A Mythology for an Approximate Word Matching: Entropy and Quality

Paulo Coelho Ventura Pinto
Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ
Centro de Tecnologia, Bloco H-319, 21945-970, Rio de Janeiro, RJ
pcoelhopinto@{cos.ufrj.br, gmail.com}

Luis Alfredo Vidal de Carvalho
Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ
Centro de Tecnologia, Bloco H-319, 21945-970, Rio de Janeiro, RJ
alfredo@cos.ufrj.br, luisalfredo@ufrj.br

ABSTRACT

The qualitative momentum, proposed herein, is used for designing a single perceptron neuron whose purpose is operate as system of checks and balances of uncertainties, these measured from similarities and dissimilarities from a pair of western European words, in order to match (or not) the pair of words. A metric of how doubtful is the classification itself is also presented. The mechanism to generate myths as counting problems to yield uncertainty is also explained and interpreted. It is established a relation between a comprehensive concept of quality and the classic Gaussian distribution: suggesting a strong connection between qualitative assessments and Shannon's information theory. This relation is also herein defined as differential quality. The research results suggest that qualitative momentum may have strong potential for improvements on record linkage techniques, for example, and on the approach of processes of qualitative nature, such as decision making problems and qualitative artificial intelligence.

KEYWORDS. Approximate String Matching, Data Quality, Artificial Intelligence.

Main Area: Other Application in OR.

1. Introduction

It is undeniable that humanity is living in the age of data science. In text mining, identifying morphological or semantic similarities of words to obtain useful information from the text content to issue value judgments is both a technical challenge as well as of scientific interest – for the ability to handle and understand a content that sometimes embodies shades of a particular culture. In the Internet social networks, an ideal intelligent agent would have to be able to deal with ambiguity, noise and even idiosyncrasies of a target group – otherwise its designers have to give up and build one that only tackles the problem in order to grasp some useful informations . In social network analysis, monitoring the life cycle of a relationships between entities (from its start up to its end) may not be easily done since such information may be signaled by unstructured data. So this information may be not readily available for immediate processing by natural language systems. Obviously, on a statistical analysis in which enormous sampling is feasible, a marginal technical improvement might not reflect in any significant leap as far as the achieved knowledge and insights are concerned. However, there are situations where the noise itself, or the odd, is expected to be unveiled: a fraudulent transaction or terrorist actions, for instance. In such cases an average has butterfingers. So the state of the art, skills and huge financial resources must be gathered together. Usually only large private institutions have the means for such an undertaking. In other cases, besides the means, only governments have the legitimacy to carry them out.

Regardless of practical needs, even a technology may have its limits. These may be sometimes predicted by the theory that rules the problem's application field for which such technology was meant. In other situations, not even a technique exists: the problem remains untouched. However, from time to time, the application of a scientific concept in one field may pave the way for technical development in another. So where not even a single possibility does exist, glimmers of uncertainty may be turned into glimpses of opportunity and may end a past of blindness – even if the mist that blurs the sight fades away bit by bit.

2. Motivation

The concept of quality is tricky to grasp even in literature or in a dictionary. It appears in different contexts and fields of knowledge. In the productive environment of goods over decades up until the 1950s, the concept of product quality was strictly linked to technical perfection (Costa et al., 2008). In the *ISO 9000:2005*, quality is the "degree to which a set of inherent characteristics fulfils requirements". So quality was also extended into the degree of user satisfaction with products and services. Where data is a corporate asset, data quality may be seen as adequacy of the data for use in the successful implementation of business processes of a company (DAMA, 2009).

In the context of pattern recognition for qualitative assignment of numerical scores, Guil and Marín (2013) use the Shafer's Theory of Evidence to derive metrics to quantify the degree of PERCEIVED quality of pattern in a prescriptive fashion. Aggarwal and Yu (2009) point out that the uncertainty has become an integrant part of data for some data sources: a challenge to data mining to deal with inherently uncertain data.

In a process of identity resolution (PIR), the quality of data from different sources can be a determinant factor for its effectiveness and efficiency. It involves both gathering data and the discovery of new knowledge, or facts, about the entities that are the real source of this data – not only data sources: the real people. As particular case of PIR, record-linkage is the gathering and comparison of file records which share some common attributes. Based on these attributes, a record-linkage technique should decide whether a set of records are related, or not, to the same entity in the real world. Examples of improved and classic techniques of matching and record-linkage can be seen in (Bilenko and Mooney, 2002) and (Herzog et al., 2007).

In the Brazilian governmental sphere, Pinto et al. (2013) and Pinto and Carvalho (2014) address the problem of record linkage, cross-checking data and data quality between personal data records from governmental nationwide databases, in particular: the records of health insurance contractual data in the Beneficiary Information System under the jurisdiction of National Regulatory

Agency for Private Health Insurance (Brazil's Ministry of Health) and the records of taxpayers in Natural Persons Register under the jurisdiction of Department of Federal Revenue of Brazil (Brazil's Ministry of Finance). In the Project of Restructuring the Registry of Beneficiaries of Health Insurances and Plans, carried out between the years 2008 and 2011, whose efforts, mainly and especially from 2010, resulted in the identification of 59,7 millions records of private health insurance beneficiary – due to their consistency to taxpayer information checked by a deterministic methodology and, as a side effect, the Brazilian government was able to assigned the National Health Card Identification Number to approximately 31 million of private health insurance beneficiaries in the Brazil (Pinto and Carvalho, 2014). In the *Brazilian operational research*, Pinto and Carvalho (2014) also suggest that Shannon's information theory may lead to a better understanding of the qualitative nature of record-linkage alike processes. They formulate the *deontic-epistemic dilemma*[1]: the existence of "conflicts between the facts and an intelligent agent's beliefs", especially, when comparing personal information such as ID numbers; names and dates of birth from data records from different jurisdictions.

3. Objectives

Primary Objective: to present a strictly theoretical model that employs Shannon's entropy to combine metric distance between strings of characters - in particular those representing western European words in the Latin alphabet - to obtain a single neuron perceptron with effectiveness to classify similar pairs of strings (machted) and dissimilar pairs of strings (not-machted). So an approach to a particular kind of approximate string matching problem. Secondary Objectives: to argue that entropy measures can be a raw material of a decision processes - not only a quantity to be maximized or minimized as an optimization parameter. To propose a magnitude that suggests a connection between a concept of quality and uncertainty by probing the coherence among empirical prescriptions, mathematical relationships and Shannon's information theory.

