Arts and Social Sciences Journal
ISSN: 2151-6200

Matos and Coelho, Arts Social Sci J 2015, 6:3
DOI: 10.4172/2151-6200.1000120

Research Article Open Access

PASVAG: A Routine to Relate Different Geographical Names

Matos V* and Coelho V

IME, Instituto Militar de Engenharia SE-6, Rio de Janeiro-BR, Brazil

Abstract

This research may contribute to the development of studies for geographically investigate or ethnic historically the origins or changes of Geographical Names. According Rostaing "The Toponymy intends to seek the origin of place names and also to study its transformations. "Quantifying these changes can reveal relationships proximity of words that the indicator is proposed. Example: change naming a neighborhood of Rio de Janeiro City- BR suffered, formerly known as "REAL ENG °" becoming in the "Realengo" neighborhood. Assign a distance value of this change can reveal never before made associations. Thus, research walked in the search for solutions that fit issues of computational linguistics and textual similarity indices applied to the Geographical Names, that these attributes become binding key-in processing queries to different databases. The goal is to enable the management and recovery of a large body of data. This indicator similarity proposed in this paper has been tested and confronted between experiments with simulated data. The main results of the experiments recognized standards in testing, and the importance of the variable noise position in the string, as well as usage limits for similarity component in the integration of databases.

Keywords: Geographical names; String; Hypothesis; Realengo

Introduction

When working with information originating from various sciences, there is a need to seek recovery strategies and exploitation of data visualization [1]. From a technological standpoint, the space used as a reference for data exploration allows adding understanding and insight in building the information [2]. Because of this, equip the information to design a geographic information system is characterized by elements such as multidisciplinary and interdisciplinary [3-5]. This heterogeneous set of such data depends on the integration of several sciences and reflects this context storage requirement of different types of data to group as logical format records constitute a Geographic Database (GDB).

For the purpose of this work is used the notion of similarity of geographical names (NG) in which admits a measure of how two strings are similar. On the premise is quantification of similarity is based on the metric space. And in turn, provides the notion of distance and relative proximity to the idea of the first law of geography "all things are similar, but closer things are more related than distant things" [6].

The growing demand of information makes the search for technological innovation tools have premised the need to generate and store large quantities of records. Thus, in the context of data replication, the theme of Geographic Database (BDG) converges to central labor issue. Treating the problem from the perspective of the universe structural [7,8], which included the concepts of reality to be stored on the computer, i.e. the BDG; it uses variables that give meaning to the question of spatiality of formal models for geographic entities. They are: geometric field that stores the registry geometry (point, line or polygon) by one or more pairs of coordinates; and field; filing its geographical name.

For the same meaning is applied to Geographic Names. It adopted the close relationship in response indicator of similarity. Therefore, if the similarity between two strings gets close to zero, it indicates symmetry between the NG, and through the text field present in BDG, called "geographical name", there is the limits of efficiency when changing similarity index to propose an acceptance threshold for similarity of Geographical Names.

Materials and Methods

The construction of the database to test the proposed indicator proposal (PASVAG), implemented in Postgres, delimited string in the set range of 1 to 20 characters which limits the maximum size of digits that you use to create the database. Through algorithms written in JAVA programming language, povoaremos the database with the simulation of a sequence of characters including noise (dissimilar character chain) in every possible position in the string. Thus, the performance is evaluated indicator similarity to character sets and changes the cardinality of its chains, the position (s) Noise (s) as well as the noise size. This routine allowed generating the comparison materials and inferring a noise in the string to which you want to compare, and so calculate the distance with controlled noise.

First

Read string1;

Read similarity;

Read size;

Assign zero to contador1;

Assign zero to contador2;

Assign to empty matrix;

Read string1 in the matrix;

While contador1 <size make

Matrix receives "@" in contador1 position; Contador2 receives increment;

***Corresponding author:** Matos V, IME, Instituto Militar de Engenharia SE-6, Rio de Janeiro-BR, Brazil, Tel: +5521997031054; E-mail: vanderlei.matos@gmail.com.br

Received June 15, 2015; **Accepted** August 13, 2015; **Published** August 21, 2015

Citation: Matos V, Coelho V (2015) PASVAG: A Routine to Relate Different Geographical Names. Arts Social Sci J 6: 120. doi:10.4172/2151-6200.1000120

While contador2 <size make Matrix receives "@" in contador2 position; Calculates the similarity matrix;

Guard the matrix and similarity; Increases contador2;

End while;

Increases contador1;

End while;

Purpose;

The Pseudo shows on your first level data entry. At this stage, the variable is loaded with the sequence of characters. In the example above s1 receives "ABCD", in the next step, triggers the counter and if verifies-the size of s1. Immediately below, the diamond is the accountant who handles the condition and repeats the subsequent expressions in the number of times the cardinality of the string. In the following level, replaces the position 1 of the array by noise (@) and increments the second counter. The following is checked the second counter condition that will process the procedures for applying the calculation of similarity and keep its results in the database.

