A Comparative of Spanish Encoding Functions
Efectiveness on Record Linkage

María del Pilar Angeles and Noemi Bailón-Miguel
Facultad de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México, México
pilarang@unam.mx, mimibailon@hotmail.com

Keywords: Data mining; Data matching; Record linkage; Data cleansing.

Abstract: Many business within big data projects suffer from duplicate data. This situation seriously impedes to managers to make well informed decisions. In the case of low data quality written in Spanish language, the identification and correction of problems such as spelling errors with English language based coding techniques is not suitable. In the case of Spanish language, written information is pronounced equal. There are phonetic techniques for duplicate detection that are not oriented to the Spanish language. Thus, the identification and correction of problems such as spelling errors in Spanish texts with such techniques is not suitable. In this paper we have implemented, modified and utilized in SEUCAD (Angeles, 2014) three Spanish phonetic algorithms to detect duplicate text strings in the presence of spelling errors in Spanish. The results were satisfactory, the Phonetic Spanish algorithm performed the best most of the time, demonstrating opportunities for an improved performance of Spanish encoding during the record linkage process.

1 INTRODUCTION

Data matching allows the following enterprise big data characteristics: a) optimizing the use of storage resources by eliminating redundant and possibly inconsistent data, hence reducing storage costs; b) enhancing the enterprise data quality through tighter governance on the consolidated hub; c) in order to execute more applications on this big data resource, you should be able to develop more powerful analytics, via MapReduce, YARN, R, and other programming frameworks.

Data de-duplication, results in better cache utilisation and less disk I/O. De-duplication is useful at any scale. In fact, most modern data warehousing products use column-based compression to achieve high de-duplication ratios and to improve performance. In the case of "text" big data, data de-duplication is highly recommendable. After all, the fastest and effective an I/O is, the least I/O required.

The way the companies handle their data makes its information more compressible too. For instance, the record linkage algorithms allow a better use of physical storage, reduce RAM, the information retrieval and its analysis are enhanced as there is no need to store the name of a person twice, besides the risk of being inconsistent.

Compression and deduplication play a key role in big data; In terms of economics, if a business system demands more storage resources than the competing systems, and the analysis takes longer, it will struggle to compete. The problem of detection and classification of duplicate records during the integration of disparate data sources affects business competitiveness. A number of encoding, comparison and classification methods have been utilized until now, but there still some work to do in terms of effectiveness and performance.

The present research was focused on the implementation and enhancement of Spanish encoding functions in order to improve the performance of the encoding phase during entity resolution when data has been written in Spanish language.

We have developed a prototype called Universal Evaluation System of Data Quality (SEUCAD) (Angeles, 2014) on the basis of the Freely Available Record Linkage System (FEBRL) (Christen, P. 2008). Within SEUCAD, there has been previously compared the Phonex, Soundex, and Modified Spanish phonetic functions in (Angeles, 2015). The Spanish phonetic coding was proposed in (Amon, 2015), which is an extended Soundex coding, where Spanish characters have been added. Besides, we have modified the Spanish Phonetic Algorithm so

Angeles M. and BailÃşn-Miguel N.
A Comparative of Spanish Encoding Functions - Efectiveness on Record Linkage.
DOI: 10.5220/0006227701050113
In *Proceedings of the Fifth International Conference on Telecommunications and Remote Sensing (ICTRS 2016)*, pages 105-113
ISBN: 978-989-758-200-4

the encryption code is resizable, and all white spaces are removed during encoding. The previous comparison showed that the modified version of the Spanish Phonetic Algorithm had a better performance in terms of precision. However, during the present research we have implemented two moreSpanish encoding functions: the Spanish Metaphone algorithm (Philips, 2000), (Mosquera, 2012), and a second version of such algorithm, which applies same code to similar sounds derived from very common misspellings.

The present paper is organized as follows: The next section briefly explains the data matching process. Section 3 explains the phonetic encoding functions proposed from previous research, the enhancements we have implemented on some of them, along with their role within de process of data matching. Section 4 presents the experiments carried out, and analyses the results. Finally, the last section concludes the main topics achieved regarding the performance of the encoding functions and the future work to be done.