4. Methodology

Discuss briefly the classical concept of Shannon's information entropy. Explain the process of myth variant generation. Give evidence that the weighting of entropies is no novelty, but tacitly present in scientific literature. Weight these myth variations using the concept of qualitative momentum and use a perceptron as decision unit to solve the approximate word matching. Tabulate some pairs of English and Portuguese words for comparison purposes as well as their matching results, a base line metric and other data of interest.

5. Basic Definitions

$\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ are, respectively, the natural (including zero), integer and real number sets. A string σ is a finite sequence of symbols belonging to a set called alphabet. If a string does not have symbols, it is denoted by the symbol ϵ (empty string). $|\sigma|$ denotes the length of a string σ – total number of alphabet symbols that makes up the string. Let σ, σ_1 and σ_2 be strings. Let x and y be such that $x, y \in \mathbb{R}$. $L(\sigma_1, \sigma_2)$ denotes the Levenshtein distance (or edit distance) between σ_1 and σ_2, which is the minimum number of edit operations (insertions, deletions or substitutions) needed to transform σ_1 into σ_2, or vice versa. $\sigma[i]$ denotes the i-th letter symbol (from left to right) of the string σ. If $i < 1$ or $|\sigma| < i$ then $\sigma[i]$ is not defined. The function $max(x, y)$ denotes the maximum number of x and y. The function $min(x, y)$ denotes the minimum number of x and y. The function $I(\sigma_1, \sigma_2)$ is the equality number between σ_1 and σ_2 and is defined as $I(\sigma_1, \sigma_2) = max(|\sigma_1|, |\sigma_2|) - L(\sigma_1, \sigma_2)$. $\Phi(z)$ is an algorithm that returns the absolute value of z, if z is a number, otherwise returns ∞. See example 5.1. $Sx(\sigma_1)$ is soundex code of σ_1. Such code is also a string. Black (2010) explains that soundex code will be formed by first letter of string σ_1

[1] Pinto and Carvalho (2014) studied role of this dilemma especially when an intelligent agent is facing up the decision whether to link, or not to unlink, two personal data records whose data are believed to concern the same entity in the real world. Although the two under assessment records are necessarily from two governmental data sources of different jurisdictions.

composed by three digits, where each digit corresponds to one of six consonant sounds of σ_1. Black (2010) also states that the soundex code was developed for the matching problem due to different spellings of people's names in US census records and points out it works best for European names. Examples of soundex code[2] obtained from Portuguese and English words are in table 3 (section 12).

Example 5.1. Let σ_1, σ_2 and σ_3 be strings such that $\sigma_1 = abaabcba$, $\sigma_2 = \epsilon$ and $\sigma_3 = c$ Then $\sigma_1[6] = c$, $|\sigma_1| = 8$, $|\sigma_2| = 0$, $|\sigma_3| = 1$, $max(|\sigma_1|, |\sigma_3|) = 8$, $min(|\sigma_1|, |\sigma_2|) = 0$, $L(\sigma_1, \sigma_3) = 7$, $I(\sigma_1, \sigma_1) = 8$, $I(\sigma_1, \epsilon) = 0$, $I(\sigma_1, \sigma_3) = 1$, $\Phi(10) = 10$, $\Phi(-3) = 3$ and $\Phi(\frac{-10}{0}) = \infty$.

Definition 5.1. Let σ_1 and σ_2 be strings. The difference indicator function is defined as:

$$u(\sigma_1, \sigma_2) = \begin{cases} 0 & \textbf{if } L(\sigma_a, \sigma_b) = 0 \\ 1 & \textbf{otherwise} \end{cases} \tag{1}$$

u function will be useful in the calculation of a degree of uncertainty for the sole fact that an intelligent agent perceives that $\sigma_a \neq \sigma_b$.

Definition 5.2. Let σ_1 and σ_2 be strings. The normalized Levenshtein distance is defined as:

$$L_n(\sigma_1, \sigma_2) = L(\sigma_1, \sigma_2)/max(|\sigma_1|, |\sigma_2|). \tag{2}$$

Definition 5.3. $N(\mu, \sigma^2)$ denotes a normal Gaussian distribution with mean μ and variance σ^2.

6. A Baseline Quality Metric between Strings

In the field of data quality, Heinrich et al. (2007) say that quantifying data quality is essential for planning quality measures in an economic manner. Montgomery (2009) proposes three properties that characterize the concept of quality in the production of goods and services. These three properties are below:

Property 6.1. Quality means fitness for use.

Property 6.2. The quality is inversely proportional to variability.

Property 6.3. The quality improvement processes are characterized by the reduction of variability.

Definition 6.1. The metric for correctness function between two strings σ_1 and σ_2 is defined as (Heinrich et al., 2007):

$$Q_c(\sigma_1, \sigma_2) = 1 - L_n(\sigma_1, \sigma_2) \tag{3}$$

The function (3) quantitatively expresses a degree similarity between σ_1 and σ_2 that may also be consistent with a qualitative judgement made due to a visual inspection of this pair of strings. if $Q'_c(\sigma_1, \sigma_2) = 1$ it is a match such that $\sigma_1 = \sigma_2$; if $Q'_c(\sigma_1, \sigma_2) = 0$ then it is undoubtedly a not-matched, for $I(\sigma_1, \sigma_2) = 0$. if $0 < Q'_c(\sigma_1, \sigma_2) < 1$ then it is consistent with a growing degree of quality from 0 up to 1 – in which it is considered both the number of editing operations and the size of the strings. Interestingly, the first attempt of Heinrich et al. (2007) was the following one:

$$Q'_c(\sigma_1, \sigma_2) = \frac{1}{L(\sigma_1, \sigma_2) + 1} \tag{4}$$

Although Heinrich et al. (2007) do not mention (Montgomery, 2009), it is not hard to see that the function (4) tries to mimic the property 6.2, since more different characters, or more missing ones, are perceived between two strings, so the greater the perceived variability is, then the quality metric should be lower in inversely proportional fashion. On the other hand, this function was excluded by the following experimental fact: $Q'_c(Eissonhour, Eisenhower) = Q'_c(Bird, Hunt) = 0.200$. This fact shows that this function is not in compliance to property 6.1. And it is was also ruled out by Heinrich et al. (2007) for its obviously lack of effectiveness. The table 3 lists some values for the function (3) when it is applied to other pair of words.

