1977; Pollock & Zamora, 1984; Petersen, 1986; Damerau & Mays, 1989). Dentro de los sistemas de corrección de errores se pueden emplear básicamente dos planteamientos (Salton, 1989):

a) **Equiparación exacta** (*exact matching*) entre las variantes y los nombres correctos prealmacenados en el diccionario.
b) **Equiparación aproximada** (*approximate matching*) para encontrar las entradas del diccionario similares a las variantes.

El primer procedimiento consistiría en crear dos diccionarios: uno con las formas correctas y otro con las variantes. Sin embargo, con este método sólo se corregiría una pequeña porción de variantes (que serían aquellas que previamente se hubieran almacenado en el diccionario correspondiente). Dentro del segundo procedimiento se aplican básicamente dos métricas: *medidas de similitud de cadenas* y *medidas de similitud fonética*. Las medidas de similitud de cadenas, '*similarity measures*', se basan generalmente en la minimización de la distancia, o en la maximización de la similitud entre las entradas del diccionario y las variantes. Para calcular el coeficiente de similitud entre dos cadenas una medida muy conocida es contar el número de *n-grams* que las dos cadenas tienen en común (Angell *et al*., 1983). Otra medida de similitud es *edit-distance* (Damerau, 1964) consistente en contar el número de inserciones, supresiones o sustituciones de caracteres necesarios para transformar una cadena en otra. Por su parte, las medidas de similitud fonética se basan en la asignación de la misma clave, o código fonético, a los nombres que se pronuncian de forma parecida.

No obstante, es bien conocido que los programas de corrección ortográfica basados en algunas de las medidas anteriores no funcionan bien cuando se trata de comprobar la ortografía de los nombres propio. La falta de normalización de este tipo de cadenas hace que

haya muchas limitaciones en las entradas de los diccionarios, e incluso la no disponibilidad de este tipo de recursos, porque la tarea de su almacenamiento se vuelve muchas veces impracticable. Estas limitaciones se deben fundamentalmente a la gran diversidad de estructuras de este tipo de cadenas originada fundamentalmente por factores históricos y culturales (Borgman & Siegfried, 1992). Todos estos obstáculos hacen que sea muy difícil emplear un solo método para el procesamiento automático de este tipo de cadenas. Un estudio sobre los métodos de normalización de variantes de nombres propios y la propuesta de un nuevo procedimiento basado en la aplicación de técnicas de estado-finito se encuentra en Gálvez y Moya-Anegón (en prensa).

De cualquier forma, aquí nos vamos a centrar en los sistemas de codificación fonética, usados habitualmente para simplificar la búsqueda en las bases de datos cuando sólo se conoce la pronunciación de un nombre propio pero no su trascripción exacta. En general, estos sistemas parten de la suposición de que los nombres que comparten la misma clave se podrían considerar similares y se han utilizado principalmente en aplicaciones que involucran la identificación de nombres personales, tales como búsquedas en bases de datos bibliográficas y bases de datos de pacientes en un hospital, así como sistemas de reservas aéreas y sistemas de censo.

4 SISTEMAS DE CODIFICACIÓN FONÉTICA

La codificación basada en la similitud fonética de los nombres personales se aplica principalmente a los nombres y apellidos para reducirlos a una forma común. La mayoría de estos sistemas se desarrollaron originariamente para el idioma inglés. Los procedimientos de codificación fonética más conocidos se encuentran los sistemas *Soundex* (Odell & Russell, 1918), *Daitch-Mokotoff Soundex* (Daitch & Mokotoff, 1985), *Phonix* (Gadd, 1988, 1990), *Metaphone* (Philips, 1990) y *NYSIIS* (Taft, 1970). En el Anexo 1 se presenta algunos enlaces a programas que implementan los algoritmos fonéticos anteriores.

El algoritmo *Soundex* desarrollado y patentado por Odell y Russell (1918) reduce, particularmente apellidos ingleses, a un código de cuatro caracteres. El primer carácter es una letra mayúscula y los tres restantes son dígitos. Knuth (1973) describe el procedimiento utilizado por *Soundex* por medio de una función que consiste en: a) la conversión de caracteres a un código fonético, tal y como aparece en la tabla 1; b) un algoritmo que sustituye todos los caracteres, excepto el primero, por su correspondiente código fonético; c)

la eliminación de cualquier repetición consecutiva de caracteres; y d) la devolución únicamente de los primeros cuatro caracteres de la cadena resultante.