Figure 1 demonstrates how was filed the records in the database. This process was repeated until strings with cardinality 20. The total universe with group sizes yielded by the law of permutations, a universe of 2,097,130 records. This small sample of the database exemplifies the distribution noise occupying all positions of possibilities.

The Table 1 illustrates a records section of how dissimilar variables were allocated in accordance with the methodology already described, comprising a permutation in leasing all the possibilities for noise values for a string with four characters cardinality. It is noteworthy that the concept of noise for this approach, symbolized by the signal "@" means the percentage of the same upon size of the string from which is compared. Then equation 1 refers:

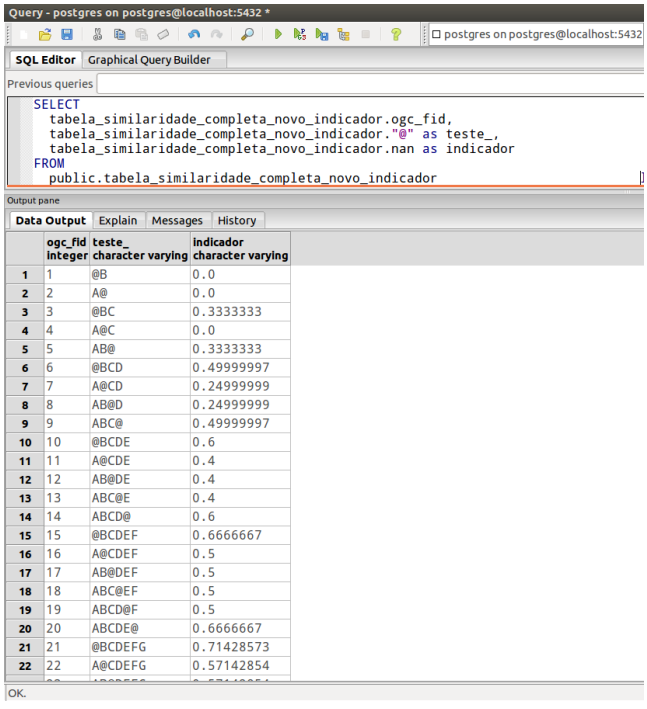

Figure 1: Bunch of data storage screen.

s2	@BCD	A@CD	AB@D	ABC@
	25%			
	@@CD	@B@D	@BC@	AB@@
	50%			
	@@@D	@@C@	@B@@	A@@@
	75%			
	@@@@			
	100%			

Table 1: Percentage of noise by character size.

s2	@BCD	@BCDEFG@	@BCDEFGH@JK@
Cardinalidade	4	8	12
Ruído	25%		

Table 2: Noise variety of demonstration in different cardinality.

R% - Noise Percentage

r - Number of dissimilar characters.

t - Number of string characters.

$$R\% = (\frac{r}{t}) \times 100 \quad (1)$$

The use of this concept provides comparability set of cardinality tracks this problem in the simulated scenario. Still on the topic above, the formulation proves the impossibility of the terms "r" and "t" assume different values of integer and smaller than "one", as the atomic unit of a string is invariably "one" character. What makes a restriction on the universe of possibilities for noise levels to a finite universe of known values. Another observation pointless the eq.1 equation is the effect of the term "t" in the composition percentage of noise. In Table 2 if we observes-that for a same percentage of noise, 25% exist values of higher concentration of dissimilar terms for a same range of noises, Logo, the higher the cardinality the chain of greater characters will be the amount of noises distributed to a same percentage. To carry out the calculation of this new indicator of similarity to the scope of the Geographic Name (NG), is Question meet a set of assumptions that are:

Assumptions

Be C a chain of characters any, $C = c_1c_2c_3 \ldots c_n$, , so that assigned a geographical name (NG) as $"C = Rio"$. And "L" a set composed by list of characters $"\mathrm{L}(\mathrm{C}) = \mathrm{L} = \{\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_3, \ldots \mathrm{c}_n\}"$. At this stage it is assumed that the set of characters belonging to the NG are the elements that compose the form of ordered lists the set "L", and thus $"\mathrm{L} = \{R, i, o\}"$.