[2]In this work $Sx(\epsilon)$ results in the string ????, otherwise it behaves according to org.apache.commons.codec.language (*version Soundex.java 1429868 2013-01-07 16:08:05Z ggregory*).

7. Counting Myths and their Variations

An intelligent agent may conjecture while comparing data why he should accept, or not, them based on system of checks and balances of uncertainties instead of using logic – in the sense of safeguard the truth – for the agent does not have access to the historical records of the process. Such historical records would track back to the real world source that would explain the changes undergone by the data. Instead he may create "new histories" or myths. In classical studies a myth is a creation account in a narrative style that tries to explain "how something that is so, but was not the case, began to be" - as "new reality" because of the forgetting of "true history" (Brandão, 2011).

Definition 7.1. A minimal sequence of operations for σ_1 and σ_2, it is a sequence of editing operations to transform the string σ_1 in σ_2, or σ_2 in σ_1, such that the number of operations is the smallest. So $L(\sigma_1, \sigma_2)$ is number of operations for such sequence.

Definition 7.2. A maximal sequence of equalities for σ_1 and σ_2, it is a sequence of equalities operations between characters of strings, respectively, σ_1 and σ_2 in which the number of equalities is the largest and there is a bijection $f(i)$ such that for every $\sigma_1[i] = \sigma_2[j]$ from this sequence then $j = f(i)$. So $I(\sigma_1, \sigma_2)$ is the largest number of equalities for such sequence.

Definition 7.3. $\Omega^-(\sigma_1, \sigma_2)$ is the number of permutations of a minimal sequence of operations for σ_1 and σ_2. Then $\Omega^-(\sigma_1, \sigma_2) = L(\sigma_1, \sigma_2)!$.

Definition 7.4. $\Omega^+(\sigma_1, \sigma_2)$ is the number of permutations of a maximal sequence of equalities for σ_1 and σ_2. Then $\Omega^+(\sigma_1, \sigma_2) = I(\sigma_1, \sigma_2)!$.

Example 7.1. To transform the string TIP into PIT it is required at least two distinct editing operations (denoted by op_1 and op_2), so 2! (two factorial) possibilities:

Permutations of op_1 and op_2
$TIP \xrightarrow{op_1} TIT \xrightarrow{op_2} PIT$
$TIP \xrightarrow{op_2} PIP \xrightarrow{op_1} PIT$

Table 1: Permutations of a minimal sequence of operations for TIP and PIT.

Example 7.2. The number of permutations of a maximal sequence of equalities to $\sigma_1 = back$ and $\sigma_2 = bak$ is 3! (six) permutations.

Permutations of $(\sigma_1[1] = \sigma_2[1], \sigma_1[2] = \sigma_2[2], \sigma_1[4] = \sigma_2[3])$
$(\sigma_1[1] = \sigma_2[1], \sigma_1[2] = \sigma_2[2], \sigma_1[4] = \sigma_2[3])$
$(\sigma_1[1] = \sigma_2[1], \sigma_1[4] = \sigma_2[3], \sigma_1[2] = \sigma_2[2])$
$(\sigma_1[2] = \sigma_2[2], \sigma_1[1] = \sigma_2[1], \sigma_1[4] = \sigma_2[3])$
$(\sigma_1[2] = \sigma_2[2], \sigma_1[4] = \sigma_2[3], \sigma_1[1] = \sigma_2[1])$
$(\sigma_1[4] = \sigma_2[3], \sigma_1[1] = \sigma_2[1], \sigma_1[2] = \sigma_2[2])$
$(\sigma_1[4] = \sigma_2[3], \sigma_1[2] = \sigma_2[2], \sigma_1[1] = \sigma_2[1])$

Table 2: Permutations of a maximal sequence of equalities for $back$ e bak.

Definition 7.5. An intelligent agent that can know the result of an equality between two strings may also believe that one of the strings is the original source of the other string. He may consider exclusively one possibility ($\sigma_1 = \sigma_2$); or two possibilities (σ_1 was transformed into σ_2 or σ_2 was transformed into σ_1). $\Omega(\sigma_1, \sigma_2) = 2^{u(\sigma_1, \sigma_2)}$ stands for this counting.

By construction, Ω^+ endorses similarities between strings. Similarly, Ω^- endorses dissimilarities. Ω is a difference detector based on the supposition of pre-existing communication process. Because Ω^+ and Ω^- are factorial they are sensitive to the size of strings. These three functions are the raw material for measuring uncertainties to be weighted in the string matching process. The values of Ω^+, Ω^- and Ω can be used for corroboration, or counteraction, and can be also weighed in a process of acceptance (or rejection). Such processes have rhetorical nature without being irrational. Each permutation of a sequence is analogous to a variation of the same myth.

8. The Shannon's Entropy as Uncertainty Measure

Definition 8.1. Let $0 < m \in \mathbb{N}$. P is a sequence of m discrete probability distribution functions. Then $P[i]$ is the i-th discrete probability distribution function of P. $P[i]_j$ is the value of the j-th probability of $P[i]$. For $1 \leq n_i \in \mathbb{N}$, if n_i is the cardinality of the domain of $P[i]$ then $\sum_{j=1}^{n_i} P[i]_j = 1$. If the results of the experiment $P[i]$ are equally likely then, for the sake of simplicity, $P[i] = 1/n_i$.

Definition 8.2. p is a single discrete probability distribution function with a domain of n outcomes. Then p is a sequence of one discrete probability distribution function, where $p_i = P[1]_i$ and $n = n_1$. If the results are equally likely then, for the sake of simplicity, $p = 1/n$.

In (Shannon, 1948), choice means the same as discrete probability distribution, since the outcome itself of probabilistic experiment is irrelevant to calculate uncertainty. However when Shannon (1948) uses the term outcome it is simply to indicate the cardinality of the domain of the probability distribution function. So discrete probability distribution function and choice are interchangeable herein. Shannon (1948) formulated the entropy of a choice p with n possible outcomes, denoted by H, as an "uncertainty measure" that should satisfy the three properties below:

Property 8.1. "H should be continuous in the p_i".

Property 8.2. "If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events".

