

DEPARTMENT OF
APPLIED PHYSICS AND ELECTRONICS
UMEÅ UNIVERISTY, SWEDEN



DIGITAL MEDIA LAB

Document distances using the Zipf distribution and a novel metric

Apostolos A. Georgakis¹
Dept. Applied Physics and Electronics
Umeå University
SE-90187, Umeå Sweden
e-mail: apostolos.georgakis@tfe.umu.se

H. Li
Dept. Applied Physics and Electronics
Umeå University
SE-90187, Umeå Sweden
e-mail: haibo.li@tfe.umu.se

DML Technical Report: DML-TR-2003:01

ISSN Number: 1652-8441

Report Date: December 1, 2003

¹This work was supported by the European Union Research Training Network (RTN) "MUHCI: Multi-modal Human Computer Interaction (HPRN-CT-2000-00111).

Abstract

A novel metric is proposed in the present report for the evaluation of the goodness-of-fit criterion between the distribution functions of two samples. We extend the usage of the proposed criterion for the case of the generalized Zipf distribution. Detailed mathematical analysis of the proposed metric, which is embodied in a hypothesis testing, is also provided.

Keywords

Zipf distribution, n-gram frequencies, bhattacharyya metric

1 Introduction

In a plethora of natural phenomena the distribution of a characteristic under consideration is heavily skewed. For example, biological, ecological, and chemical systems sometimes tend to exhibit an exponential decaying model. Web site popularity, web access statistics, Internet traffic, population and growth of cities also comply with the same decaying model. Furthermore, many references can be found in bibliometrics, informetrics and library science. A plethora of distributions exists in the literature that are capable to model the above phenomena; with the most prevalent among them the well-know *Zipf* distribution [1, 3].

The Zipf distribution rely on an empirical law discovered by Estoup in 1916 and named after the Harvard linguistic professor G. K. Zipf (1902-1950). This distribution relates the frequency of occurrence of an event α and the rank, m_α , of the event when the rank is determined by the above frequency of occurrence. The relationship is the power-law function:

$$P(\alpha) \sim 1/m_\alpha^\theta \quad (1)$$

with the exponent θ to be close to unity. The probability distribution in Eq. (1) is an instance of a power law. Zipf's law is an experimental law, not a theoretical one. The causes of Zipfian distributions in real life are a matter of some controversy. However, Zipfian distributions are commonly observed in many kinds of phenomena.

Initially the Zipf distribution was confined to the linguistic community and associated the frequency of word in a document with its rank [4, 7]. The prerequisite for the above law to be applicable in linguistics is that the size of the document to be fairly large.

2 Document distance

Its is generally admissible that the contextual "similarity" between documents (regardless of their size) can be based on their structural textual elements, namely the words forming these documents. The previous fact is the basic principle behind the *vector space model* (VSM) [6]. In the VSM, the available textual data are encoded into a numerical form and are represented by numerical vectors. Furthermore, it is generally agreed upon that the contextual similarity between documents exists also in their vectorial representation. Since the Zipf distribution of a document employs the frequencies of the words forming that particular document, it is justified to evaluate the contextual similarity based on the numerical encoding produced by the particular distribution.

A novel distance measure will be provided in the current chapter. In Appendix A it will be proven that the proposed distance measure is also a metric. This metric is used in order to evaluate the similarity between Zipf distributed vectors. The suggested metric can be easily proven that it is computationally superior (faster) than the Euclidean distance. For example, for two N_w -dimensional vectors, the computational cost of the suggested metric is N_w multiplications, a bit-shift operation and N_w additions compared to N_w multiplications and $(2N_w - 1)$ addition of the Euclidean distance.

Furthermore, by exploiting the fact that the vectors under consideration are distributed according to the Zipf law enables us to extend the suggested metric towards the direction of a statistical hypothesis. The hypothesis under consideration is whether two Zipf distributed vectors, and subsequently two documents, are similar or not. For this reason a detailed distribution is provided for the proposed metric along with a detailed proof. Also, two distribution tables are supplied for the proposed metric to make the chapter self-content.

In what follows, section 2.1 provides a description of the proposed metric and section 2.2 describes the process of incorporating the Zipf distribution in the proposed metric. It also provides a detailed proof for the evaluation of the distribution associated with the proposed metric. Following, section 2.3 provides the hypothesis testing for the evaluation of the similarity between two Zipf distributed vectors.