2 RELATED WORK

The data matching process is mainly concerned to the record comparison among databases in order to determine if a pair of records corresponds to the same entity or not (Christen, 2012). It is also called record linkage o de-duplication. This process in general terms consists on the following tasks:

A standardization process (Christen, 2012), which refers to the conversion of input data from multiple databases into a format that allows correct and efficient record correspondence between two data sources.

Phonetic encoding is a type of algorithm that converts a string into a code that represents the pronunciation of that string. Encoding the phonetic sound of names avoids most problems of misspellings or alternate spellings, a very common problem on low quality of data sources.

The indexing process aims to reduce those pairs of records that are unlikely to correspond to the same real world entity and retaining those records that probably would correspond in the same block for comparison; consequently, reducing the number of record comparisons. The record similarity depends on their data types because they can be phonetically, numerically or textually similar. Some of the methods implemented within our prototype SEUCAD are for instance, Soundex (Odell, 1918), Phonex, Phonix (Christen, 2012), NYSIIS (Borgman, 1992), Double Metaphone (Philips, 2000).

Field and record comparison methods provide degrees of similarity and define thresholds depending on their semantics or data types. In the prototype, the algorithms Qgram, Jaro - Winkler Distance (Jaro, 1989), (Winkler, 1990), longest common substring comparison are already implemented.

The classification of pairs of records grouped and compared during previous steps is mainly based on the similarity values that were already obtained, since it is assumed that the more similar two records are, there is more probability that these records belong to the same entity of the real world. The records are classified into matches, not matches or possible matches.

The aim of the following section is to briefly explain the phonetic encoding functions that we have implemented and enhance in order to quantify and compare their performance during the record linkage process.

3 PHONETIC ENCODING PROPOSALS TO COMPARE

3.1 Phonetic Coding Functions

Phonetic encoding is a type of algorithm that converts a string (generally assumed to correspond to a name) into a code that represents the pronunciation of that string. Encoding the phonetic sound of names avoids most problems of misspellings or alternate spellings, a very common problem on low quality of data sources.

3.2 Spanish Phonetic

The Spanish phonetic coding function compared in the present document is a variation of the Soundex algorithm. Soundex is a phonetic encoding algorithm developed by Robert Russell and Margaret Odell in (Odell, 1918), and patented in 1918 and 1922. It converts a word in a code (Willis, 2002). The Soundex code is to replace the consonants of a word by a number; if necessary zeros are added to the end of the code to form a 4-digit code. Soundex choose the classification of characters based on the place of articulation of the English language.

The limitations of the Soundex algorithm have been extensively documented and have resulted in several improvements, but none oriented to the

Spanish language. Furthermore, the dependence of the initial letter, the grouping articulation point of the English language, and the four characters coding limit are not efficient to detect common misspellings in the Spanish language. The Spanish phonetic coding was proposed in (Amon, 2012), it is an extended Soundex coding, where Spanish characters have been added. In general terms the algorithm is as follows:

The string is converted to uppercase with no consideration of punctuation signs. The symbols "A, E, I, O, U, H, W" are eliminated from the original word. Assign numbers to the remaining letters according to Table 1.

Table 1: Spanish Coding

Characters	Digit
P	0
B, V	1
F, H	2
T, D	3
S, Z, C,X	4
Y, LL, L	5
N, Ñ, M	6
Q, K	7
G, J	8
R, RR	9

We have modified the Spanish Phonetic Algorithm (Angeles, 2014) so the encryption code is resizable, and all white spaces are removed during encoding. This model allows us to analyse a larger number of cases where we can have misspellings. The modified Spanish phonetic algorithm is called as soundex_sp in our SEUCAD prototype.

3.3 The Spanish Metaphone Algorithm

The Metaphone is a phonetic algorithm for indexing words by their English sounds when pronounced, it was proposed by Lawrence Philips in 1990 (Philips, 2000). The English Double-Metaphone algorithm was implemented by Andrew Collins in 2007 who claims no rights to this work. The Metaphone port adapted to the Spanish Language is authored by Alejandro Mosquera in (Mosquera, 2012); we have implemented this function and called as Esp_metaphone in our SEUCAD prototype. Some of the changes applied in order to adjust to the Spanish language are shown in Table 2, which considers typical cases of the Spanish language with letters such as á, é, í, ó, ú, ll, ñ, h.