Property 8.3. "If a choice be broken down into two successive choices, the original $H(p)$ should be the weighted sum of the individual values of H."

Theorem 8.1. Let $k \in \mathbb{R}$ and $k > 0$. The only family of functions satisfying the properties 8.1, 8.2 and 8.3 is (Shannon, 1948):

$$H(p) = k \sum_{i=1}^{n} p_i \log(1/p_i) \tag{5}$$

Corollary 8.1. In (Sethna, 2011) a special case of the function (5) where $p = 1/n$ is called counting entropy, denoted herein by H_c.

$$H(p) = k \sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{\frac{1}{n}}) = k \log(n) = k \log(1/p) = H_c(p) \tag{6}$$

The constant k is related to the changing in the base logarithmic function and of establishing a unit to entropy measurement. For $k = 1$ and logarithm base 2, the unit of entropy is called bit (Shannon, 1948) and (Sethna, 2011). Both values are herein the standards.

Example 8.1. Let p, p', p'' and p''' be choices. p'' e p''' are two successive choices corresponding to p' as far as the overall probabilities of the outcomes A_1, A_2 and A_3 are concerned and the outcome A' is a precondition to reach p''' from p''. Also if p'' e p''' are repeated over and over expecting the final outcomes A_1, A_2 and A_3 (A' is not taken into account as final outcome), $2/3$ will be the weight for the entropy of p''' and the weight for entropy of p'' is 1 as far as uncertainty is concerned. So it is to say that p' was broken in p'' and p'''. The trees in figure 1 depict the relation among p'' and p''' and also provide further information about these and the other choices.

Property 8.2 is quite intuitive: more equiprobable possibilities, the greater shall be the uncertainty about the choice: $\log(2) = H(p) < H(p') = \log(3)$. Property 8.3 ensures that $H(p') = \log(3) = H(p'') + \frac{2}{3}H(p''')$. The understanding lies in modifying a sequence of choices

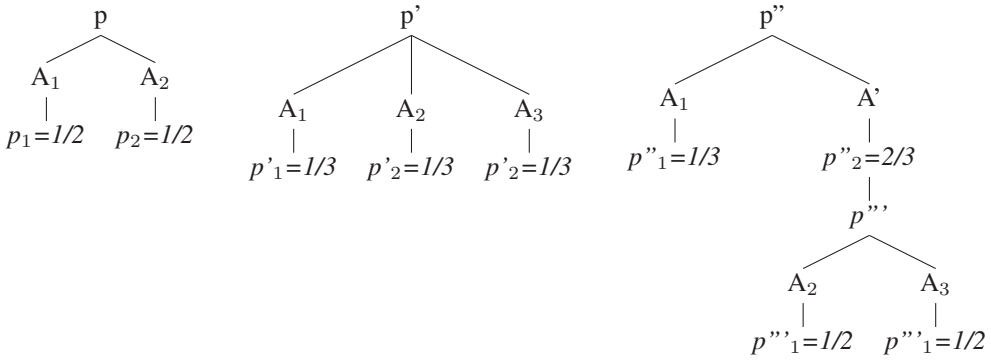

Figure 1: Four choices in tree-like representation: p, p', p'' e p'''.

and their local uncertainty, while conserving final uncertainty since more choices by themselves do not entail less nor more uncertainty. So the properties of Shannon's entropy[3] [4] provide a way to measure a quantity that may be conserved in some processes involving choices.

9. The Qualitative Momentum

Definition 9.1. Let P a sequence of m discrete probability distribution functions (or choices) and W a sequence of m real numbers $(w_1, ..., w_i, ..., w_m)$. The qualitative momentum $Q(P, W)$ is the weighted uncertainties as following:

$$Q(P, W) = \sum_{i=1}^{m} w_i \, H(P[i]) \tag{7}$$

Unlike Shannon entropy, $Q(P, W)$ may assume negative values. It can be even zero. The immediate motivation for the definition of the qualitative momentum is that in probability problems involving the equiprobable event counting, the counting entropy of the original problem can often be rewritten as multiples of $\log(n)$ or by adding or subtracting terms like $\log(n!)$, where $n \in \mathbb{N}$. That suggests that the uncertainty of a counting problem can be interpreted as the weighting of the uncertainties due to other counting problems. The second motivation is provided in section 10.

Example 9.1. S_r^n denotes the number distinct sequences of r selected elements each at a time from n elements with replacement. Then $\mathsf{S}_r^n = n^r$. The counting entropy of $p = \frac{1}{\mathsf{S}_r^n}$ is the following:

$$log(\mathsf{S}_r^n) = \log(n^r) = r \cdot \log(n) = \sum_{i=1}^{r} \log(n) \tag{8}$$

Example 9.2. P_n denotes the number distinct sequences of n distinct elements. Then $\mathsf{P}_n = n!$. C_k^n denotes the number of distinct sets of k elements selected from n distinct elements. Then $\mathsf{C}_k^n = \frac{n!}{k!(n-k)!}$. The counting entropy of $p = \frac{1}{\mathsf{C}_k^n}$ is the following:

$$log(\mathsf{C}_k^n) = log(\frac{n!}{k!(n-k)!}) = log(\mathsf{P}_N) - log(\mathsf{P}_K) - log(\mathsf{P}_{N-K}) \tag{9}$$

In the example 9.2, the negative signed terms are linked to decrease in the counting entropy when growing in absolute figures and the positive signed term are linked to increase in the counting entropy when growing in absolute figures. Even the share of the term $-log(\mathsf{P}_{N-K})$, in which $N - K$ can be considered a pseudo parameter A (if $A = N - K$), can be taken into account in that way since increasing values of A contribute to the decrease in the net counting entropy.

[3] See the classic (Verdú, 1998) for an historical perspective on the development of the concept of entropy and further information.

[4] So Shannon's entropy is not a fancy measurement function with a bunch of logarithms that appeared out of thin air.

The rightmost expressions of the equations (7), (8) and (9) have an analogous formal characteristic: a weighted sum of entropies. However, (7) can never be considered a generalization of Shannon entropy, equation (5), because a qualitative momentum can be negative. But the expression (7) may allow us to envision a mathematical model in which possibilities can counteract and catalyze one another in a decision process – but not annihilate themselves.