2.1 Proposed metric

Let us suppose that $\mathcal{X}^N = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ is a collection of N_w -dimensional random vectors, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iN_w})^T$ with cumulative probability density function $f_{\mathbf{x}}(i)$. Let also x_{im} denote the univariate random variable with distribution function $f_i(m)$, where $f_i(m)$ corresponds to the probability of the m th element of the i th vector, that is, $f_i(m) = P(x_{im})$, where

$$\sum_{m=1}^{N_w} P(x_{im}) = 1. \quad (2)$$

We further assume that the probabilities in Eq. (2) follow the Zipf distribution. In order to assess whether two vectors drawn independently from the set \mathcal{X}^N are of the same ‘‘shape’’, one needs to compare their distribution functions. For this purpose a novel metric is introduced. Let \mathbf{x}_i and \mathbf{x}_j denote two vectors randomly drawn from the set \mathcal{X}^N , ($\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}^N$). The hypothesis whose validity is to be tested is:

H_0 : The two cumulative distribution functions are ‘‘identical’’ $\Rightarrow f_{\mathbf{x}}(i) = f_{\mathbf{x}}(j)$

or equivalently

H_0 : $f_i(m) \cong f_j(m)$, for almost each m ,

against the negation of H_0 . If the null hypothesis is true, the population distributions are identical and the two samples are drawn from the same population, meaning that the vectors \mathbf{x}_i and \mathbf{x}_j should be regarded as instances of the same population. Therefore, allowing for statistically neglectful sampling variations, under H_0 there should be reasonable agreement between the two distributions. The proposed criterion between the i th and j th distributions, henceforth denoted by D_{ij} , is defined as:

$$D_{ij} = h(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_i \circ \mathbf{x}_j) = (\mathbf{x}_i + \mathbf{x}_j + g(\mathbf{x}_i, \mathbf{x}_j)) (\mathbf{x}_i + \mathbf{x}_j + g(\mathbf{x}_i, \mathbf{x}_j))^T, \quad (3)$$

where the notion (\circ) corresponds to the Hadamard product between two vectors and $g(\mathbf{x}_i, \mathbf{x}_j)$ corresponds to the N_w -dimensional vector whose the k -th element is $\sqrt{P(x_{ik})P(x_{jk})}$ (square root of the *Hadamard product* between the vectors \mathbf{x}_i and \mathbf{x}_j).

From Eq. (3), the following form for the variable D_{ij} , derives:

$$D_{ij} \triangleq \sum_{m=1}^{N_w} \frac{(f_i(m) + \sqrt{f_i(m)f_j(m)})^2}{f_i(m)} \quad (4)$$

$$= \sum_{m=1}^{N_w} \frac{(f_i^2(m) + f_i(m)f_j(m) + 2f_i(m)\sqrt{f_i(m)f_j(m)})}{f_i(m)}$$

$$= \sum_{m=1}^{N_w} \left(f_i(m) + f_j(m) + 2\sqrt{f_i(m)f_j(m)} \right)$$

$$= \sum_{m=1}^{N_w} (f_i(m) + f_j(m)) + \sum_{m=1}^n \left(2\sqrt{f_i(m)f_j(m)} \right)$$

$$= 2 + 2 \sum_{m=1}^{N_w} \left(\sqrt{f_i(m)f_j(m)} \right) \quad (5)$$

$$= 2 + 2L_{ij}^B \quad (6)$$

From Eq. (5) is evident that only the square roots of the x_{im} and x_{jm} are needed. Therefore, instead of storing the actual values for the x_{im} and x_{jm} one can only retain the square roots of them. In that way there is no need

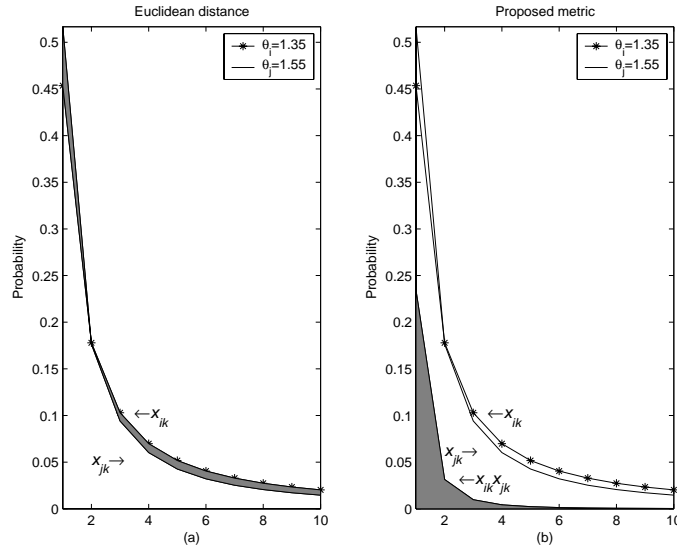


Figure 1: The divergence viewed under different metrics. The grayed area corresponds to the divergence measured from: (a) the Euclidean distance, which relies on the shaded area between the distribution functions, and (b) the proposed metric that is based on the shaded area in the bottom left side of the plot.

to evaluate the square roots each time one needs to evaluate the value of the random variable D_{ij} , thus limiting the computations cost to just N_w multiplications, a bit-shift operation (the multiplication by 2) and N_w additions. Appendix A proves that the proposed distance measure satisfies also the three properties of a metric, so it will be referred as a metric, henceforth.

