

A SOM VARIANT FOR HEAVILY SKEWED VECTORS

Apostolos Georgakis and Haibo Li

Digital Media Laboratory (DML)
Department of Applied Physics and Electronics,
Umeå University, SE-90187, Sweden
apostolos.georgakis@tfe.umu.se.

ABSTRACT

A variant of the self organizing maps algorithm is proposed in this paper for organization and retrieval of documents that exhibit highly skewed word distributions. The proposed variant replaces the Euclidean distance employed by the standard self organizing maps algorithm with a novel metric.

1. INTRODUCTION

A crucial issue for the efficient *document organization* (DO) is the evaluation of the contextual similarity between documents. In the information retrieval community is generally admissible that automatic DO can be grounded on the contextual “similarity” between documents. And the contextual similarity can be based on the document’s structural elements, namely the words forming the documents. This paper provides a method for evaluating the document similarity by addressing the issue of a proper distance between documents represented as vectors in the *vector space model* (VSM) [1, 2].

In the VSM, the available textual data are represented by numerical vectors whose elements are related to the frequency of appearance of the words forming the documents. Furthermore, it is generally agreed upon that the contextual similarity between documents exists also in their vectorial representation. Moreover, it has been shown that the word frequency histograms exhibit a heavily skewed pattern (when the frequencies are arranged into descending order). A plethora of distributions has been proposed in the literature that are capable to model the above phenomenon; the most prevalent among them is the well-know *Zipf* distribution [3, 4].

The Zipf distribution rely on an empirical law discovered by Estoup in 1916. 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 has the form of a power-law function:

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

with the exponent θ to be close to unity.

A novel distance measure will be proposed in the current paper. The proposed distance measure can also be

proven to be a metric (see Appendix A). 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 than the Euclidean distance which is the prime metric employed in information retrieval tasks.

Furthermore, 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.

In what follows, section 2 provides a description of the proposed metric. Subsection 2.1 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, subsection 2.2 provides the hypothesis testing for the evaluation of the similarity between two Zipf distributed vectors. Furthermore, sections 3 and 4 describes the SOM algorithm and its variant which relies on the proposed metric and finally, section 5 provides the experimental results for the comparison of the standard SOM and its Zipf variant.

2. 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 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, and $\sum_{m=1}^{N_w} f_i(m) = 1$. 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 . The hypothesis whose validity is to be tested is:

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

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:

$$D_{ij} = \| (\mathbf{x}_i + \mathbf{x}_j + g(\mathbf{x}_i, \mathbf{x}_j)) \|^2, \quad (2)$$

where $g(\mathbf{x}_i, \mathbf{x}_j)$ 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. (2), the following form for the variable D_{ij} derives:

$$D_{ij} \triangleq \sum_{m=1}^{N_w} \frac{(f_i(m) + \sqrt{f_i(m)f_j(m)})^2}{f_i(m)} \quad (3)$$

$$= 2 + 2 \sum_{m=1}^{N_w} \left(\sqrt{f_i(m)f_j(m)} \right) \quad (4)$$

$$= 2 + 2L_{ij}^B \quad (5)$$

From Eq. (4) 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, thus limiting the computations cost to just N_w multiplications, a bit-shift operation and N_w additions. From Eq. (4) is also 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:

$$f_i(m) = \begin{cases} \neq 0, & \text{when } f_j(m) = 0 \\ = 0, & \text{elsewhere} \end{cases} \quad \forall m \quad (6)$$

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

$$H_0: D_{ij} \text{ tends towards the value four.}$$

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

2.1. The Zipf distribution and the proposed metric

In order to evaluate the hypothesis test mentioned in section 2 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 x_{im} . For

¹The vectors used in this figure were artificially generated.

the case under consideration the probability of the random variable is:

$$f_i(m) = \frac{1}{m^{\theta_i} H_{N_w, \theta_i}}, \quad (7)$$

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. Equation (7) is the well known *generalized Zipf* distribution [3].