Corollary 9.1. Let $Q_1(P_1, W_1)$ and $Q_2(P_2, W_2)$ be qualitative momenta and $a_1, a_2 \in \mathbb{R}$. Then $a_1 \cdot Q_1(P_1, W_1) + a_2 \cdot Q_2(P_2, W_2)$ is also a qualitative momentum.

10. Differential Entropy, Qualitative Momentum and Quality

Here the concept of differential entropy is used to suggest a relationship between the differences of entropy and the properties of quality suggested by Montgomery (2009), see section 6.

Definition 10.1. Let $\mathbb{X}$ be a continuous random variable with the probability density function (PDF) $p_{\mathbb{X}}$. The differential entropy of $\mathbb{X}$ is (Shannon, 1948) and (Haykin, 2009):

$$H_{dif}(\mathbb{X}) = \int_{-\infty}^{+\infty} p_{\mathbb{X}}(x) \log(\frac{1}{p_{\mathbb{X}}(x)})\, dx \tag{10}$$

Example 10.1. The differential entropy of $N(\mu, \sigma^2)$ (Shannon, 1948):

$$H_{dif}(N(\mu, \sigma^2)) = \log(\sqrt{2\pi e} \cdot \sigma) \tag{11}$$

Definition 10.2. Let $\mathbb{X}$ be a continuous random variable with PDF $p_{\mathbb{X}}$. H_x is a discretization process ruled by x to subject $\mathbb{X}$ then $p_k = p_{\mathbb{X}}(x_k) \cdot \Delta x$ and the domain of $p_{\mathbb{X}}(x)$ is divided into an countably infinite number of intervals (x_k, x_{k+1}) such that $x_k = k \cdot \Delta x$ and $\Delta x \to 0^+$ for all $k \in \mathbb{Z}$, such that (by analogy to equation (5)):

$$H_x(\mathbb{X}) = \lim_{\Delta x \to 0^+} \sum_{k=-\infty}^{+\infty} p_k \cdot \log(\frac{1}{p_k}) \tag{12}$$

In (Haykin, 2009):

$$\lim_{\Delta x \to 0^+} \sum_{k=-\infty}^{+\infty} p_k \cdot \log(\frac{1}{p_k}) = \int_{-\infty}^{+\infty} p_{\mathbb{X}}(x) \log(\frac{1}{p_{\mathbb{X}}(x)})\, dx + \lim_{\Delta x \to 0^+} \log(\frac{1}{\Delta x}) = H_{dif}(\mathbb{X}) + \lim_{\Delta x \to 0^+} \log(\frac{1}{\Delta x}) \tag{13}$$

Haykin (2009) points out that the term $\lim_{\Delta x \to 0^+} \log(\frac{1}{\Delta x})$ can be considered as a mere reference because the entity of interest in stochastic systems studies is the entropy difference between two terms with the same reference. However the concept of the x ruled processes, proposed herein, avoids the disadvantage of ignoring $\lim_{\Delta x \to 0^+} \log(\frac{1}{\Delta x})$ as a residue when time-like parameters are not a concern, but does not at all exclude time-like parameters from being object of analysis in other studies.

Example 10.2. H_x is a discretization process ruled by x to subject $\mathbb{X}$ and $\mathbb{Y}$. Then:

$$H_x(\mathbb{X}) - H_x(\mathbb{Y}) = H_{dif}(\mathbb{X}) - H_{dif}(\mathbb{Y}) + \lim_{\Delta x \to 0^+} \log(\frac{1}{\Delta x}) - \lim_{\Delta x \to 0^+} \log(\frac{1}{\Delta x}) = H_{dif}(\mathbb{X}) - H_{dif}(\mathbb{Y}) + \lim_{\Delta x \to 0^+} \log(\frac{\Delta x}{\Delta x}) = \tag{14}$$

$$= H_{dif}(\mathbb{X}) - H_{dif}(\mathbb{Y}) + \lim_{\Delta x \to 0^+} \log(1) = H_{dif}(\mathbb{X}) - H_{dif}(\mathbb{Y}) \tag{15}$$

Example 10.3. H_x is a discretization process ruled by x to subject $N(\mu_a, \sigma_a^2)$ and $N(\mu, \sigma^2)$. Then:

$$H_x(N(\mu_a, \sigma_a^2)) - H_x(N(\mu, \sigma^2)) = H_{dif}(N(\mu_a, \sigma_a^2)) - H_{dif}(N(\mu, \sigma^2)) = \log(\frac{\sigma_a}{\sigma}) = \log(\sigma_a) - \log(\sigma) \tag{16}$$

The concept differential entropy, a well-established concept in literature, necessarily depends on the difference between entropy measures from at leat two PDFs to be mathematically meaningful. Example 16 is enlightening: the difference between uncertainties arising from normal gaussian distribution is connected to the standard deviation of these distributions and is independent of their means. Furthermore, depending on the values of σ_a and σ the difference between differential entropies can be positive, negative or zero, respectively for $\frac{\sigma_a}{\sigma} > 1$, $\frac{\sigma_a}{\sigma} < 1$ e $\frac{\sigma_a}{\sigma} = 1$.

At the end of the section 8, it was reported that the choice of logarithmic base to calculate entropy is just a matter of choosing the entropy unit of measurement. To use the natural logarithm, it suffices to multiply the Shannon entropy for $\frac{1}{log(e)}$. Let c be a constant such that $c = \frac{k}{log(e)}$ and $0 < k \in \mathbb{R}$ and multiplying the equation (16) by c then:

$$k \cdot \frac{H_{dif}(N(\mu_a, \sigma_a^2)) - H_{dif}(N(\mu, \sigma^2))}{log(e)} = k \cdot \ln(\frac{\sigma_a}{\sigma}) \tag{17}$$

From equation (17), where $c = \frac{k}{log(e)}$ and

$$\lim_{\sigma_a \to \sigma} k \cdot \frac{H_{dif}(N(\mu_a, \sigma_a^2)) - H_{dif}(N(\mu, \sigma^2))}{log(e)} = \lim_{\sigma_a \to \sigma} k \cdot \ln(\frac{\sigma_a}{\sigma}) \tag{18}$$

then it follows the *diferencial* equation:

$$c \cdot \, dH_{dif} = \frac{k}{\sigma} \, d\sigma \tag{19}$$

Definition 10.3. Q'_{dif} is a differential quality if, only if, Q'_{dif} is a quantity such that $dQ'_{dif} = -c \cdot \, dH_{dif}$. Where $0 < c \in \mathbb{R}$ and c is a constant.