Table 2: Spanish Metaphone

Char	Replacement
á	A
ch	X
C	S
é	E
í	I
ó	O
ú	U
ñ	NY
ü	U
b	V
Z	S
ll	Y

3.4 Modified Spanish Metaphone Coding Function

In Spanish language there are words such as “obscuro”, “oscuro” or “combate”, “convate” that should share the same code because even they are written different, their sound is similar and the misspelling is common. The second version of Esp_metaphone contains the following enhancements:

The Royal Academy of the Spanish Language reviewed words that originally were written with “ps” as “psicología”, and introduced some changes, because "the truth is that in Castilian the initial sound ps is quite violent, so the ordinary, both in Spainand in America, it is simply pronounced as “sicologia”. Moreover, our language, differing French or English, is not greatly concerned to preserve the etymological spelling; He prefers the phonetic spelling and therefore tends to write as it is pronounced." (Toscano-Mateus, 1965). Words that begin with "ps" can be written and pronounced as "s", and are called silent letters; for example, words “psicólogo” and “sicólogo”. We have added some cases to the Spanish Metaphone algorithm in order to consider these possible variations in Spanish written words and to assign the same code in both cases. Therefore, in case there is a word that starts with “ps”, it will be replaced by “s”. A special case with silent letter is presented with words like “oscuro” and “obscuro”, where both words have the same meaning so that the use of both is correct. In this case both its meaning and pronunciation is usually the same. Then, in case there is a word that starts with “bs”, it shall be replaced by “s”. One case of a common misspelling in Spanish language is given with words like “tambien” and “tanbien” were the latter is orthographically wrong, but phonetically is very similar to the former, and in case of typos, the letter “n” is close to letter “m” in a keyboard. Thus, we have decided to replace "mb" by "nb" and assign the same code. We have decided to replace "mp" by "np" and assign the same code in case of words such as “tampoco” and “tanpoco”. The words that begin with “s” followed by a consonant are replaced by 'es' such as “scalera” and “escalera”. Table 3 shows the additions contained in the Spanish Metaphone version 2.

Table 3: Modified Spanish Metaphone

Char	Replacement
mb	nb
mp	np
bs	S
ps	s

Table 4 shows coding from Metaphone and Metaphone_v2, the former is not able to apply the same code to words “psiquiatra“, “siquiatra“; “oscuro“, “obscuro“; “combate“, “convate“, “conbate“. All these words have the same meaning and in order to identify duplicates they should have the same code. In the case of code generated by Metaphone_v2 the code is the same, although there are not identical texts because of spelling mistakes but same meaning.

Table 4: Spanish Metaphone and Spanish Metaphone V2 coding

Word	Metaphone	Metaphone_v2
Caricia	KRZ	KRZ
Llaves	YVS	YVZ
Paella	PY	PY
Cerilla	ZRY	ZRY
Empeorar	EMPRR	ENPRR
Embotellar	EMVTYR	ENVTYR
Hoy	OY	OY
Xochimilco	XXMLK	XXMLK
Psiquiatra	PSKTR	ZKTR
siquiatra	SKTR	ZKTR
Obscuro	OVSKR	OZKR
Oscuro	OSKR	OZKR
Combate	KMBT	KNVT
Convate	KNVT	KNVT
Conbate	KNBT	KNVT
Comportar	KMPRTR	KNPRTR
Conportar	KNPRTR	KNPRTR
Zapato	ZPT	ZPT
Sapato	SPT	ZPT
Escalera	ESKLR	ESKLR
scalera	ESKLR	ESKLR