From Eq. (25) is obvious that $D_{ij} \in [2, 4]$, where D_{ij} equals four when $f_i(m) = f_j(m), \forall m$. On the other hand D_{ij} equals two only in the extreme case where the distributions of the i th and j th RVs are of the following form:

$$P(x_{im}) = \begin{cases} \neq 0, & \text{when } P(x_{im}) = 0 \\ = 0, & \text{elsewhere} \end{cases} \quad \forall m \quad (7)$$

in which case the product $f_i(m)f_j(m)$ equals to zero and therefore D_{ij} tends towards the value two. So the closer the pdf of the i th RV is to the pdf of the j th RV the larger is the value of L_{ij} and subsequently, the value of D_{ij} tends toward the value of four. So the hypothesis test mentioned earlier is transformed into:

H_0 : D_{ij} is statistically equal to four

H_1 : The negation of H_0

It must be noted here that Eq. (4) resembles the *Chi-square goodness-of-fit* test proposed by Pearson but there is no other resemblance with that particular test. In fact, since Chi-square uses the maximum divergence between the distribution under considerations this might lead to unexpected results in case when the distributions differ in just two samples out of the N_w samples comprising the N_w -dimensional vectors.

Figure 1 depicts the areas used by the proposed metric and the Euclidean distance in evaluating the similarity between the distributions¹.

2.2 The Zipf distribution and the proposed metric

In order to evaluate the hypothesis test mentioned in section 2.1 it is needed to compute the probability density function of the random variable D_{ij} . In doing so one must first determine the probability of the random variable

¹The vectors used in this figure were artificially generated.

x_{im} . For the case under consideration the probability of the random variable is:

$$f_i(m) = \frac{1}{m^{\theta_i} H_{N_w, \theta_i}}, \quad (8)$$

where θ_i is a parameter dependent on the data-set under consideration and H_{N_w, θ_i} is the so-called N_w th Harmonic number of order θ_i which is a normalizing factor equal to:

$$H_{N_w, \theta_i} = \sum_{m=1}^{N_w} \frac{1}{m^{\theta_i}}. \quad (9)$$

Equation (8) is the well known *generalized Zipf* distribution [1].

The first step towards the computation of the distribution of the variable D_{ij} is to evaluate the distribution of the elements of the random vector $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2}, \dots, z_{ijN_w}) = \mathbf{x}_i \circ \mathbf{x}_j = (x_{i1}x_{j1}, x_{i2}x_{j2}, \dots, x_{iN_w}x_{jN_w})$. Since for the formation of the m th element of \mathbf{z}_{ij} it is needed to multiply the corresponding m th elements in both \mathbf{x}_i and \mathbf{x}_j this leads to the following: $P(z_{ijm}) = P(x_{im}x_{jm})$. In the previous expression the random variable x_{im} is independent of the variable x_{jm} since they refer to two different random vectors, which leads to: $P(z_{ijm}) = P(x_{im})P(x_{jm})$.

For the evaluation of the probability of z_{ijm} it is needed first to determine the cdf for m a given number, where $m \in N$. Lets denote this distribution by $F_{ij}(m)$:

$$\begin{aligned} F_{ij}(m) &= P(\text{until the } m\text{th element of } \mathbf{z}_{ij}) & (10) \\ &= F_i(m) \cdot F_j(m) = \sum_{s=1}^m f_i(s) \cdot \sum_{t=1}^m f_j(t) \\ &= \sum_{s=1}^m P(x_{is}) \cdot \sum_{t=1}^m P(x_{jt}) \\ &= \sum_{s=1}^m \frac{1}{s^{\theta_i} H_{N_w, \theta_i}} \cdot \sum_{t=1}^m \frac{1}{t^{\theta_j} H_{N_w, \theta_j}} \\ &= \frac{1}{H_{N_w, \theta_i} H_{N_w, \theta_j}} \cdot \sum_{s=1}^m \frac{1}{s^{\theta_i}} \cdot \sum_{t=1}^m \frac{1}{t^{\theta_j}} \end{aligned} \quad (11)$$

where $F_i(m)$ and $F_j(m)$ are the cdfs of the i th and j th RVs respectively. The next step is to find the pdf for the random variable z_{ijm} , that is:

$$\begin{aligned} f_{ij}(m) &= P(z_{ijm}) = F_{ij}(m) - F_{ij}(m-1) \\ &= a_{ij} \left[\sum_{s=1}^m \sum_{t=1}^m \frac{1}{s^{\theta_i}} \cdot \frac{1}{t^{\theta_j}} - \sum_{s=1}^{m-1} \sum_{t=1}^{m-1} \frac{1}{s^{\theta_i}} \cdot \frac{1}{t^{\theta_j}} \right] \end{aligned} \quad (12)$$