Corollary 10.1. Let be $N(\mu, \sigma^2)$ and $0 < \frac{k}{log(e)} = c \in \mathbb{R}$. A differential quality Q'_{dif} for $N(\mu, \sigma^2)$ obeys:

$$dQ'_{dif} = \frac{-k}{\sigma} \, d\sigma \tag{20}$$

Let $N(\mu, \sigma^2)$ be the model of a process whose standard deviation σ is the metric of variability and mean μ is the standard of fitness-for-use. Globally Q'_{dif} always increases as σ decreases, and vice versa – for $\sigma > 0$ and $k > 0$ always. Locally, σ has an inertial effect on $\Delta Q'_{dif}$ for a small $\Delta\sigma$. Thus, Q'_{dif} is partial consistent with the proposition 6.2 in the sense that Q'_{dif} always decreases with increase in the variability, but Q'_{dif} is not inversely proportional to variability. Moreover, the concept of reducing variability is the concept of quality improvement process, see proposition 6.3. The concept of μ of an normal gaussian distribution would be consistent property 6.1, as a quality process decreases the variability then the greater the likehood of achieving the fitness-for-use vicinities (i.e. reduncing error) as long as there is no μ drifting. So these results suggest that the relationship $dQ'_{dif} = -c \cdot \, dH_{dif}$ is quasi-adherent to the quality properties recommended by Montgomery (2009) as far as $N(\mu, \sigma^2)$ is concerned. If $k = log(e)$ then the unit of Q'_{dif} is the bit since $c = 1$ and differential entropy, equation (10), is also in bit units.

It is noteworthy herein that, in the field of cognitive sciences, the rightmost expressions of equations (17 and 19) are analogous to the negative symmetric of Weber–Fechner law (only from a formal perspective). This observation suggests a possible connection to the differential quality of $N(\mu, \sigma^2)$, entropy and studies on sensorial perception. For further information on the Weber–Fechner law, see (Masin et al., 2009).

11. A Perceptron: Matching a Pair of Words

Definition 11.1. Let $Q(P, W)$, σ_a and σ_a, and w be, respectively, a qualitative momentum, two strings and $w \in \mathbb{R}$ such that $P[1] = 1/\Omega^+(\sigma_a, \sigma_b)$, $P[2] = 1/\Omega^-(\sigma_a, \sigma_b)$, $P[3] = 1/\Omega(\sigma_a, \sigma_b)$, $W = (w, -w, -w)$. $Q(\sigma_a, \sigma_b, w)$ is a qualitative dipole for strings σ_a and σ_b, if $Q(\sigma_a, \sigma_b, w) = Q(P, W)$. $Q(\sigma_a, \sigma_b) = Q(\sigma_a, \sigma_b, 1)$, if w is omitted as a parameter.

By construction, a qualitative dipole is intuitively a metric that weights uncertainties due to the similarities and dissimilarities between two strings and also the fact that they strings are different or the same, as shown by equations (21).

$$Q(P,W) = w\log(\Omega^{+}(\sigma_a,\sigma_b)) - w\log(\Omega^{-}(\sigma_a,\sigma_b)) - w\log(\Omega(\sigma_a,\sigma_b)) = w\log(\frac{\Omega^{+}(\sigma_a,\sigma_b)}{\Omega^{-}(\sigma_a,\sigma_b)\Omega(\sigma_a,\sigma_b)}) \quad (21)$$

As a suggestion, the classic supervised learning algorithm of the perceptron may be employed to search for weights of a qualitative momentum. If the training instances are not linearly separable, a strategy that combines the pocket algorithm and multiple training (the restart with different synaptic weights chosen at random) may be used in that search. The ADALINE can also be used, since there is neither morphological nor functional difference from the perceptron. The sole difference lies in the training algorithm because the first uses the local induced field value and the second the classification value to calculate the error. Both neural learning classification models may allow a synaptic weight search that meets both the classification effectiveness and fine-grained tuning by using low learning rates in the delta rule.

Definition 11.2. Let $n \in \mathbb{N}$, $n > 0$. $x = (x_1, ..., x_i, ..., x_n)$ is an instance, $w = (w_1, ..., w_i, ..., w_n)$ is the vector of synaptic weights and $\Sigma(w, x)$ an induced local field function (ILFF) such that $\Sigma(w, x) = \sum_{i=1}^{n} w_i \cdot x_i$. $\nu(w, x)$ is a single perceptron neuron defined as following:

$$\nu(w,x) = \begin{cases} +1 & \textbf{if } \Sigma(w,x) \geq 0 \\ -1 & \textbf{otherwise} \end{cases} \quad (22)$$

Definition 11.3. Let $Q(\sigma_1, \sigma_2)$ be a qualitative dipole. $\nu_Q(\sigma_1, \sigma_2)$ is a single neuron perceptron such that: x is an instance for σ_1 and σ_2 defined as $x = (Q(\sigma_1, \sigma_2), Q(Sx(\sigma_1), Sx(\sigma_2)), 1)$; w is the vector of synaptic weights such that $w = (\frac{1}{2}, 1, 0)$; $\nu_Q = +1$ e $\nu_Q = -1$ are, respectively, the classification matched, or not-matched, for the pair σ_1 and σ_2; The ILFF of ν_Q is denoted as $\Sigma_Q(\sigma_1, \sigma_2) = \Sigma(w, x)$. $D_Q(\sigma_1, \sigma_2) = \Phi(\frac{1}{\Sigma_Q(\sigma_1,\sigma_2)})$ is the doubt degree (a metric of how doubtful is the classification itself).

Intuitively, $\nu_Q(\sigma_1, \sigma_2)$ weights lexical uncertainties favoring phonetic uncertainties on the classification process: $\Sigma_Q(\sigma_1, \sigma_2) = \frac{1}{2}Q(\sigma_1, \sigma_2) + Q(Sx(\sigma_1), Sx(\sigma_2))$. By the definition 11.1 and the corollary 9.1 then $\Sigma_Q(\sigma_1, \sigma_2)$ is also a qualitative momentum. $D_Q(x, w)$ measures how unsure the perceptron should be about its own classification. Suppose $\Sigma_Q(\sigma_1, \sigma_2) \approx 0$ then in this case minor changes in one of the strings may be crucial to regard the pair either as matched or not matched. if $\Sigma_Q(\sigma_1, \sigma_2) = 0$, although a decision is made, then doubt should be out of scale.