Tabla 1– Códigos fonéticos de *Soundex*

Código	Caracteres
0	a e h i o u w y
1	b f p v
2	c g j k q s x z
3	d t
4	L
5	m n
6	R

Cada nombre en la base de datos se clasificaría en algunos de estos tres rangos con respecto a la consulta: (*i*) *idénticos*; (*ii*) *diferentes pero compartiendo el mismo código*; y (*iii*) *no-relacionados*. El resultado de la aplicación del sistema *Soundex* a determinados apellidos ingleses se muestra en la tabla 2. Una modificación del algoritmo *Soundex* se realizó en *Extended Soundex Algorithm*. En este sistema el primer carácter se trata de la misma forma que los caracteres restantes, así el código que utiliza es puramente numérico, y esto da lugar a que la equiparación de los nombres con una codificación similar sea más rápida.

Tabla 2 – Resultado de la aplicación de *Soundex*

	Apellidos	Variantes	Códigos
Match	Appelt	Apelt	(A143, A143)
	Hobbs	Hubbs	(H120, H120)
Mismatch	Appelt	Appell	(A143, A140)
	Hobbs	Hobds	(H120, H130)

El sistema *Soundex* se usa actualmente por el *National Archives and Records Administration* (NARA) de EE.UU, pero tiene dificultades al aplicar el algoritmo a los apellidos judíos, germánicos o eslavos. Para solucionar estos problemas se creó el sistema *Daitch-Mokotoff Soundex* desarrollado en 1985 por Randy Daitch y Gary Mokotoff (publicado un año después en *Avotaynu*, el diario de la genealogía judía, en un artículo titulado "*The Jewish Soundex: a revised format*")[1]. El sistema Daitch-Mokotoff codifica todos los sonidos en cifras, formando un código de 6 dígitos. Las letras o sonidos tienen diferentes valores si están al principio de la palabra, en el centro, variando si anteceden a una vocal o no.

[1] Para una información adicional sobre el sistema *Daitch-Mokotoff Soundex*, véase MOKOTOFF, G., AMDUR, S. Where once we walked. **Avotaynu**, p. 567-569, 2002.

Si no se alcanza a tener las 6 cifras, se completa con ceros, hasta llegar a los 6 dígitos, por ejemplo:

Blejsman 784660

Kestenbojm 543676

Drukker 395900

Fried 793000

Sharon 496000

Otra adaptación del método *Soundex*, pero en este caso, a los nombres franceses la constituye el denominado *Henry Code*. Este sistema también clasifica los nombres en códigos de tres letras, pero produce muchos fallos porque a menudo modifica la estructura fonética de las cadenas analizadas, además de generar falsas correspondencias entre nombres completamente diferentes, o no establecer una relación entre nombres similares (Bouchard & Pouyez, 1980).

No obstante, el auténtico problema de los sistemas anteriores es que no son capaces de establecer algún tipo de *ordenación* entre las cadenas similares. Este problema se resuelve con una variante de *Soundex*, denominada *Phonix* (Gadd, 1988, 1990), cuyo algoritmo es más complejo que sus predecesores. El método de codificación *Phonix* se basa en la sustitución de todos los caracteres menos el primero por valores numéricos, con una leve variación, como se muestra en la tabla 3, y en la eliminación de todas las apariciones del valor '0'. La novedad que introduce *Phonix* es que realiza previamente unas 163 transformaciones de grupos de letras que normalizan las cadenas (por ejemplo, el carácter '*X*' se transforma en '*ECS*', además si la primera letra es una vocal o la consonante '*Y*' la transforma en '*V*'). Sin embargo, la aportación más importante de este sistema de codificación es que computa los sonidos finales, y como consecuencia de esto es capaz de establecer tres rangos de similitud constituidos por palabras que concuerdan: en los *sonidos finales*, en los *prefijos de los sonidos finales*, o con *sonidos finales distintos*.