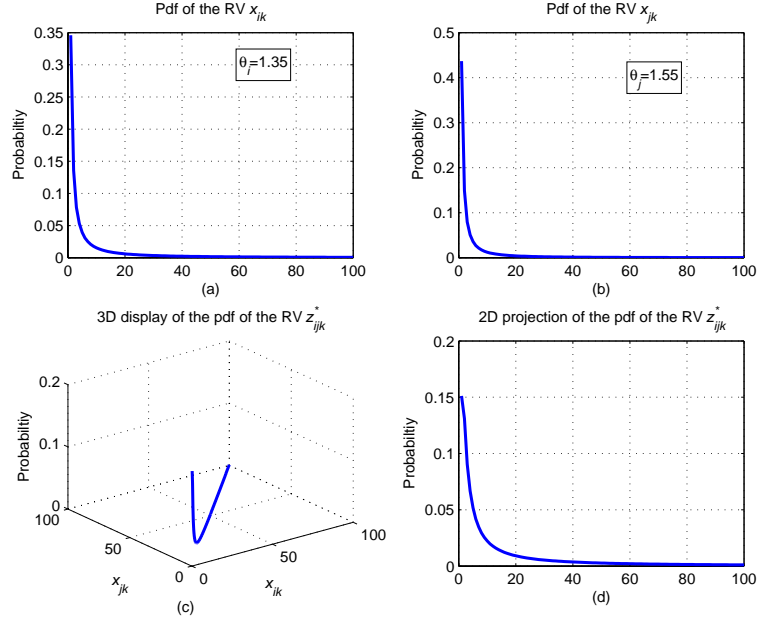


Figure 2: The probability density function for the Zipf distribution for $N_w = 100$ and for (a) $\theta_i = 1.35$, (b) $\theta_j = 1.55$, and (c) the product z_{ijm}^* .

where a_{ij} denotes the fraction $1 / (H_{N_w, \theta_i} H_{N_w, \theta_j})$. From (12) derives:

$$\begin{aligned}
 f_{ij}(m) &= a_{ij} \left[\frac{1}{m^{(\theta_i + \theta_j)}} + \frac{1}{m^{\theta_i}} \sum_{t=1}^{m-1} \frac{1}{t^{\theta_j}} + \frac{1}{m^{\theta_j}} \sum_{s=1}^{m-1} \frac{1}{s^{\theta_i}} \right] \\
 &= a_{ij} \left[\frac{1}{m^{(\theta_i + \theta_j)}} + \frac{H_{N_w, \theta_j}}{m^{\theta_i}} F_j(m-1) + \frac{H_{N_w, \theta_i}}{m^{\theta_j}} F_i(m-1) \right] \Rightarrow \\
 f_{ij}(m) &= \begin{cases} a_{ij}, & m = 1 \\ a_{ij} \left[\frac{1}{m^{(\theta_i + \theta_j)}} + \frac{H_{N_w, \theta_j}}{m^{\theta_i}} F_j(m-1) + \frac{H_{N_w, \theta_i}}{m^{\theta_j}} F_i(m-1) \right], & \forall m \in \{2, N_w\} \\ 0, & elsewhere \end{cases} \quad (13)
 \end{aligned}$$

Figure 2 depicts the process of obtaining the distribution of the random variable z_{ijm} .

After the computation of the pdf for z_{ijm} it is needed to compute the density function of the random variable $\sqrt{z_{ijm}}$. This is due to the fact that D_{ij} is a linear combination of $\sqrt{z_{ijm}}$. Let z_{ijm}^* denote the square root of z_{ijm} , that is, $z_{ijm}^* = \sqrt{z_{ijm}}$, where $m \in \{1, 2, \dots, N_w\}$. Since the sample space for the RV z_{ijm} is the set $Z_1 = \{1, 2, \dots, N_w\}$, the sample space corresponding to z_{ijm}^* is the set $Z_2 = \{1, \sqrt{2}, \dots, \sqrt{N_w}\}$. It must be noted here that the cardinality of the set Z_2 is equal to N_w since each element of the set Z_2 is the square root of the set Z_1 . So z_{ijm} is a discrete RV then the RV z_{ijm}^* is of the same pdf as the RV z_{ijm} [5] and if $f_{ij}^*(m)$ denotes the pdf of the RV z_{ijm}^* , then, $f_{ij}^*(m) = f_{ij}(m), \forall m$.

The final step is to evaluate the pdf of the random variable $L_{ij} = \sum_{m=1}^{N_w} \sqrt{z_{ijm}} = \sum_{m=1}^{N_w} z_{ijm}^*$. For a large value

of N_w and due to the central limit theorem (CLT) the pdf of the above sum tends toward the normal distribution with mean value μ and variance σ^2 [5]. The mean value is:

$$\begin{aligned}
 \mu &= \mathbb{E}[L_{ij}] = \mathbb{E}\left[\sum_{m=1}^{N_w} z_{ijm}^*\right] = \sum_{m=1}^{N_w} \mathbb{E}[z_{ijm}^*] \\
 &= N_w \mathbb{E}[z_{ijm}^*] = N_w \sum_{m=1}^{N_w} \sqrt{m} f_{ij}^*(m) \\
 &= N_w a_{ij} \sum_{m=2}^{N_w} \sqrt{m} \left[\frac{1}{m^{(\theta_i+\theta_j)}} + \frac{H_{N_w, \theta_j}}{m^{\theta_i}} F_j(m-1) + \frac{H_{N_w, \theta_i}}{m^{\theta_j}} F_i(m-1) \right] \\
 &= N_w a_{ij} \sum_{m=1}^{N_w} \left[\frac{1}{m^{(\theta_i+\theta_j)-0.5}} + \frac{H_{N_w, \theta_j}}{m^{\theta_i-0.5}} F_j(m-1) + \frac{H_{N_w, \theta_i}}{m^{\theta_j-0.5}} F_i(m-1) \right] \\
 &= N_w a_{ij} \begin{bmatrix} H_{N_w, (\theta_i+\theta_j)-0.5} + \\ H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_i-0.5}} + \\ H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_j-0.5}} \end{bmatrix} \tag{14}
 \end{aligned}$$