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 $\mathbf{z}_{ij} = (z_{ij1}, \dots, z_{ijN_w}) = \mathbf{x}_i \circ \mathbf{x}_j = (x_{i1}x_{j1}, \dots, x_{iN_w}x_{jN_w})$. 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 $F_{ij}(m)$:

$$F_{ij}(m) = P(\text{until the } m\text{th element of } \mathbf{z}_{ij}) \quad (8)$$

$$= F_i(m) \cdot F_j(m)$$

$$= \frac{1}{H_{N_w, \theta_i} H_{N_w, \theta_j}} \cdot \sum_{s=1}^m \frac{1}{s^{\theta_i}} \cdot \sum_{t=1}^m \frac{1}{t^{\theta_j}} \quad (9)$$

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:

$$f_{ij}(m) = a_{ij} \sum_{s=1}^m \sum_{t=1}^m \frac{1}{s^{\theta_i}} \cdot \frac{1}{t^{\theta_j}} - a_{ij} \sum_{s=1}^{m-1} \sum_{t=1}^{m-1} \frac{1}{s^{\theta_i}} \cdot \frac{1}{t^{\theta_j}} \quad (10)$$

where a_{ij} denotes the fraction $1/(H_{N_w, \theta_i} H_{N_w, \theta_j})$. From (10) it derives:

$$f_{ij}(m) = \begin{cases} a_{ij} & , u_1 \\ a_{ij} \left[\frac{1}{m^{(\theta_i + \theta_j)}} + \frac{H_{N_w, \theta_i}}{m^{\theta_i}} F_j(m-1) + \frac{H_{N_w, \theta_j}}{m^{\theta_j}} F_i(m-1) \right] & , u_2 \\ 0 & , u_3 \end{cases} \quad (11)$$

where u_1, u_2, u_3 corresponds to $\{m=1\}, \{\forall m \in \{2, N_w\}\}$ and $\{0\}$ respectively. 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

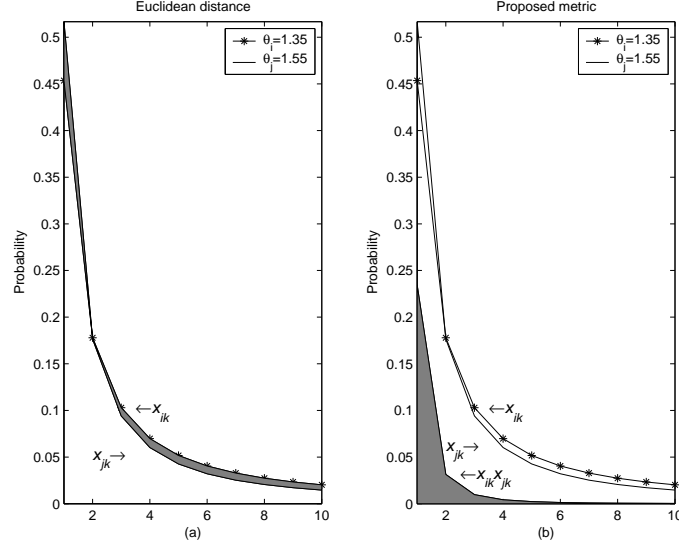


Figure 1. The divergence viewed under different metrics. The grayed area corresponds to the divergence measured from: (a) the Euclidean distance, which relies on the shaded area between the distribution functions, and (b) the proposed metric that is based on the shaded area in the bottom left side of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

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 the evaluation of 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:

$$\mu = N_w a_{ij} \begin{bmatrix} H_{N_w, (\theta_i + \theta_j) - 0.5} + \\ H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_i - 0.5}} + \\ H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_j - 0.5}} \end{bmatrix} \quad (12)$$

and the variance is:

$$\sigma^2 = N_w E \left[(z_{ijm}^*)^2 \right] + 2N_w (N_w - 1) E \left[z_{ijm_1}^* z_{ijm_2}^* \right] - \mu^2 \quad (13)$$