Tabla 3–Códigos fonéticos de *Phonix*

Código	Caracteres
0	a e h i o u w y
1	b p
2	c g j k q
3	d t
4	l
5	m n
6	r
7	f v
8	s x z

Un algoritmo de codificación fonética parecido a los anteriores lo constituye el sistema *Metaphone* (Philips, 1990). Se trata de un sistema de codificación especialmente diseñado para el inglés americano. El algoritmo de *Metaphone* elimina las vocales, aunque éstas permanecen si son la primera letra de una palabra, reteniendo solamente las consonante, que se reducen a 16 consonantes sin incluir los dígitos (aunque hay excepciones como '0' para representar el sonido '*TH*'): B X S K J T F H L M N P O W Y. Además, se elimina la repetición de los caracteres consecutivos. Las transformaciones realizadas por las reglas del sistema *Metaphone* serían las siguientes:

B ⇒ B excepto en el final de una palabra
C ⇒ X si aparece en - cia-, - ch -
S si aparece en - ci-, - ce-, - cy -
K en el resto de los casos
D ⇒ J si está dentro de - dge-, - dgy-, - dgi -
T en el resto de los casos
F ⇒ F
G ⇒ silencio, si aparece en - gh -
J si está delante de - i-, - e-, - y -
K en el resto de los casos
H ⇒ silencio, si aparece después de vocal y no seguida por vocal
H en el resto de los casos
J ⇒ J
K ⇒ silencio, si aparece después de - c -
K en el resto de los casos
L ⇒ L
M ⇒ M
N ⇒ N
P ⇒ F si aparece delante de - h -
P en el resto de los casos
Q ⇒ K
R ⇒ R
S ⇒ X si aparece antes de - h - o dentro de - sio-, - sia -
S en el resto de los casos
T ⇒ X si aparece en - tia-, - tio
0 si aparece delante de - h -
T en el resto de los casos
V ⇒ F
W ⇒ silencio, si no está seguida por vocal
W si está seguida por vocal
X ⇒ KS
Y ⇒ silencio, si no está seguida por vocal
Y si está seguida por vocal
Z ⇒ S

Los códigos *Metaphone* estarían constituidos por cadenas que representarían aproximadamente cómo un nombre sonaría cuando se pronuncia usando las reglas de pronunciación de la lengua inglesa. El resultado de la aplicación del algoritmo *Metaphone* se muestra en la tabla 4.

Tabla 4–Resultado de la aplicación de *Metaphone*

	Apellidos	Variantes	Códigos
Match	Appelt	Apelt	(APLT, APLT)
	Hobbs	Hubbs	(HBS, HBS)
Mismatch	Appelt	Appell	(APLT, APL)
	Hobbs	Hobds	(HBS, HBTS)

Otro código fonético fue el propuesto por Taft (1970) y desarrollado por the *New York State Division of Criminal Justice*. El sistema de codificación presentado por Taft se denomina *New York State Identification and Intelligence Systems* (NYSIIS) y se basa en la reducción de los nombres a un código de hasta 6 letras. Las reglas utilizadas por el algoritmo NYSIIS para la codificación fonética son las siguientes:

1) El primer carácter de la clave fonética corresponde
al primer carácter del nombre

2) Traduce los primeros carácteres del nombre
MAC ⇒ MCC
PH ⇒ FF
KN ⇒ NN
K ⇒ C
SCH ⇒ SSS

3) Traduce los últimos caracteres del nombre
EE ⇒ Y
IE ⇒ Y
DT, RT, RD, NT, ND ⇒ D
Si el último carácter es S, eliminar;
Si el último carácter es A, eliminar;
Si los últimos caracteres son AY, sustituir por Y.

En su trabajo, Taft compara el *NYSIIS* algoritmo con *Soundex* y concluye que *NYSIIS* tiene una ratio de precisión del 98.72%, mientras la precisión de *Soundex* es del 95.99%. Sin embargo, Taft pone de manifiesto que tanto Soundex como *NYSIIS* sólo tratan las variantes de los nombres producidas por errores fonéticos. 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 trascripción, formas abreviadas, o variantes originadas por la distinta ordenación de las

secuencias de los componentes que forman los nombres personales. El resultado de la aplicación del algoritmo NYSIIS a un grupo de apellidos se muestra en la tabla 6.