and the variance is:

$$\begin{aligned}
 \sigma^2 &= \mathbb{E}[(L_{ij} - \mu)^2] = \mathbb{E}[(L_{ij})^2] - \mu^2 \\
 &= \mathbb{E}\left[\left(\sum_{m=1}^{N_w} z_{ijm}^*\right)^2\right] - \mu^2 \\
 &= \mathbb{E}\left[\sum_{m=1}^{N_w} (z_{ijm}^*)^2 + 2 \sum_{\substack{m_1, m_2=1 \\ m_1 \neq m_2}}^{N_w} z_{ijm_1}^* z_{ijm_2}^*\right] - \mu^2 \\
 &= \sum_{m=1}^{N_w} \mathbb{E}[(z_{ijm}^*)^2] + 2 \sum_{\substack{m_1, m_2=1 \\ m_1 \neq m_2}}^{N_w} \mathbb{E}[z_{ijm_1}^* z_{ijm_2}^*] - \mu^2 \\
 &= N_w \mathbb{E}[(z_{ijm}^*)^2] + 2N_w(N_w - 1) \mathbb{E}[z_{ijm_1}^* z_{ijm_2}^*] - \mu^2 \tag{15}
 \end{aligned}$$

At this point, and without loss of generality, it can be regarded that the RVs $z_{ijm_1}^*$ and $z_{ijm_2}^*$ are independent. Having this postulate:

$$\mathbb{E}[z_{ijm_1}^* z_{ijm_2}^*] = \mathbb{E}[z_{ijm_1}^*] \mathbb{E}[z_{ijm_2}^*] \tag{16}$$

The first term on the right side of the variance equation is equal to:

$$\begin{aligned}
 E \left[(z_{ijm})^2 \right] &= a_{ij} \sum_{m=1}^{N_w} (\sqrt{m})^2 \left[\begin{array}{l} \frac{1}{m^{(\theta_i+\theta_j)}} + \\ \frac{H_{N_w, \theta_j}}{m^{\theta_j}} F_j(m-1) + \\ \frac{H_{N_w, \theta_i}}{m^{\theta_i}} F_i(m-1) \end{array} \right] \\
 &= a_{ij} \left[\begin{array}{l} H_{N_w, (\theta_i+\theta_j)-1} + \\ H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_j-1}} + \\ H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_i-1}} \end{array} \right] \quad (17)
 \end{aligned}$$

whereas the second term equals to:

$$E \left[z_{ijm_1}^* \right] E \left[z_{ijm_2}^* \right] = \mu^2 \quad (18)$$

So the total variance of the random variable L_{ij} is:

$$\begin{aligned}
 \sigma^2 &= N_w a_{ij} \left[H_{N_w, (\theta_i+\theta_j)-1} + H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_j-1}} + H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_i-1}} \right] \\
 &+ [2N_w(N_w-1) - 1] \mu^2 \\
 &= N_w a_{ij} \left[H_{N_w, (\theta_i+\theta_j)-1} + H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_j-1}} + H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_i-1}} \right] \\
 &+ N_w a_{ij} [2N_w(N_w-1) - 1] \left[\begin{array}{l} H_{N_w, (\theta_i+\theta_j)-0.5} + \\ H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_j-0.5}} + \\ H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_i-0.5}} \end{array} \right] \\
 &= N_w a_{ij} \left[H_{N_w, (\theta_i+\theta_j)-1} + [2N_w(N_w-1) - 1] H_{N_w, (\theta_i+\theta_j)-0.5} \right] \\
 &+ 2a_{ij} N_w^2 (N_w-1) H_{N_w, \theta_j} \left[\sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_j-1}} + \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_j-0.5}} \right] \\
 &+ 2a_{ij} N_w^2 (N_w-1) H_{N_w, \theta_i} \left[\sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_i-1}} + \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_i-0.5}} \right] \quad (19)
 \end{aligned}$$

which is equal to:

$$\begin{aligned}
 \sigma^2 &= N_w a_{ij} \sum_{m=1}^{N_w} \frac{1 + [2N_w(N_w-1) - 1] m^{-0.5}}{m^{(\theta_i+\theta_j)-1}} \\
 &+ 2a_{ij} N_w^2 (N_w-1) H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1) [1 - m^{-0.5}]}{m^{\theta_j-1}} \\
 &+ 2a_{ij} N_w^2 (N_w-1) H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1) [1 - m^{-0.5}]}{m^{\theta_i-1}} \quad (20)
 \end{aligned}$$

Finally, the pdf of the RV $D_{ij} = 2 + 2L_{ij}$ has to be computed. Given the fact that L_{ij} is normally distributed we get the following pdf for the RV D_{ij} :

$$f_{D_{ij}}(t) = \frac{f_{L_{ij}}\left(\frac{t-2}{2}\right)}{2} = \frac{1}{\sqrt{8\pi\sigma}} \exp\left\{-\frac{1}{8\sigma^2}(t-2-2\mu)^2\right\} \quad (21)$$

where μ and σ are the expected value and the standard deviation of the random variable L_{ij} .