4 EXPERIMENTS

We have been developed and executed a set of experiments within the record linkage process through four scenarios; each scenario contains a different data-source. These experiments are aimed to identify for each data-set which encoding function has the best performance. The performance of the record linkage process is measured in terms of how many of the classified matches correspond to true real-world entities, while matching completeness is concerned with how many of the real-world entities that appear in both databases were correctly matched (Christen, 2012), (Churches, 2002). Each of the record pair corresponds to one of the following categories: True positives (TP): These are the record pairs that have been classified as matches and are true matches. These are the pairs where both records refer to the same entity. False positives (FP): These are the record pairs that have been classified as matches, but they are not true matches. The two records in these pairs refer to two different entities. The classifier has made a wrong decision with these record pairs. These pairs are also known as false matches. True negative (TN): These are the record pairs that have been classified as non-matches, and they are true non-matches. The two records in pairs in this category do refer to two different real-world entities. False negatives FN): These are the record pairs that have been classified as non-matches, but they are actually true matches. The two records in these pairs refer to the same entity. The classifier has made a wrong decision with these record pairs. These pairs are also known as false non-matches. Precision: calculates the proportion of how many of the classified matches (TP + FP) have been correctly classified as true matches (TP). It thus measures how precise a classifier is in classifying true matches (Odell, 1918). It is calculated as: precision= TP/(TP+FP). F-measure graph: An alternative is to plot the values of one or several measures with regard to the setting of a certain parameter, such as a single threshold used to classify candidate records according to their summed comparison vectors, as the threshold is increased, the number of record pairs classified as non-matches increases (and thus the number of TN and FN increases), while the number of TP and FP decreases.

An ideal outcome of a data matching project is to correctly classify as many of the true matches as true positives, while keeping both the number of false positives and false negatives small. Based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), different quality measures can be calculated. However, most classification techniques require one or several parameters that can be modified and depending upon the values of such parameters, a classifier will have a different performance leading a different numbers of false positives and negatives. Figure 1 shows the structure and sample source data utilized for experimentation.

nombre	apellido_pat	apellido_mat	calle
santiago		gonzalez	calle de san gumersindo
david	hernandez	cruz	calle de arnedillo
jessica	perez	martinez	calle de barbara de braganz
martha	sanchez	lopez	calle de jordi sole tura
patricia	garcia	aviles	calle de santa maria reina
alfonso	garcia	hernandez	calle del iridio
adriana	vazquez	gonzalez	calle de jose espelius
tania	mendez	lopez	plaza de arguelles
vicente		reyes	calle de infiesto
angelica	hernandez	brito	calle de los hermanos carpi
maria elena	perez	ramirez	calle de los bascones
isaac	martinez	gutierrez	calle de julia garcia boutan
berenice	ramirez	reyes	calle de elvira barrios
alejandro	alonso	flores	calle de la anunciacion
enrique	cordero	ramirez	calle del gladiolo

Figure 1: Sample of data source

The configuration of indexing, comparison and classification for all scenarios has been the same and repeated for each encoding function (Esp-Metaphone, Esp_metaphone_v2 and Soundex_sp). Such configuration is presented as follows:

1. **Indexing:**

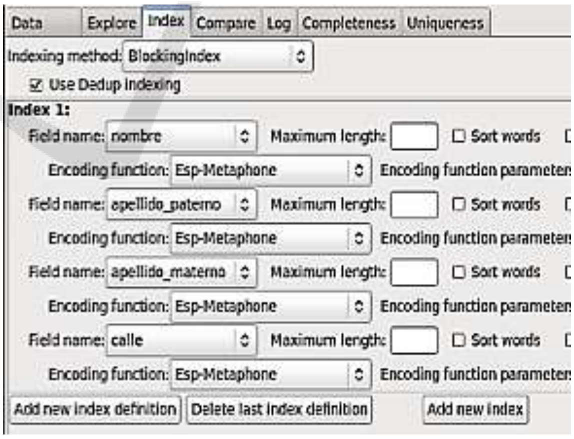

Figure 2: Indexing and encoding configuration

Fields that form the record require to be encoded and indexed in order to avoid a large number of comparisons between records whose fields are not even similar. Then, during the coding phase, we have executed for each experiment one of the coding functions: esp-metaphone, esp_metaphone_v2 or soundex_sp. We have chosen ”Blocking index” as indexing method based on fields: “nombre”, ”apellido paterno”, ”apellido materno”, ”calle”.

Figure 2 shows the configuration utilized for indexing and encoding methods.

2. Comparison: Once records have been ordered and grouped in terms of the previous fields specified. Each encoded field will be compared.

In order to obtain quality measures during the comparison step, we have chosen an exact function ”Str-Exact”, with “nombre”, ”apellido paterno”, ”apellido materno”, ”calle” fields.