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:

$$E \left[z_{ijm_1}^* z_{ijm_2}^* \right] = E \left[z_{ijm_1}^* \right] E \left[z_{ijm_2}^* \right] \quad (14)$$

The first term on the right side of the variance equation is

equal to:

$$E \left[(z_{ijm})^2 \right] = a_{ij} \begin{bmatrix} H_{N_w, (\theta_i + \theta_j) - 1} + \\ H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1)}{m^{\theta_i - 1}} + \\ H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1)}{m^{\theta_j - 1}} \end{bmatrix} \quad (15)$$

whereas the second term equals to:

$$E \left[z_{ijm_1}^* \right] E \left[z_{ijm_2}^* \right] = \mu^2 \quad (16)$$

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

$$\begin{aligned} \sigma^2 &= N_w a_{ij} \sum_{m=1}^{N_w} \frac{1 + [2N_w(N_w - 1) - 1] m^{-0.5}}{m^{(\theta_i + \theta_j) - 1}} \\ &+ 2a_{ij} N_w^2 (N_w - 1) H_{N_w, \theta_j} \sum_{m=1}^{N_w} \frac{F_j(m-1) [1 - m^{-0.5}]}{m^{\theta_i - 1}} \\ &+ 2a_{ij} N_w^2 (N_w - 1) H_{N_w, \theta_i} \sum_{m=1}^{N_w} \frac{F_i(m-1) [1 - m^{-0.5}]}{m^{\theta_j - 1}} \end{aligned} \quad (17)$$

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} :

$$f_{D_{ij}}(t) = \frac{1}{\sqrt{8\pi\sigma}} \exp \left\{ -\frac{1}{8\sigma^2} (t - 2 - 2\mu)^2 \right\} \quad (18)$$

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

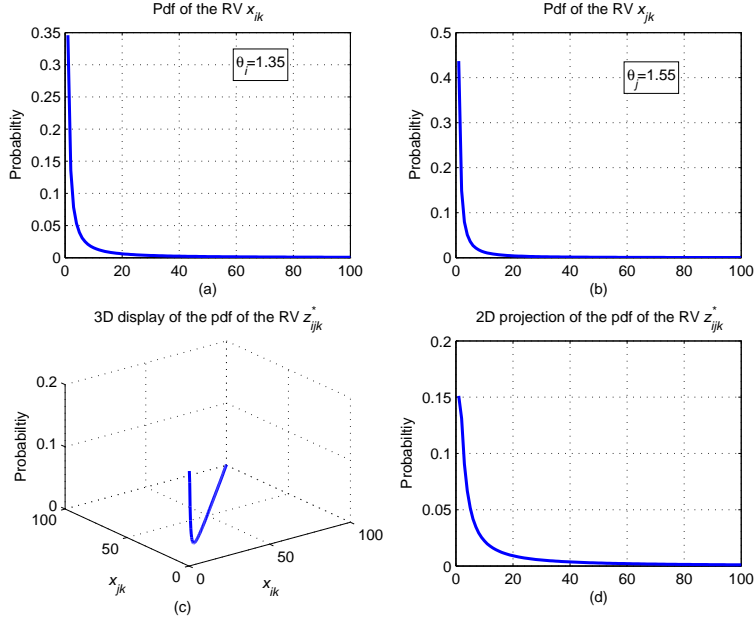


Figure 2. The probability density function for the Zipf distribution for $N_w = 100$ and for (a) $\theta_i = 1.35$, (b) $\theta_j = 1.55$, and (c) the product z_{ijm}^* .