Tabla 6–Resultado de la aplicación de *NYSIIS*

	Apellidos	Variantes	Códigos
Match	Appelt	Apelt	(APALT, APALT)
	Hobbs	Hubbs	(HAB, HAB)
Mismatch	Appelt	Appell	(APALT, APAL)
	Hobbs	Hobds	(HAB, HABD)

5 CONSIDERACIONES FINALES

La gran diversidad de variantes que tienen los nombres personales dan origen a errores en las consultas a las bases de datos, por esta razón se han desarrollado distintos algoritmos de normalización basados en medidas de similitud. Entre estos procedimientos, se encuentran los métodos de codificación fonética utilizados con la finalidad de generar claves fonéticas para los nombres personales cuando éstos se introducen en las bases de datos. Las claves se almacenan en los índices, como parte del registro en las bases de datos, a modo de claves correctas. En el momento de la consulta a la base de datos, los nombres que aparecen en la consulta se codifican con el mismo algoritmo utilizado en los índices. Este procedimiento permite establecer una comparación entre aquellos nombres que comparten el mismo código y, con ello, se lograría identificar determinadas variantes de nombres personales.

Aunque la correspondencia fonética incrementa el número de equiparaciones, o '*matches*', potenciales, las medidas de similitud en las que se basan los métodos fonéticos están limitadas para identificar errores de traducción o transliteración, usos de signos de puntuación, o variaciones en los formatos de un mismo nombres personales (tales como, '*Vorhees, Ellen M.*', '*Ellen M. Vorhees'*, '*E. M. Vorhees'*, o '*Vorhees EM'*). Por esta razón, es necesario complementar y combinar la equiparación fonética con otros métodos capaces de establecer nombres similares según su ortografía, como son los *métodos n-grams*, o *edit-distances*. Otra limitación es que los sistemas de codificación fonética se desarrollaron para nombres y apellidos en idioma inglés, por lo que es necesario realizar modificaciones según el idioma que se va a emplear. Finalmente, se puede afirmar que, a pesar de que las claves

fonéticas simplifican la búsqueda en las bases de datos cuando se conoce la pronunciación pero no la trascripción de los nombres, la incorporación de reglas fonéticas suplementarias en aplicaciones distintas al idioma inglés hace que estos sistemas tengan una complejidad adicional. De cualquier forma, como ya se ha mencionado, no sólo los errores sino los factores culturales están detrás de muchas cuestiones que se plantean en la identificación de los nombres personales y, en consecuencia, un sólo método no es capaz de solucionar este difícil problema.

REFERENCIAS

ANGELL, R. C., FREUND, G. E., WILLETT, P. Automatic spelling correction using a trigram similarity measure. **Information Processing & Management**, v. 19, n. 4, p. 255-261, 1983.

BALUJA, S., MITTAL, V., SUKTHANKAR, R. Applying machine learning for high performance name-entity extraction. **Computational Intelligence**, v. 16, 2000.

BLAIR, C. R. A program for correcting spelling errors. **Information and Control**, v. 3, p. 60-67, 1960.

BORGMAN, C. L., SIEGFRIED, S. L. Getty's synoname and its cousins: a survey of applications of personal name-matching algorithms. **Journal of the American Society for Information Science**, v. 43, n. 7, p. 459-476, 1992.

BOUCHARD, G., POUYEZ, C. Name variations and computerized record linkage. **Historical Methods**, v. 13, n. 2, p. 119-125, 1980.

CHINCHOR, N. Named entity task definition, version 3.5. In: SEVENTH MESSAGE UNDERSTANDIG CONFERENCE. **Proceedings…** Fairfax, VA: Morgan Kaufmann, 1997

DAITCH-MOKOTOFF SOUNDEX SYSTEM. Disponível em: <http://www.jewishgen.org.>

DAMERAU, F. J. A technique for computer detection and correction of spelling errors. **Communications of the ACM**, v. 7, n. 4, p. 171-176, 1964.

DAMERAU, F. J., MAY, E. An examination of undetected typing errors. **Information Processing & Management**, v. 25, n. 6, p. 659-664, 1989.

GADD, T. N. Fisching for werds: Phonetic retrieval of written text in information systems. **Program: Automated Library and Information Science**, v. 22, n. 3, p. 222-237, 1988.

GADD, T. N. (1990). PHONIX: the algorithm. **Program: Automated Library and Information Science**, v. 24, n. 4, p. 363-366.

GALVEZ, C., MOYA-ANEGÓN, F. Approximate personal name-matching through finite-state graphs. **Journal of the American Society for Information Science** (en prensa).

GAIZAUSKAS, R., *et. al*. University of Sheffield: description of the LaSIE system as used for MUC-6. In: Sixth Message Understanding Conference. **Proceedings…**Columbia, MD:

Morgan Kaufmann, 1995.