But since the random variable D_{ij} is confined in the interval $[2, 4]$ ($D_{ij} \in [2, 4]$), Eq. (21) obviously underestimates the true pdf of D_{ij} . The accurate form of the pdf is:

$$f_{D_{ij}}(t) = \begin{cases} 0, & -\infty \leq t \leq 2 \\ \frac{\exp\left\{-\frac{1}{8\sigma^2}(t-2-2\mu)^2\right\}}{\int_2^4 \exp\left\{-\frac{1}{8\sigma^2}(t-2-2\mu)^2\right\} dt}, & 2 \leq t \leq 4 \\ 0, & 4 \leq t \leq +\infty \end{cases} \quad (22)$$

which is the so-called *truncated* normal distribution [5]. Equation (22) can be simplified in the following form:

$$f_{D_{ij}}(t) = \begin{cases} \frac{\exp\left\{-\frac{1}{8\sigma^2}(t-2(1+\mu))^2\right\}}{\sqrt{2\pi\sigma} \operatorname{erf}\left(\frac{x}{\sqrt{2\sigma}} - \frac{2(1+\mu)}{\sqrt{2\sigma}}\right)\Big|_{x=2}}^4, & 2 \leq t \leq 4 \\ 0, & \text{elsewhere} \end{cases} \quad (23)$$

where $\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-t^2} dt$ denotes the so-called *error function* [2].

2.3 Hypothesis test evaluation

To assess whether the i th and j th distributions are of the same ‘‘shape’’, the RV D_{ij} will be employed in a hypothesis test. The hypothesis is as follows:

H_0 : The i th and j th pdf are statistically identical which equals to $D_{ij} \rightarrow 4$.

H_1 : The i th and j th pdf are not identical.

Given a pre-defined significant level α , the rejection region for the above hypothesis test is formulated as follows:

$$\begin{aligned} \alpha &= P(D_{ij} \leq z_\alpha) = \int_2^{z_\alpha} f_{D_{ij}}(t) dt \\ &= \frac{\operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2\sigma}} - \frac{z_\alpha}{2\sqrt{2\sigma}}\right) - \operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2\sigma}} - \frac{1}{\sqrt{2\sigma}}\right)}{\operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2\sigma}} - \frac{2}{\sqrt{2\sigma}}\right) - \operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2\sigma}} - \frac{1}{\sqrt{2\sigma}}\right)}. \end{aligned} \quad (24)$$

In Eq. (24) the only unknown is the parameter z_α . After evaluating the parameter z_α the null hypothesis is accepted if the expression $z_\alpha \leq D_{ij}(t)$ is true otherwise its rejected. Figure 3 depicts the distribution of the RV D_{ij} along with the support regions for the hypothesis H_0 and the alternative hypothesis H_1 .

Appendix B provides a brief description of the computation of the critical values for the acceptance or the rejection of the null hypothesis along with two tables with critical values computed by the proposed Eq. (27).

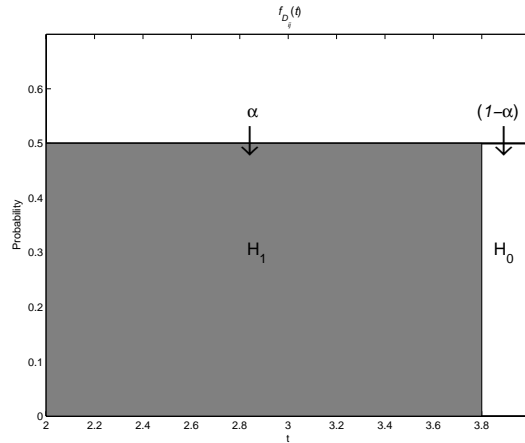


Figure 3: The support regions for the null and the alternative hypothesis for the RV D_{ij} for $N_w = 2000$, $\theta_i = 1.35$ and $\theta_j = 1.55$ at a significant level of $\alpha = 0.90$. (Important remark: Although the above graph implies a uniform distribution this is not true. The slope of the line in the graph approaches zero but still is significantly different than this value.)

3 Conclusions

The present report provides a preliminary mathematical analysis on a novel metric, that is also introduced in the report, for the evaluation of the contextual similarity between documents. The proposed metric is computationally superior than the Euclidean distance which is oftenly employed in similar tasks. Further investigation will be performed towards the direction of the biasness of the introduced metric (investigate whether the proposed metric is biased or not).

A Is it a metric?