Figure 3 shows the comparison specification for the experiments.

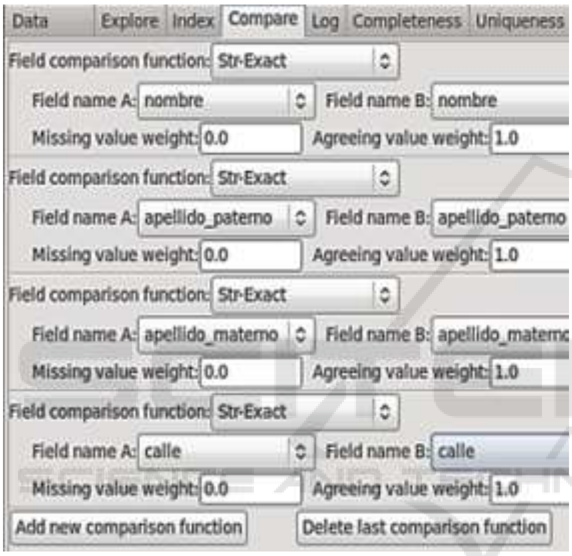

Figure 3: Comparison by String Exact method

3. Classification: In the case of pairs of record classification, we have selected the Optimal Threshold method, with a minimized false method of Positives and negatives, and a bin width of 40 for the range of values to be considered for the output graphic.

Figure 4 shows the classification configuration for the experiments.

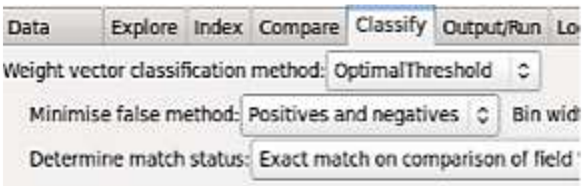

Figure 4: Classification by Optimal Threshold

4.1 Scenario I

The first file was generated with a total length of 1000 records, 100 duplicated records, one duplicated record for an original record as maximum, one change field per item as maximum, one maximum record modification, with a uniform probability distribution for duplicates.

The quality metrics obtained for each encoding method are presented in Table 5.

Table 5: Quality Metrics for Scenario I

Encode Method	Total Classif.	TP	FP	Precision	F-measure
Metaphone_sp	68	65	3	0.95588	0.977443
Metaphone_v2	69	66	3	0.95652	0.977777
Soundex_sp	76	73	3	0.96052	0.979865

According to the outcomes obtained from the first scenario, we can observe that in the case of the Modified Spanish coding function (soundex_sp), there were 76 record pairs classified, with 73 duplicated record pairs as true positives and 3 record pairs as false positives. Therefore, this method was 96% precise, slightly higher than the rest.

4.2 Scenario II

The second data source contained a total length of 5000 records, 500 duplicated records, one duplicated record for an original record as maximum, one change field per item as maximum, one maximum registry modification, with a uniform probability distribution for duplicates.

The quality metrics obtained for each encoding method are presented in Table 6.

Table 6: Quality Metrics for Scenario II

Encode Method	Total Classif.	TP	FP	Precision	F-measure
Metaphone_sp	320	319	1	0.9968	0.9984
Metaphone_v2	341	340	1	0.99706	0.99853
soundex_sp	353	352	1	0.99716	0.99581

From Table 6 we can observe that the Modified Spanish function classified 353 record pairs, with 352 duplicated record pairs as true positives and 1 record pair mistakenly classified as true match, corresponding then as one false positive. Therefore, this method was 99.7% precise, with more records classified than the Metaphone_sp and Methaphone_v2 with 320 and 341 records classified respectively.

4.3 Scenario III

The third data source contained a total length of 10000 records, 5000 duplicated records, one duplicated record for an original record as maximum, one change field per item as maximum, one maximum registry modifications, with a uniform probability distribution for duplicates. The process of record linkage under this scenario showed that the Modified Spanish coding function classified 3622 record pairs out of a total of 5000 potentially to detect, with 3620 duplicated record pairs as true positives and 2 record pairs mistakenly classified as true match. Therefore, this method was 99.94% precise. The Metaphone_sp and Methaphone_v2 phonetic functions obtained less records classified and more false positives than Spanish soundex function. The quality metrics obtained for each encoding method are presented in Table 7.