Ergo reformulate elements of the set "L" on form of bigrams composing the set "B". This recast elements is expressed by the formation rule given by the position of the elements in $"c_{n-1}c_n"$, and thus $"B(L) = B = \{c_1c_2, c_2c_3, \ldots c_{n-1}c_n\}"$. That is, the set "B" is rearranged as $"B = \{Ri, io\}"$.

In sequence, adopts-if the use of cardinality of the sets, i.e. the number of elements within a set "W" any, if denotes by $|\mathrm{W}|$. Therefore, the cardinality of "W" is nothing more than the set number of objects. Using the example of the set $"\mathrm{L} = \{R, i, o\}"$ the "L" set cardinality is given by $|\mathrm{L}| = 3$.

The following proposition consists of data two sets any "W" and "Z". Be the difference $"(W - Z)"$ the set composed of elements that are in the set "W" and are not in the set "Z". Thus, $"(W - Z) \neq (Z - W), \forall W \neq Z"$. These information forms the basis for understanding the subtraction

operation of set theory. A practical example of this use by NG, either the sets: $e\,L_2 = \{R,i,c,o\}$ $e\,L_2 = \{R,i,c,o\}$. Figure 2 illustrates through the Venn diagram the difference between sets by subtracting the portion not hatched is observed as a result of an empty set. However, in Figure 2 is changed order of subtraction of the joint which results in a unitary assembly. Therefore, the proposition for the whole set "$L_1 \neq L_2$", subtracting "$(L_1 - L_2) \neq (L_2 - L_1)$".

And last, I is an index of similarity that meets the requisites of metric space.

Motion for Similarity Indicator for Geographic Names

In calculating this new indicator, addresses the mapping of the string of two units: characters and bigrams. In characters, it is taken as noise the character set that are present in L_2 and L_1 not. We told these differences, we will treat the set of characters that are present in L_1 and not in L_2, so that these differences are summed, termed as noise in the characters (R_c)

$$R_c = |L_1 - L_2| + |L_2 - L_1| \tag{2}$$

Then, referred to as noise ratio (T_c) will be the reason (2) in (3) in the following formulation.

$$T_c = \frac{R_c}{|L_1| + |L_2|} \tag{3}$$

In this step the process repeats with new mapping unit string into bigrams. Made the sum of the differences is found the amount of noise in the bigrams (R_B).

$$R_B = |B_1 - B_2| + |B_2 - B_1| \tag{4}$$

In this other step is calculated noise ratio in bigrams (T_B) applying the ratio (eq 4) (eq. 5).

$$T_B = \frac{R_B}{|L_1| + |L_2|} \tag{5}$$

Finally the indicator is made by applying the variables (EQ. 3) and (EQ 5) to the variables of the equation (EQ. 6).

$$I = (1 - T_c) \times (1 - T_B) \tag{6}$$

Thus, to satisfy the condition of metric space, the proposed indicator meets this premise reverse. Therefore, the distance between two strings is é $f(C_1, C_2) = 1 - I$, if the indicator (I) obtains maximum similarity will result in 1, but applied to the similarity function will zero distance between the string comparisons.

As seen previously, the formulation of the indicator similarity to NG must meet the following conditions.

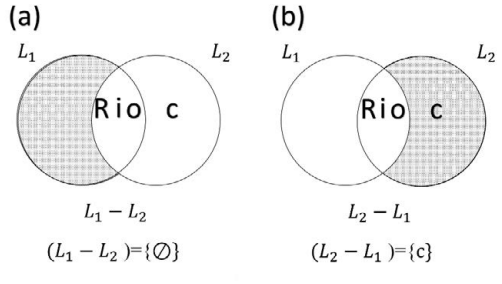

Figure 2: Example with use of Venn diagrams.

Para que : $\begin{cases} R_C \in N : |L_1| + |L_2| \geq 0 \\ R_B \in N : |B_1| + |B_2| \geq 0 \end{cases}$

Once the test if the indicator formulation meets the expected properties, it was found the maximum and minimum limits of similarity. For example, the comparison of two strings in the amount of resulting noise null means that all the elements fit together, therefore, we have a case of maximum similarity value: $\begin{cases} R_C = 0 \\ R_B = 0 \end{cases}$, applying the Equation (eq.3) and (eq.5).

$$T_c = \frac{0}{|L_1| + |L_2|} = 0$$

$$T_B = \frac{0}{|B_1| + |B_2|} = 0$$

The null result for noise rates when applied in (eq.6).

$$I = (1-0) \times (1-0) = 1 : f(C_1, C_2) = 1 - I = 0$$

Results in the maximum similarity are validation demonstrates that the proposed indicator fits the requirements of metric space as it obeys the positivity condition and zero distance between the strings.