In order to prove that the proposed statistic, D_{ij} , is also a metric distance the following has to be proven:

Positiveness: Since $f_i(m)$ and $f_j(m)$ for $m = 1, 2, \dots, N_w$ contains the total probability mass of the i th and j th

RV the following stems out:

$$\begin{aligned}
 & \left. \begin{aligned} 0 \leq x_{im} < 1 \text{ and } \sum_{m=1}^{N_w} x_{im} = 1 \\ 0 \leq x_{jm} < 1 \text{ and } \sum_{m=1}^{N_w} x_{jm} = 1 \end{aligned} \right\} \Rightarrow \\
 & 0 \leq x_{im}x_{jm} \leq 1 \Rightarrow \\
 & 0 \leq \sqrt{x_{im}x_{jm}} \leq 1 \Rightarrow \\
 & 0 \leq \sum_{m=1}^{N_w} \sqrt{x_{im}x_{jm}} \leq 1 \Rightarrow \\
 & 0 \leq 2L_{ij} \leq 2 \Rightarrow \\
 & 2 \leq 2 + 2L_{ij} \leq 4 \Rightarrow \\
 & 2 \leq D_{ij} \leq 4
 \end{aligned} \tag{25}$$

In case where $i = j$ then $L_{ii} = \sum_{m=1}^{N_w} \sqrt{x_{im}x_{jm}} = \sum_{m=1}^{N_w} x_{im} = 1 \Rightarrow D_{ii} = 2 + 2L_{ii} = 4$.

Symmetry: Since $x_{im}x_{jm} = x_{jm}x_{im} \Rightarrow D_{ij} = D_{ji}$.

Triangular inequality: In order to prove the triangular inequality it can be proven that:

$$D_{ij} + D_{jm} \geq D_{im} \Rightarrow 2 + 2L_{ij} + 2 + 2L_{jm} \geq 2 + 2L_{im} \Rightarrow 1 + L_{ij} + L_{jm} \geq L_{im} \tag{26}$$

which is obvious since $L_{ij}, L_{jm} \geq 0$ and $1 \geq L_{im}$.

B Critical values

The critical values for the hypothesis test associated with the RV D_{ij} are computed using the following:

$$\begin{aligned}
 \frac{1}{2\sqrt{2}\sigma} (2(1+\mu) - z_\alpha) & \int_0^{e^{-t^2}} dt = \frac{\alpha\sqrt{\pi}}{2} \operatorname{erf}\left(\frac{(\mu-1)}{\sqrt{2}\sigma}\right) + \frac{(1-\alpha)\sqrt{\pi}}{2} \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) \Rightarrow \\
 z_\alpha & = 2(1+\mu) - 2\sqrt{2}\sigma \operatorname{erf}^{-1}\left(\alpha \operatorname{erf}\left(\frac{(\mu-1)}{\sqrt{2}\sigma}\right) + (1-\alpha) \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right)\right),
 \end{aligned} \tag{27}$$

where erf^{-1} is the inverse of the error function [2]. Using Eq. (27) and pre-defined significance levels two tables of critical values for hypothesis test were computed. Table I² corresponds to a significance level of $\alpha = 0.90$ when the dimensionality of the feature vectors is $N_w = 2000$, whereas, table II³ corresponds to a significance level of $\alpha = 0.95$ under the same feature vector dimensionality.

²The values of the table a scaled version of the original values. Original values = 3.8+scaled value*10⁻⁹.

³The values of the table a scaled version of the original values. Original values = 3.8+scaled value*10⁻⁹.

References

- [1] References on zipf's law. <http://linkage.rockefeller.edu/wli/zipf>.
- [2] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Pubns., 10th edition, 1974.
- [3] L. A. Adamic. Zipf, Power-laws, and Pareto - a ranking tutorial. <http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html>.
- [4] D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.
- [5] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1984.
- [6] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, 1983.
- [7] R. B. Yates and B. R. Neto. *Modern Information Retrieval*. ACM Press, 1999.

Table I: Distribution tables for the RV D_{ij} and for $N_w = 2000$ at $\alpha = 0.90$ (10% Confidence Level).

1.05	1.10	1.15	1.20	1.25	1.30	1.35	1.40	1.45	1.50	1.55	1.60	1.65	1.70
1.05	1.2081	1.2117	1.2146	1.2183	1.2234	1.2256	1.2277	1.2314	1.2314	1.2328	1.2365	1.2372	1.2365
1.10		1.2161	1.2197	1.2263	1.2299	1.2336	1.2372	1.2408	1.243	1.2452	1.2459	1.2481	1.2488
1.15			1.2263	1.2328	1.2379	1.2438	1.2481	1.251	1.2547	1.2576	1.259	1.2612	1.2627
1.20				1.2394	1.2467	1.2539	1.2583	1.2634	1.2667	1.271	1.2736	1.2761	1.2776
1.25					1.2561	1.2619	1.2685	1.2754	1.2812	1.2852	1.2889	1.2929	1.2958
1.30						1.2725	1.2816	1.2885	1.2954	1.3012	1.3056	1.3092	1.3132
1.35							1.2929	1.3023	1.31	1.3161	1.3231	1.3282	1.3322
1.40								1.3151	1.3238	1.3329	1.3405	1.3471	1.3529
1.45									1.3383	1.3489	1.358	1.366	1.3733
1.50										1.3642	1.3751	1.3853	1.3938
1.55											1.3915	1.4027	1.4135
1.60												1.4195	1.4313
1.65													1.4484