But since the random variable D_{ij} is confined in the interval $[2, 4]$ ($D_{ij} \in [2, 4]$), Eq. (18) obviously underestimates the true pdf of D_{ij} . The accurate form of the pdf is:

$$f_{D_{ij}}(t) = \begin{cases} 0 & , u_1 \\ \frac{\exp\{-\frac{1}{8\sigma^2}(t-2-2\mu)^2\}}{\int_2^4 \exp\{-\frac{1}{8\sigma^2}(t-2-2\mu)^2\} dt} & , u_2 \\ 0 & , u_3 \end{cases} \quad (19)$$

where u_1, u_2 and u_3 corresponds to $\{-\infty \leq t \leq 2\}$, $\{2 \leq t \leq 4\}$ and $\{4 \leq t \leq +\infty\}$ respectively. Equation (19) is the so-called *truncated* normal distribution [5]. Equation (19) can be simplified in the following form:

$$f_{D_{ij}}(t) = \begin{cases} \frac{\exp\{-\frac{1}{8\sigma^2}(t-2(1+\mu))^2\}}{\sqrt{2\pi}\sigma \operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma} - \frac{2(1+\mu)}{\sqrt{2}\sigma}\right)_{x=2}}, & 2 \leq t \leq 4 \\ 0, & \text{elsewhere} \end{cases} \quad (20)$$

where $\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-t^2} dt$ denotes the so-called *error function* [6].

2.2. Hypothesis test evaluation

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

$$\begin{aligned} \alpha &= P(D_{ij} \leq z_\alpha) = \int_2^{z_\alpha} f_{D_{ij}}(t) dt \\ &= \frac{\operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2}\sigma} - \frac{z_\alpha}{2\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2}\sigma} - \frac{1}{\sqrt{2}\sigma}\right)}{\operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2}\sigma} - \frac{2}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{(1+\mu)}{\sqrt{2}\sigma} - \frac{1}{\sqrt{2}\sigma}\right)} \end{aligned} \quad (21)$$

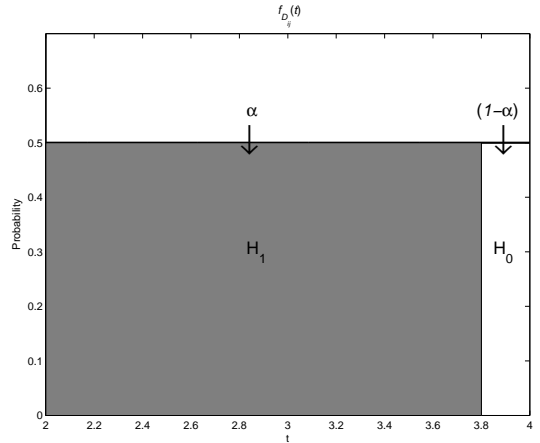


Figure 3. The support regions for the null and the alternative hypothesis for the RV D_{ij} for $N_w = 2000$, $\theta_i = 1.35$ and $\theta_j = 1.55$ at a significant level of $\alpha = 0.90$.

In Eq. (21) 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 . It should be noted here that although the graph in question 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. SELF-ORGANIZING MAP

Let \mathcal{W} denote the set of reference vectors of the neurons on the computational layer of the SOM, $\{\mathbf{w}_i(t) \in \mathbb{R}^{N_w}$,

$l = 1, 2, \dots, Q$ }, where the parameter t denotes discrete time (number of iterations) and Q corresponds to the total number of neurons. During the training phase, the algorithm tries to identify the *winning* reference vector, $\mathbf{w}_s(t)$, to a specific feature vector \mathbf{x}_h . The index of the winning reference vector is given by: $s = \arg \min \|\mathbf{x}_h - \mathbf{w}_l(t)\|, \forall l = 1, 2, \dots, Q$, where $\|\cdot\|$ denotes the Euclidean distance.

The reference vector of the winner as well as the reference vectors of the neurons in its neighborhood are modified towards \mathbf{x}_h using the following equation: $\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + a(t) [\mathbf{x}_h - \mathbf{w}_i(t)]$, where $a(t)$ is the *learning rate*, which is a monotonically decreasing parameter and the index i denotes the neurons in the neighborhood of the winner neuron.

4. ZIPF VARIANT

Let the weight vector of each neuron, during the initialization phase of the algorithm, be chosen as one of the available vectors in \mathcal{X}^N . In the Zipf SOM variant the Euclidean norm is replaced with the Eq. (5) in identifying the winner neuron.