12. Experimenting on Perceptron ν_Q

Table 3 is the result of a small experiment in order to quantitatively record the results of computations on approximate string matching using the concepts engendered in the previous sections. The words in this table are especially from the Portuguese and English languages. Most from English, though. Columns L, L_n, Q_c, Ω^+, Ω^-, Ω, Σ_Q, D_Q and ν_Q are functions of the columns σ_1 and σ_2 (the pair of strings to be classified as matched or not matched), for example: $\Omega^+ = \Omega^+(\sigma_1, \sigma_2)$. In particular, the column Q_c is the metric proposed in (Heinrich et al., 2007). Columns Σ_Q, D_Q and ν_Q are respectively the values of the ILFF, the doubt metric and the perceptron decision according to definition 11.3. Heinrich et al. (2007) did not explicitly give a threshold for acceptance or rejection for Q_c, however the 19^o pair in table 3 was also valued $Q_c = 0.600$ in (Heinrich et al., 2007). The 19^o pair, as far as data quality is concerned, was considered acceptable too. So, as a baseline to the reader, $Q_c \geq 0.600$ is a match, otherwise not.

Both Q_c and ν_Q were effective in not matching pairs whose soundex code were totally different: the 19^o and 22^o pairs. When the pairs were identical, both methods were equally effective

in matching them: 7^o, 13^o, 18^o, 23^o and 36^o. For the pairs mentioned in this paragraph, the lowest and highest values of D_Q were $0,056$ and $0,218$.

For the 9^o pair (*pile* and *mile*), $Q_c = 0,750$ e $\nu_Q = +1$. So it is a false positive for both. Such occurrence was expected since no semantic analysis is made. But it is noteworthy that $D_Q = 0,421$.

From the 31^o up to the 36^o pair, there are only comparisons with a well-know English (or American) name $Jessica$ with its Portuguese homophonic versions, excepted for the 36^o pair which is a matched by lexical equality. For these 6 pairs, ν_Q classifies 5 as matched while Q_c, 4. The 31^o pair might be an outlier because $D_Q = 0,535$ and a higher value was expected. Two plausible computational explanations are the low capacity of the soundex algorithm to deal with vowels; and the employed edit distance algorithm considers an alphabetic letter and its diacritically-marked version[5] as different as two distinct alphabetic letters. In other words: the low ability to create myths figuring diacritic marks and vowel sounds. For 40^o pair ($\sigma_1 = Jessica$ and $\sigma_2 = Celina$) both are names have the same vowel letter sequence (i.e. *e-i-a*) and $D_Q = 0,197$ what is consistent with original purpose of soundex coding algorithm to privilege consonant letters.

At last, for $D_Q \geq 1$, there are four pairs: 8^o, 14^o, 32^o e 39^o. it is noteworthy that either are pairs of homophones, whose spelling is quite different, or of quite misspelt words.

Table 3: **The experimental results: the perceptron ν_Q, the baseline Q_c and other data.**

N^o	σ_1	σ_2	$Sx(\sigma_1)$	$Sx(\sigma_2)$	$\lvert\sigma_1\rvert$	$\lvert\sigma_2\rvert$	L	L_n	Q_c	Ω^+	Ω^-	Ω	Σ_Q	D_Q	ν_Q
1^o	cocho	coxo	C200	C200	5	4	2	0,400	0,600	6	2	2	4,877	0,205	+1
2^o	nós	noz	N200	N200	3	3	2	0,667	0,333	1	2	2	3,585	0,279	+1
3^o	xeque	cheque	X200	C200	5	6	2	0,333	0,667	24	2	2	2,877	0,348	+1
4^o	cacei	cassei	C200	C200	5	6	2	0,333	0,667	24	2	2	5,877	0,170	+1
5^o	cacei	xeque	C200	X200	5	5	5	1,000	0,000	1	120	2	-2,368	0,422	-1
6^o	conselho	concelho	C524	C524	8	8	1	0,125	0,875	5040	1	2	10,235	0,098	+1
7^o	conselho	conselho	C524	C524	8	8	0	0,000	1,000	40320	1	1	12,235	0,082	+1
8^o	ship	cheap	S100	C100	4	5	3	0,600	0,400	2	6	2	0,292	3,419	+1
9^o	pile	mile	P400	M400	4	4	1	0,250	0,750	6	1	2	2,377	0,421	+1
10^o	hear	here	H600	H600	4	4	2	0,500	0,500	2	2	2	4,085	0,245	+1
11^o	replace	replacements	R142	R142	7	12	5	0,417	0,583	5040	120	2	6,781	0,147	+1
12^o	I	Z	I000	Z000	1	1	1	1,000	0,000	1	1	2	1,085	0,922	+1
13^o	I	I	I000	I000	1	1	0	0,000	1,000	1	1	1	4,585	0,218	+1
14^o	I	eye	I000	E000	1	3	3	1,000	0,000	1	6	2	-0,208	4,819	-1
15^o	kill	queue	K400	Q000	4	5	5	1,000	0,000	1	120	2	-4,953	0,202	-1
16^o	weak	week	W200	W200	4	4	1	0,250	0,750	6	1	2	5,377	0,186	+1
17^o	one	own	O500	O500	3	3	2	0,667	0,333	1	2	2	3,585	0,279	+1
18^o	one	one	O500	O500	3	3	0	0,000	1,000	6	1	1	5,877	0,170	+1
19^o	Eissonhour	Eisenhower	E256	E256	10	10	4	0,400	0,600	720	24	2	6,538	0,153	+1
20^o	complement	compliment	C514	C514	10	10	1	0,100	0,900	362880	1	2	13,320	0,075	+1
21^o	complement	Eissonhour	C514	E256	10	10	10	1,000	0,000	1	3628800	2	-16,980	0,059	-1
22^o	complement	Eisenhower	C514	E256	10	10	10	1,000	0,000	1	3628800	2	-16,980	0,059	-1
23^o	Eisenhower	Eisenhower	E256	E256	10	10	0	0,000	1,000	3628800	1	1	15,480	0,065	+1
24^o	fir	fur	F600	F600	3	3	1	0,333	0,667	2	1	2	4,585	0,218	+1
25^o	tip	pit	T100	P300	3	3	2	0,667	0,333	1	2	2	-2,000	0,500	-1
26^o	plain	plane	P450	P450	5	5	2	0,400	0,600	6	2	2	4,877	0,205	+1
27^o	write	right	W630	R230	5	5	4	0,800	0,200	1	24	2	-3,792	0,264	-1
28^o	whole	hole	W400	H400	5	4	1	0,200	0,800	24	1	2	3,377	0,296	+1
29^o	Eissonhour	ϵ	E256	????	10	1	10	1,000	0,000	1	3628800	2	-16,980	0,059	-1
30^o	Bird	Hunt	B630	H530	4	4	4	1,000	0,000	1	24	2	-3,792	0,264	-1
31^o	Jessica	Gécika	J220	G220	7	6	5	0,714	0,286	2	120	2	-1,868	0,535	-1
32^o	Jessica	Gecika	J220	G220	7	6	4	0,571	0,429	6	24	2	0,085	11,770	+1
33^o	Jessica	Jecika	J220	J220	7	6	3	0,429	0,571	24	6	2	5,085	0,197	+1
34^o	Jessica	Jecica	J220	J220	7	6	2	0,286	0,714	120	2	2	7,038	0,142	+1
35^o	Jessica	Jesica	J220	J220	7	6	1	0,143	0,857	720	1	2	8,831	0,113	+1
36^o	Jessica	Jessica	J220	J220	7	7	0	0,000	1,000	5040	1	1	10,735	0,093	+1
37^o	Jessica	Jessca	J220	J200	7	6	1	0,143	0,857	720	1	2	5,831	0,172	+1
38^o	Jessica	Gessca	J220	G200	7	6	2	0,286	0,714	120	2	2	1,453	0,688	+1
39^o	Jessica	Gesca	J220	G200	7	5	3	0,429	0,571	24	6	2	-0,500	2,000	-1
40^o	Jessica	Celina	J220	C450	7	6	4	0,571	0,429	6	24	2	-5,085	0,197	-1