1.75	1.80	1.85	1.90	1.95	2.00
1.05	1.2372	1.2379	1.2379	1.2394	1.2394
1.10	1.2488	1.2503	1.2503	1.251	1.2503
1.15	1.2641	1.2649	1.2656	1.2659	1.2659
1.20	1.279	1.2805	1.2805	1.2823	1.282
1.25	1.2972	1.2983	1.2998	1.3001	1.3012
1.30	1.3151	1.3172	1.3187	1.3201	1.3216
1.35	1.3362	1.338	1.3409	1.3427	1.3445
1.40	1.3569	1.3609	1.3634	1.3656	1.3682
1.45	1.3787	1.3831	1.3871	1.3898	1.3918
1.50	1.4005	1.406	1.4113	1.4144	1.4173
1.55	1.4216	1.4287	1.4344	1.4389	1.4424
1.60	1.4418	1.4502	1.4575	1.4628	1.4673
1.65	1.4604	1.4704	1.4784	1.4851	1.49
1.70	1.4769	1.4882	1.4975	1.5059	1.5119
1.75		1.5046	1.5153	1.5241	1.5315
1.80			1.5301	1.5402	1.5483
1.85				1.5539	1.5626
1.90					1.5752
1.95					1.593
2.00					

Table II: Distribution tables for the RV D_{ij} and for $N_w = 2000$ at $\alpha = 0.95$ (5% Confidence Level).

	1.05	1.10	1.15	1.20	1.25	1.30	1.35	1.40	1.45	1.50	1.55	1.60	1.65	1.70
1.05														
1.10		6.3825	6.397	6.4116	6.4552	6.448	6.4625	6.4771	6.4916	6.5062	6.4989	6.5207	6.528	6.5207
1.15			6.4188	6.4334	6.4698	6.4843	6.5134	6.5207	6.5498	6.5571	6.5644	6.5716	6.5935	6.6007
1.20				6.4625	6.5134	6.528	6.5571	6.5935	6.608	6.6226	6.6299	6.6517	6.6517	6.6662
1.25					6.5425	6.5789	6.6153	6.6444	6.6808	6.6844	6.7099	6.7244	6.739	6.7463
1.30						6.6226	6.659	6.699	6.7317	6.7608	6.7826	6.8045	6.8227	6.8372
1.35							6.7172	6.7608	6.7936	6.8409	6.8736	6.8918	6.91	6.9354
1.40								6.8227	6.8736	6.9063	6.9464	6.99	7.0046	7.0373
1.45									6.9464	6.9864	7.0373	7.081	7.1101	7.1392
1.50										7.0591	7.121	7.1646	7.2083	7.2483
1.55											7.201	7.2556	7.3102	7.3538
1.60												7.3393	7.4029	7.4593
1.65													7.4902	7.5557
														7.643
1.75	6.5207	6.5425	6.5425	6.5425	6.5425	6.5425	6.5425	6.528	6.5498	6.5425	2.00	6.5425	6.5425	6.5425
1.80	6.6007	6.608	6.608	6.608	6.608	6.608	6.608	6.608	6.6007	6.6007	6.6007	6.6007	6.6007	6.6007
1.85	6.659	6.7499	6.7608	6.7608	6.7608	6.7608	6.7608	6.7608	6.6844	6.6844	6.6844	6.6844	6.6844	6.6844
1.90	6.8445	6.8518	6.8518	6.8554	6.8554	6.8554	6.8554	6.859	6.8663	6.8663	6.8663	6.8663	6.8663	6.8663
1.95	7.0482	7.0628	7.0628	7.0773	7.0773	7.0773	7.0773	7.0846	6.9718	6.9718	6.9718	6.9718	6.9718	6.9718
2.00	7.161	7.1828	7.1828	7.1937	7.1937	7.1937	7.1937	7.2083	7.2119	7.2119	7.2119	7.2119	7.2119	7.2119
	7.2738	7.3029	7.3029	7.3211	7.3211	7.3211	7.3211	7.3338	7.3484	7.3484	7.3484	7.3484	7.3484	7.3484
	7.3938	7.4193	7.4193	7.4448	7.4448	7.4448	7.4448	7.4666	7.4811	7.4811	7.4811	7.4811	7.4811	7.4811
	7.4993	7.5393	7.5393	7.5703	7.5703	7.5703	7.5703	7.5957	7.6121	7.6121	7.6121	7.6121	7.6121	7.6121
	7.6066	7.6539	7.6539	7.6921	7.6921	7.6921	7.6921	7.7231	7.7413	7.7413	7.7413	7.7413	7.7413	7.7413
	7.7049	7.7594	7.7594	7.8067	7.8067	7.8067	7.8067	7.8377	7.8631	7.8631	7.8631	7.8631	7.8631	7.8631
	7.7922	7.8559	7.8559	7.905	7.905	7.905	7.905	7.9468	7.9814	7.9814	7.9814	7.9814	7.9814	7.9814
								8.0432	8.0814	8.0814	8.0814	8.0814	8.0814	8.0814
									8.1287	8.1287	8.1287	8.1287	8.1287	8.1287
										8.2015	8.2015	8.2015	8.2015	8.2015
											8.2778	8.2778	8.2778	8.2778
												8.3506	8.3506	8.3506
													8.407	8.407