When the limit being the minimum of similarity or the maximum distance, when the geographical names are quite dissimilar, that is, there is no similarity between the pair of strings, in which case the amount of noise is greater than zero and equal be conditional as

$$|L_1| = a; |L_2| = b; (L_1 - L_2) = a; (L_2 - L_1) = b; R_C = a + b,$$

where a ≠ b ≠ 0 by applying the equation (eq. 3) and (eq 5).

$$T_c = \frac{a+b}{a+b}$$

$$R_B = a + b - 2$$

$$T_B = \frac{a+b-2}{a+b-2} = 1$$

$$|B_1| = |L_1| - 1 = a - 1$$

$$|B_2| = |L_2| - 1 = b - 1$$

$$(B_1 - B_2) = a - 1$$

$$(B_2 - B_1) = b - 1$$

Soon

I=0

It provides us the value of the indicator, even if not known the value of T_B, but by multiplying T_C, will have zero multiplied by any amount resulting in zero.

This final result demonstrates that the function of the compared strings obtained maximum distance from one another, which means no similarity.

Calculation Methods

Unless the conditions of use of the similarity to the specificity of the NG will need to make measurable how different they are NG. For this, we will take the notion of distance to measure the similarity. So the shorter the distance, the more similar the NG are, and the greater

the distance, the lower the similarity between them. By applying the necessary requirements to ensure a metric space, these properties are attributed to the indicator.

Let's see if the proposed indicator meets the conditions to print its results in the metric space. Be the first property the metric space positivity condition: $f(C_1, C_2) \geq 0$ for all C_1, C_2 in X. In the example of its application: Be I("BANANA", "ANANAIS") the similarity of two strings C_1 and C_2 (Table 3).

The results found indicated in the last row of the table. Satisfies the first condition of metric space in which its value is positive. The following test demonstrates that the indicator by finding a maximum similarity, when the distance between the strings is zero. Let's see if C1=C2 meets the distance (C1, C2)=0. Be I("BANANA", "BANANA") (Table 4).

When you view the zero in response, it is understood that the distance between the words is null proving more this property. To validate the third property, symmetry f(C_1, C_2)=f(C_2, C_1), will reuse the example of the first property entering the inverted variable is I ("Ananais", "Banana") the similarity of two strings C_1 and C_2 (Table 5).

At the end of the calculation found that the result is the same as example1, meaning that no matter the order of the input variables of the similarity of function, because its result will be the same. In this check it is concluded that the indicator meets the symmetry condition metric space. Following is proven the last condition, triangular inequality, to contemplate the premises of metric space. Be C_1, C_2 and C_3, "Banana", "Ananais" and "TOMATE" strings which you want to calculate their distances (Tables 4 and 6).

Below we find the distance C2, C3 (Tables 5 and 7).

With these results it can be seen in the graph below the validation of the triangle inequality property, d(C1, C2) ≤ d(C1, C3) + d(C2, C3). Visually checked in Figure 2 is that the sum of f distances f(ABACAXI, BANANA) + f(ABACAXI, ANANAIS) is greater than the distance f(ANANAIS, BANANA) (Figure 3).

	C1		C2	=
Ex	BANANA		ANANAIS	
L	{B,A,N,A,N,A}		{A,N,A,N,A,I,S}	
B	{BA,AN,NA,AN,NA}		{AN,NA,AN,NA,AI,IS}	
Rc		$\|L_1-L_2\|=1+\|L_2-L_1\|=2$		3
R_B		$\|B_1-B_2\|=1+\|B_2-B_1\|=2$		3
T_C		$Rc/\|L_1\|+\|L_2\|$		3/|6+7|
T_B		$R_B/\|B_1\|+\|B_2\|$		3/|5+6|
I		(1-Tc)×(1-T_B)		0,559
f(d)		f(1-I)		0,441

Table 3: Example of calculating the indicator proposed to prove the positivity condition.