13. Conclusions

The objectives were achieved especially with the handcrafted perceptron, however derived from de qualitative momentum, with decision-making ability to a string approximate matching problem – with a metric of degree of doubt. The results suggest a clear cut amongst uncertainty, doubt and data quality although they may be related. The differential quality is directly proportional

[5]Examples of letters and its diacritically-marked version: *a*/*ã*, *c*/*ç*, *a*/*à*, *e*/*é* and *u*/*ü*.

and symmetrical to the differential entropy and almost captured a set of quality properties as far as processes obeying a classical normal Gaussian distribution are concerned. It is noteworthy that: the bit as a quality metric unit of differential quality and the ubiquity of the Gaussian in many fields of knowledge. These suggest a broad application of the concepts herein. As far as quality is concerned, the Gaussian is also a mathematical cornerstone in both theoretical and experimental studies: from assembly lines to software metrics. Myth counting not only gave another interpretation to a particular case of string matching but it did not restrict the approach to a specific algorithm. Instead it suggests the blending of effectiveness of other algorithms other than soundex, such as metaphone, if they allow counting that can account for similarities and dissimilarities. The results suggest that record linkage may benefit from greater diversities of methods and theories along side with statistics to assess performance which involves cognitive abilities and artificial intelligence.

References

Aggarwal, C. C. and Yu, P. S. (2009). A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):609–623.

Bilenko, M. and Mooney, R. J. (2002). Learning to combine trained distance metrics for duplicate detection in databases. Technical report, Artificial Intelligence Lab - University of Texas at Austin.

Black, P. E. (2010). "soundex", in Dictionary of Algorithms and Data Structures [online], Vreda Pieterse and Paul E. Black, eds. (accessed 4 January 2015) Available from: http://www.nist.gov/dads/HTML/soundex.html.

Brandão, J. S. (2011). *Mitologia grega*. Number vol. 1. Vozes, 23 edition.

Costa, A., Epprecht, E., and Carpinetti, L. (2008). *Controle estatístico de qualidade*. Atlas, 2 edition.

DAMA (2009). *The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK*, chapter Data Quality Management, pages 291–316. Technics Publications, LLC, USA.

Guil, F. and Marín, R. (2013). A theory of evidence-based method for assessing frequent patterns. *Expert Systems with Applications*, 40(8):3121 – 3127.

Haykin, S. (2009). *Neural Networks and Learning Machines*, chapter Maximum-Entropy Learning For Blind Source Separation, pages 479–482. Number v. 10 in Neural networks and learning machines. Prentice Hall, 3rd edition.

Heinrich, B., Kaiser, M., and Klier, M. (2007). How to measure data quality? – a metric based approach. Montreal. 28th International Conference on Information Systems (ICIS).

Herzog, T., Scheuren, F., and Winkler, W. (2007). *Data Quality and Record Linkage Techniques*, chapter 8, pages 81–92. Springer.

Masin, S. C., Zudini, V., and Antonelli, M. (2009). Early alternative derivations of fechner's law. *Journal of the History of the Behavioral Sciences*, 45(1):56–65.

Montgomery, D. (2009). *Introdução ao controle estatístico da qualidade*. LTC.

Pinto, P. C. V. and Carvalho, L. A. V. (2014). Record linkage e conferência de dados: uma avaliação de metodologias de cruzamento de dados cadastrais governamentais. In *XLVI Simpósio Brasileiro de Pesquisa Operacional*, pages 671–682, Salvador, BA, Brasil.

Pinto, P. C. V., Cerceau, R., Mesquita, R., and Carvalho, L. A. V. (2013). Conferência eletrônica de dados cadastrais governamentais por critérios qualitativos. In *IX Simpósio Brasileiro de Sistemas de Informação: trilhas técnicas*, pages 803–814, João Pessoa, PB, Brasil.

Sethna, J. P. (2011). *Entropy, Order Parameters, and Complexity*, chapter 5, pages 82–90. Clarendon Press - Oxford, Laboratory of Atomic and Solid State, Corwell University, Ithaca, NY.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Verdú, S. (1998). Fifty years of shannon theory. *IEEE Transactions on Information Theory*, 44(6):2057–2078.