Table 7: Quality Metrics for Scenario III

Encode Method	Total Classif.	TP	FP	Precision	F-measure
Metaphone_sp	3333	3324	9	0.997299	0.9986
Metaphone_v2	3489	3480	9	0.99742	0.9987
Soundex_sp	3622	3620	2	0.99944	0.9997

4.4 Scenario IV

The fourth file has a total length of 1000 records, 100 duplicated records, one duplicated record for an original record as maximum, two changed fields per item as maximum, three maximum registry modifications, with a uniform probability distribution for duplicates.

The Modified Spanish coding function allowed that 964 record pairs could be classified; the total number of duplicates was actually 2500 records. However, this method did not present any false positive. The rest of the phonetic algorithms were 99% precise with two false positives, but the number of classified records was lower than those with Soundex_sp. The outcomes obtained for each encoding method under scenario IV are presented in Table 8.

Table 8: Quality Metrics for Scenario III

Encide Mrtod	Total Clas s	TP	FP	Precision	measure
Metapho-ne_sp	812	810	2	0.998753 6	0.998766
Metapho-ne_v2	884	882	2	0.99773	0.99886
Soundex_ sp	964	964	0	1	1

4.5 Analysis of Outcomes

According to the outcomes shown in previous section, we can observe that the Modified Spanish Phonetic algorithm was always more precise than the rest of the algorithms. Therefore, the Modified Spanish-Phonetic algorithm allows a higher proportion of how many of the classified matches (TP+FP) have been correctly classified as true matches.

The Spanish phonetic algorithm allows a total similarity greater than the remaining algorithms in all cases, because is more effective codifying Spanish words.

The Spanish phonetic algorithm achieved a slightly higher f-measure than the two versions of the Spanish Metaphone algorithm.

The graphics presented in this section, have been generated according to the variation of the coding function in order to observe the behaviour of the algorithms.

The precision obtained from each encode method for all the scenarios have been compared, graphed and shown in Figure 5, which shows the trend of the contribution of each encoding method to the precision of the classification.

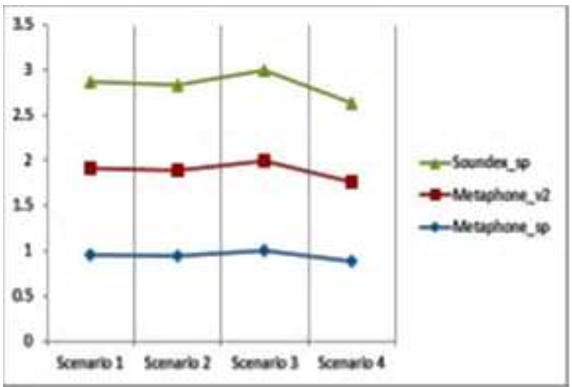

Figure 5: Precision of encode function

Figure 6 shows the trend of the contribution of each encoding method to the completeness of the classification.

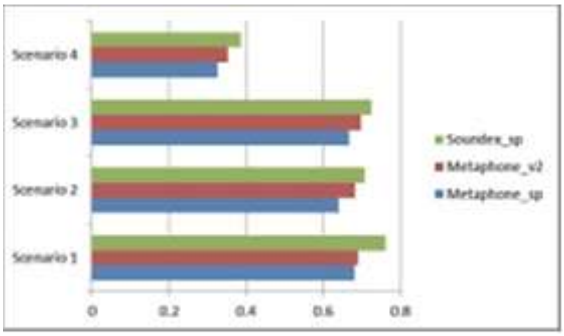

Figure 6: Completeness of each encoding method per scenario.

In other words, the proportion of record pairs classified against the total number of duplicates per scenario.

According to the outcomes shown in previous section, we can observe that the Modified Spanish Phonetic algorithm was always more precise than the two versions of Metaphone. Therefore, the Modified Spanish-Phonetic algorithm allows a higher proportion of true matches. The Spanish phonetic algorithm allows a total similarity greater than the remaining algorithms in all cases, because is more effective codifying Spanish words. The Spanish phonetic algorithm achieved a slightly higher f-measure than the rest. As we can observe from Figure 6, the Spanish phonetic algorithm obtained a larger number of pairs of records classified than the rest of the phonetic algorithms.