HALL, P. A. V., DOWLING, G. R. (1980). Approximate string matching. **Computing Surveys**, v. 12, n. 4, p. 381-402, 1980.

KNUTH, D. **The art of computer programming: sorting and searching**. Reading, Massachusetts : Addison-Wesley, 1973

MUC-4. In: FOURTH MESSAGE UNDERSTANDING CONFERENCE. **Proceedings…**McLean, VA: Morgan Kaufmann, 1992.

MUC-6. In: SIXTH MESSAGE UNDERSTANDING CONFERENCE. **Proceedings…**Columbia, MD: Morgan Kaufmann, 1995.

MUC-7. In: SEVENTH MESSAGE UNDERSTANDING CONFERENCE. **Proceedings…**Fairfax, Virginia: Morgan Kaufmann, 1997.

ODELL, M. K., RUSSELL, R. C. **U. S. Patent Numbers 1261167 (1918) and 1435663 (1922)**. Washington, D.C.: U.S. Patent Office, 1918.

PETERSEN, J. L. A note on undetected typing errors. **Communications of the ACM**, v. 29, n. 7, 1986.

PHILIPS, L. 1990. Handing on the Metaphone. **Computer Language**, v. 7, n. 12, p. 39-43, 1990.

POLLOCK, J. J., ZAMORA, A. Automatic spelling correction in scientific and scholarly text. **Communications of the ACM**, v. 27, n. 4, p. 358-368, 1984.

RAVIN, Y., WACHOLDER, N. 1996. Extracting names from natural-language text. **IBM Research Report 20338**, 1996

RISEMAN, E. M., ELRICH, R. W. Contextual word recognition using binary digrams. **IEEE Transactions on Computers**, v. 20, n. 4, p. 397-403, 1971.

SALTON, G. **Automatic text processing: the transformation, analysis and retrieval of information by computer**. Reading, Massachusetts: Addison-Wesley, 1989.

TAFT, R. L. **Special Report nº. 1**. Albany, New York: Bureau of Systems Development, New York State Identification and Intelligence Systems (NYSIIS), 1970.

THOMPSON, P., DOZIER, C.C. Name recognition and retrieval performance. In: Strzalkowski, T. (Ed.). **Natural language information retrieval.** Dordrecht: Kluwer Academic Publishers, 1999, p. 25-74.

ULLMANN, J. R. A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors. **The Computer Journal**, v. 20, n. 2, p. 141-147, 1977

ZAMORA, E., POLLOC, J., ZAMORA, A. The use of trigrams analysis for spelling error detection. **Information Processing and Management**, v. 17, n. 6, p. 305-316, 1981.

ABSTRACT
The need to identify the variants of personal names is a well-known problem in applications such as information retrieval systems (IRS), digital libraries, databases of patients in a hospital, the electronic systems of air reserves, or the systems of census. The phonetic codification methods constitute one of the procedures for the solution of this problem, permitting to obtain canonical or normalized names. These systems are included inside the general techniques of approximate string matching. In this work a revision of the processes is carried out that utilize the *Soundex*, *Daitch-Mokotoff Soundex*, *Phonix*, *Metaphone* and NYSIIS systems for the assignment of phonetic keys. The phonetic codification permits reduce to a common form those personal names that are similar in its pronunciation; performance simpler the string matching due to that the common code is stored instead of the complete name. Nevertheless, these systems are dependent of the language utilized, doing necessary the execution of modifications according to the language on the one that apply.

KEYWORDS: Phonetic codification. Personal name-matching. Name-matching techniques.

ANEXO 1–LISTA DE ALGUNOS *SOFTWARES* QUE IMPLENTAN LOS SISTEMAS DE CODIFICACIÓN FONÉTICA

Sistema	URL
Soundex	http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm#Algorithm
Metaphone	http://www.wbrogden.com/phonetic/index.html
Double Metaphone Algorithm.	http://aspell.sourceforge.net/metaphone/
New York State Identification and Intelligence System (NYSIIS)	http://www.dropby.com/NYSIIS.html
NameSearch®	http://www.name-searching.com/Working/Name_Search.htm
Daitch-Mokotoff Soundex	http://www.jewishgen.org/InfoFiles/soundex.html
JewishGen's JOS Calculator	http://www.jewishgen.org/jos/jossound.htm

Originais recebidos em 22/01/2006.