	C1		C2	=
Ex	BANANA		BANANA	
L	{B,A,N,A,N,A}		{B,A,N,A,N,A}	
B	{BA,AN,NA,AN,NA}		{BA,AN,NA,AN,NA}	
Rc		|L1-L2 |=0+|L2-L1|=0		0
RB		|B1-B2|=0+|B2-B1|=0		0
TC		Rc/|L1|+|L2|		0/|6+6|
TB		RB/|B1|+|B2|		0/|5+5|
I		(1-Tc)×(1-TB)		1
f(d)		f(1-I)		0,0

Table 4: Example of calculating the indicator proposed to prove the null distance condition.

	C1		C2	=
Ex	ANANAIS		BANANA	
L	{A,N,A,N,A,I,S}		{B,A,N,A,N,A}	
B	{AN,NA,AN,NA,AI,IS}		{BA,AN,NA,AN,NA}	
Rc		|L2-L1 |=2+|L1-L2|=1		3
RB		|B2-B1|=2+|B1-B2|=1		3
TC		Rc/|L1|+|L2|		3/|7+6|
TB		RB/|B1|+|B2|		3/|6+5|
I		(1-Tc)×(1-TB)		0,559
f(d)		(1-I)		0,441

Table 5: Example of calculating the indicator proposed proving the symmetry.

	C1		C3	=
Ex	ANANAIS		ABACAXI	
L	{A,N,A,N,A,I,S}		{A, B, A, C, A, X, I}	
B	{AN,NA,AN,NA,AI,IS}		{AB, BA, AC, CA, AX, XI}	
Rc		|L3-L1|=3+|L1-L3|=3		6
RB		|B3-B1|=6+|B1-B3|=6		12
TC		Rc/|L1|+|L2|		6/|7+7|
TB		RB/|B1|+|B2|		12/|6+6|
I		(1-Tc)×(1-TB)		0,0
f(d)		f(1-I)		1

Table 6: Example of calculating the indicator proposed proving the condition of triangular inequality.

	C1		C3	=
Ex	BANANA		ABACAXI	
L	{B,A,N,A,N,A}		{A, B, A, C, A, X, I}	
B	{BA,AN,NA,AN,NA}		{AB, BA, AC, CA, AX, XI}	
Rc		|L3-L2|=4+|L2-L3|=5		9
RB		|B3-B2|=4+|B2-B3 |=5		9
TC		Rc/|L1|+|L2|		9/|6+7|
TB		RB/|B1|+|B2|		9/|5+6|
I		(1-Tc)×(1-TB)		0,05
f(d)		f(1-I)		0,95

Table 7: Example of calculating the indicator proposed proving the condition of triangular inequality.

Results

For the realization of similarity tests in the string with simulated data, an important point is the definition of the sequence of known characters that will be the basis for the studied models. This representation allows you to play on a smaller scale, the universe of possibilities that an alphanumeric digit occupy the composition of an NG. The design of the tests with simulated data base was provided both to discover other parameters on the study problem, as the test marker proposed similarity and to identify their advantages and limitations. In this simulation model, the capture of noise effect on strings was a great value greatness to formulate a theoretical model and refine knowledge of markers on the similarity of NG.

Data simulated results

In Figure 4 is observed the effect of noise of 40% (two dissimilar characters) to cardinality of six characters. To the extent that it varies the position of the noise observed different patterns of responses:

1) The first position when the noise occupies the ends of the string, namely the sign "@" symbol noise grouped at the beginning, end, or occupying its two ends. This pattern gives the maximum cutting pattern for this range of cardinality and noise. This feature has a positive effect for the treatment of addresses because the indicator demonstrates tolerant exchange of address types, "Street or Avenue", framed at the beginning of the string.

2) The next standard, obtained of the noise occupying one end, or grouped within the string. This pattern shows an average cut in relation to other positions noise.

3) In the hard cut pattern is observed that noise manifests itself within the chain, occupying positions spaced.

Given the above, we note that for any cardinality there is a pattern of response that distinguishes the various known noise positions. This makes it possible adoption to any geographical name.

The following were checked outliers marker similarity. In Figure 5 it is shown the performance of the similarity function for cardinality of twenty characters with minimal cutting pattern. By imposing the premise we "one" for maximum similarity and absence of noise. However, according to the behavior of the represented function, the similarity score comprises measuring an amount of dissimilar terms of

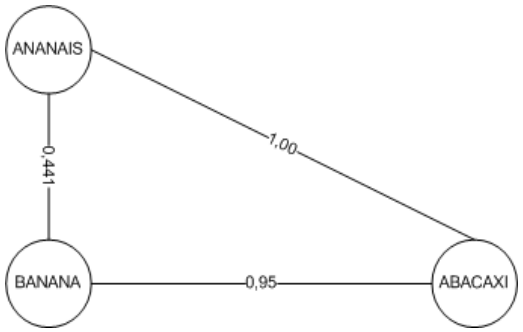

Figure 3: Graph representing the distances C1, C2, C3.