5 CONCLUSION

There are very real costs derived from duplicated customer data within big data.

Depending on the functional area (marketing, sales, finance, customer service, healthcare, etc.) and the business activities undertaken, high levels of duplicate customer data can cause hundreds of hours of manual reconciliation of data, sending information to wrong addresses, and decrease confidence in the company, increase mailing costs, increase resistance to implementation of new systems result in multiple sales people, sales teams or collectors calling on the same customer.

The present work has evaluated the record linkage outcomes under a number of different scenarios, where the true match status of record pairs was known. We have obtained precision, recall, and f-measure because they are suitable measures to assess data matching quality.

The Modified Spanish Soundex function presented a better performance than the rest of the phonetic functions during most of the experiments. However, it takes the longest execution time with a difference of some milliseconds.

It is important to be aware that the performance of a de-duplication system or technique is dependent on the type and the characteristics of the involved data sets, having good domain knowledge is relevant in order to achieve good matching or deduplication results.

We have previously concluded in (Angeles, 2015) that the Modified Spanish Phonetic algorithm was always more precise and complete than Soundex y Phonex.

Under a new set of experiments we have carried out against a Spanish version of the Metaphone algorithm and an enhanced version of the Spanish Metaphone, the Modified Spanish Phonetic algorithm still having the best performance in terms of precision in the majority of the cases we have experimented during the present research.

ACKNOWLEDGEMENTS

This work is being supported by a grant from Research Projects and Technology Innovation Support Program (Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica, PAPIIT, UNAM Project 1N114413 named Universal Evaluation System Data Quality (Sistema Evaluador Universal de Calidad de Datos).

REFERENCES

Angeles, P., et al., 2014. Universal evaluation system data quality, In DBKDA 2014 : *The Sixth International Conference on Advances in Databases, Knowledge, and Data Applications,* vol. 32, pp. 13–19.

Angeles, P., J. García-Ugalde, A. Espino-Gamez, and J. Gil-Moncada, Comparison of a Modified Spanish Soundex, and Phonex Coding function during datamatching process, *In International Conference on Informatics, Electronic and Vision, ICIEV, Kytakyushu, Fukuoka Japan,ISBN:978-1-4673 6901-5,* DOI:10.1109/ICIEV.2015.7334028, IEEE, pp.1-6,2015.

Borgman, C.L. & S. L. Siegfried, 1992. Getty's synonym TM and its cousins: A survey of applications of personal name-matching algorithms. In *Journal of the American Society for Information Science* 43((7)), 459-476.

Christen, P., 2008. Febrl A Freely Available Record Linkage System with a Graphical User Interface. *Second Australasian Workshop on Health Data and Knowledge Management* (HDKM 2008), 80, 17-25. Figure 6: Completeness of each encoding method per scenario.

Christen, P., 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection. *Springer Data-Centric Systems and Applications.*

Churches, T., P. Christen, K. Lim, & J. X. Zhu, 2002. Preparation of name and address data for record linkage using hidden Markov models. In BMC Medical Informatics and Decision Making 2 (1), 9.

Cohen, W. W., P. Ravikumar, & S. E. Fienberg, 2003. The book, *A comparison of string distance metrics for name matching tasks*, 73-78.

Rahm E. & H. Do, 2000. Data cleaning: Problems and current approaches. In *IEEE Data Engineering 23* (4), 3-13.

Amon F. M. I. & J. Echeverria, 2012. Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español. In *Ingenierías Universidad de Medellin 11 (20),* 120 138.

Jaro, M. A., 1989. Advances in record-linkage methodology applied to matching the 1985 Census of. Tampa, Florida. In *Journal of the American Statistical Association, 84,* 414-420.

Mosquera, A., E. Lloret, & P. Moreda, 2012. Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation. In *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility*, 9 -14.

Odell, M. & R. Russell, 1918. The book, *The soundex coding system*. (1261167).

Philips, L., 2000. The double metaphone search algorithm. In *C/C++ Users J 18 (6),* 38-43.

Toscano-Mateus, H., J. B. Powcrs, (ed.), 1965. *Hablemos del lenguaje.*

Winkler, W., 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 354-359.