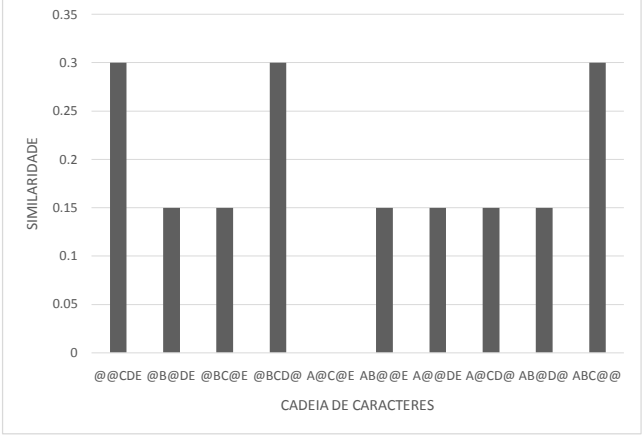

Figure 4: Width of the results due to the positioning of the noise.

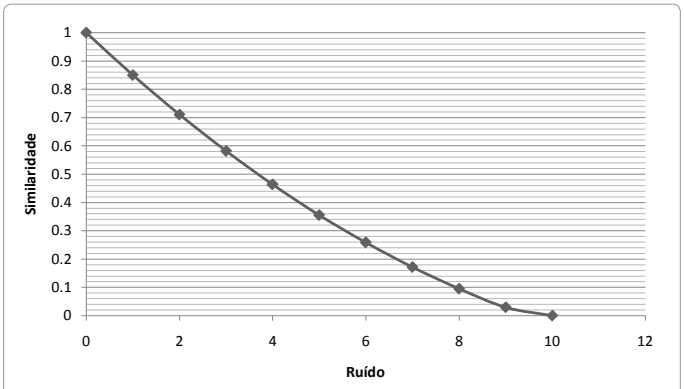

Figure 5: Leasing the minimum points to a chain of twenty characters.

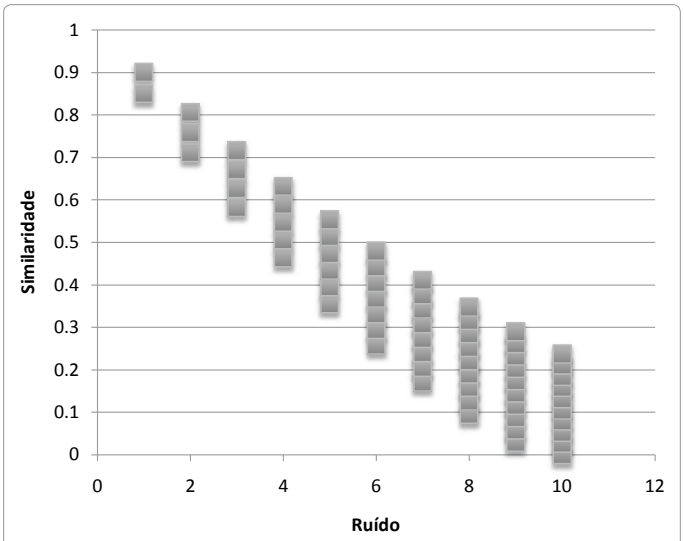

Figure 6: Behavior of similarity to noise variation and constant cardinality.

cardinality up to half of the string (values above 50% noise).

The amplitude indicator similarity to-noise variation was investigated and shown in Figure 6, through the similarity score. As the noise is increased and features of smaller amounts of similarity, the greater the amplitude between maximum and minimum values of class similarity. Example is the marker in the position that admits only noise Rc=1: The marker has a similarity Rc=[8.5, 9.0].

For a second outlet point values for the three position for abscissa values have their order Rc(3)=[5.8, 7.2].

Upon the behavior of the tested similarity function, one can create a rule for quantifying the amplitude response as a function of cardinality and the amount of noise. The amount of similarity is always results the amount of "noise + 1". That is, a chain of cardinality of twenty characters with two obtains a noise amplitude similarity three valid results. This expression is valid up to the limit indicator of similarity, that is, half of its cardinality.

Another important observation is shown in Figure 7. When you enter a five-character noise, the rule of verification established to quantify the number of similar responses is the same. Highlighted by the red rectangle, the standard six answers spreads no matter how much rises the cardinality. With this data we find that the outlet values

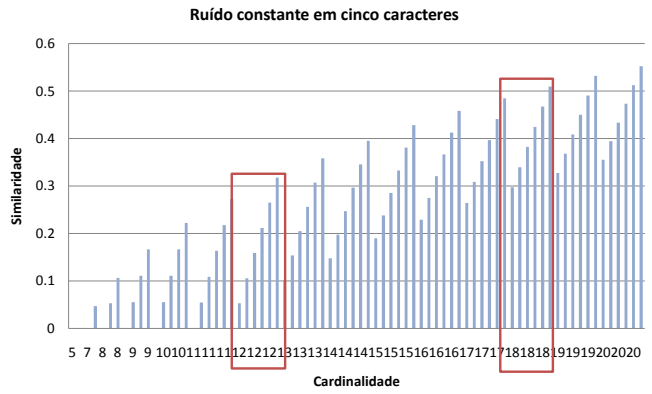

Figure 7: Graph illustrating the spread of the five characters in different noise cardinalities.

Cardinalidade	FIXA	
Ruído	>	<
Similaridade	<	>
Confiança do relacionamento	<	>

Figure 8: Sintese of relationships between variables, cardinality, noise and similarity.

Amostra	*Similaridade*	*Cardinalidade*
RUA GENERAL CORREIA ***DE*** CASTRO JARDIM AMÉRICA RUA GENERAL CORREIA E CASTRO JARDIM AMÉRICA	0.946	44
RUA ANTONIO TIBURCIO DE SOU***Z***A ANCHIETA RUA ANTONIO TIBURCIO DE SOUSA ANCHIETA	0.909	38
RUA GENERAL GAL***L***IENI BONSUCESSO RUA GENERAL GALIENE BONSUCESSO	0.856	31

Figure 9: Geographical Names ratio to 0.9 similarity.

of the triangular table cardinality for up to twenty characters can estimate values of any cardinality.

Taxonomia noise in geographical names

From the maximum and minimum intervals listed in the triangular matrix available in the appendix, were listed below some noise patterns studied from simulated data. This use of the triangular matrix provides support for the user has the conditions to infer situations expected in similarity query, i.e., quantify how flexible is the response pattern in the query by NG.

Figure 8 summarizes the relationship between strategy records similarity, considering a cardinality of a NG any fixed, the higher the amount of noise, the lower the similarity index, then lower confidence for a relationship. Since the smaller the amount of noise, the higher the similarity index and increased confidence to the linkage between records. So, starting to break similarity, I=[0.85,9.0], with the interference of a lower noise 5% (relative to EQ.5 expression).

In this first pattern is underscored a subtle difference for writing terms in Figure 9. To achieve this level of similarity, the smaller the amount of noise and higher cardinality only the variable similarity is sufficient to rely on the relationships between data.

In response patterns with 0.81 similarity, with a percentage of noise 15 to 20% (relative to eq.5 expression), a new pattern in which the marker relates similarity incomplete NG. Exemplified in Figure 10 the track in question contains the name “PARQUE” or “VILA”, even if present in only one of the records the relationship of both is possible. However, this range covers the need to know one more variable in addition to the similarity, because it just is not enough to rely on the relationship between the NG.

For the range of values, I=[0.76, 0.70], it is essential to consider the size of the string and noise. In the examples of Figure 10, it is observed that for this amount of similarity with smaller cardinalities that the example of Figure 11 shows how much diminishes the trust between relationships. In the meantime the cardinality variables and similarity reveal another kind of relationship, one to many: a record of the IPP database table happens to be related to more than one record of the place names of CNEFE table. This limit is unfeasible the decision of relationship, because trust even trimmed the similarity variables,

Amostra	*Similaridade*	*Cardinalidade*
RUA ENGENHEIRO ARMINDOR ANGEL ANCHIETA RUA ENGENHEIRO ARMINDOR ANGEL ***PARQUE*** ANCHIETA	0.816	38
RUA CAIRUCU VALQUEIRE RUA CAIRUCU ***VILA*** VALQUEIRE	0.814	21

Figure 10: Geographical Names ratio to 0.8 similarity.

Amostra	*Similaridade*	*Cardinalidade*
RUA J SANTA CRUZ RUA G SANTA CRUZ RUA K SANTA CRUZ	0.769	16
RUA DO ALHO PENHA CIRCULAR RUA DO ARROZ PENHA CIRCULAR	0.702	26

Figure 11: Geographical Names similarity ratio to 0.7.

Amostra	*Similaridade*	*Cardinalidade*
RUA GUACUPI COELHO NETO RUA ***ACEGUA*** COELHO NETO	0.659	23
RUA ALBERTO CAVALCANTI ***RIO DE JANEIRO*** RUA ALBERTO CAVALCANTI ***RECREIO DOS BANDEIRANTES***	0.605	35
RUA ***ADOLFO BERGAMINI*** ENGENHO DE DENTRO RUA ***DOUTOR LEAL*** ENGENHO DE DENTRO	0.500	35

Figure 12: Geographical Names ratio to less than 0.7 similarity.

cardinality and noise are low.

For range of values below 0.65 to the limit of the similarity metric 0.5, we find the following: cardinality and low similarity with loud noise which rules out any relationship.

To cardinality and high noise ratio, with low similarity, the relationship of the sought records is possible as long as analyzed over an external variable, its geographical position (Figure 12).

Discussion

In short, when dealing with smaller similarity values from 0.6 to balanced recovery of their records can be performed since the noise is considered to be greater than eight characters. Because even with low sensitivity and correct classification rate (43%), their low percentage of error for incorrect classifications (6%) minimizes the risk of erroneous classification.

Before concluding the hypothesis that there is a threshold of similarity between Geographic Names, the experiments show that it is necessary to consider the position occupied by the noise in the string. In addition, one must wonder if these noises are clustered or dispersed. This renders a high degree of relevance for the acceptance scale value of similarity in the tested databases. To set a lower limit based on experiments and complete a minimum of similarity can be considered two parameters: The first parameter is observed in Figures 4 and 11. At the point of minimum similarity of a chain of twenty characters, the maximum amount of noise that the indicator measures is equivalent to 50% of the cardinality of the string. The second point, to set a minimum limit, is analysis of different cardinality sizes from the Geographic Names. A string cannot be twice the size of your relational pair, as well as consider this difference in cardinality should pay attention to the noise between them. Example:

I (Avenida Guilherme Maxwell, Rio de Janeiro, AvG Macwell, Rio de Janeiro)=0.507 Therefore, the maximum of 62% difference considers altercation cardinality and/or noise.

For strings of the same cardinality, the number of possible results considering the variation of the noise position is the amount of noise plus one. This property is important for future shows is value mappings.

Conclusion

The verification of the applicability of the proposed indicator to assess the recovery of records in different databases, demonstrated quite effectively for correct classification of NG pairs. The results indicated that the smaller similarity value ranges than 0.9 and greater than 0.8, the use of a variable to refine the retrieval of records retrieved increase confidence in the information. This use provides for greater cardinalities of ten characters, the positive effect to correct positive ratings genuine. For smaller similarity range 0.6 and 0.5 higher, the additional variable noise increases efficiency for correct classification. When considering noise values greater than eight characters, the incorrect classifications index drops to 6%. This shows an asymmetry property with high efficiency for negative impostor classification. Reached the expected goal, the similarity coefficients maximized queries that require textual comparisons, as in the georeferencing process addresses.

References

1. Camara G (2001) Introduction à Ciência da Geoinformação. São José dos Campos, INPE.
2. Bishop ID (1994) The role of visual realism in communicating and understanding spatial change and process. In Hearnshaw HM and Unwin DJ (eds), Visualization in Geographic Information Systems. John Wiley, Chichester.
3. Camara G, Vinhas L, Clodoveu D, Fred F, Tiago C (2009) Geographical information engineering in the 21st century. In: Navratil G (eds), Research trends in geographic information science. Springer, Dordrecht pp: 203-218
4. Camara G (2007) Territorios digitais: as novas fronteiras do Brasil. In:Seminário Preparatório para a 3º Conferência Nacional de Ciência e Tecnologia.
5. Camara G (2005) Representação computacional de dados geográficos. Casanova M, Camara G, Davis C, Vinhas L, Queiroz G (ed.) Banco de Dados Geograficos. Mundo GEO, Curitiba, Brasil.
6. Miller HJ (2004) Tobler's first law and spatial analysis. Annals of the Association of American Geographers 94: 284-289.
7. Ferreira, Carine Reis (2006) GeoBrasil: Tutorial sobre Bancos de Dados Geograficos (eds), INPE.
8. Ferreira KR, Casanova MA, Queiroz GR, Oliveira OF (2011) Arquiteturas e linguagens. Livro do curso- BDG.