In identifying the index of the winner neuron with respect to a specific, randomly selected, feature vector from \mathcal{X}^N , the distances between all the reference vectors of the NN and the feature vector under consideration are assessed. If the null hypothesis, defined in subsection 2.2, is validated to be “true”, then the reference vector corresponds to the winner neuron². After the identification of the winner neuron and for either single or multiple winners the updating procedure is identical to the standard SOM updating procedure.

With the advance of time the significance level in the hypothesis testing for the winner identification is increased from 0.01 to 0.025 and finally to 0.05. As a result, fewer neurons are labeled as winner neurons. Multiple assignment of the same document, during the early iterations of the algorithm, are permitted in order to achieve a global ordering of the indicator vectors and accordingly the documents on the lattice. Finally, prior to the completion of the training phase and in order to fine tune the map, the identification of the winning neuron is achieved by:

$$s = \max \{D_{ij}\}, \forall q \in \{1, 2, \dots, Q\}, \quad (22)$$

that is, we select the neuron whose distance from the feature vector under consideration is maximum. Eq. (22) can be used as an ultimate mean to resolve ambiguities when multiple winners are still determined.

5. EXPERIMENTAL RESULTS

To test the proposed metric against the SOM algorithm the Reuters-21578 corpus was used [7]. The documents are marked up using *SGML* tags and are manually annotated using 135 topic categories according to their content (some of the 135 categories are not used at all and in some cases the documents are multiply annotated).

²It must be noted that one or more than one winner neurons can be identified from such a test.

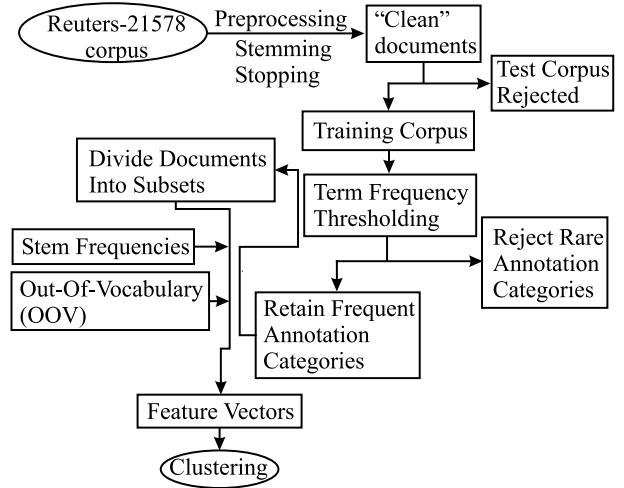


Figure 4. Block diagram from the unprocessed Reuters-21578 corpus through the formation of the feature vectors until the clustering algorithms that will partition the set of feature vectors into collection of “similar” vectors.

Prior to the construction of the feature vectors a series of actions were taken in order to remove any unwanted textual information from the documents. First, the SGML tags were removed. Subsequently, numbers and punctuation marks were also removed. Text cleaning was also applied in order to remove some common English words in a process called *stopping*. Subsequently, *stemming* was performed. Stemming refers to the elimination of the word suffixes so that the vocabulary shrinks, although it keeps the informative context of the text. For our purposes the commonly used Porter stemmer was applied [8].

Furthermore, the resulted corpus was partitioned into two distinct sets, a *training* and a *testing* set, according to the recommended *Modified Apte* split of the collection [7]. The first set, the training corpus, contains 8.762 documents whereas the second set, the testing set contains 3.009 documents.

Subsequently, from the documents that were retained we selected those belonging to the 25 most frequent categories. The documents in each category were randomly divided into non-overlapping subsets. Each of the subsets contained between 50 – 100 documents. The above step is justified by the fact that the Zipf distribution is applicable to large documents and the document subsets serves towards this direction. Through the above process 144 document subsets were formed. A detailed block diagram of the above procedure can be seen in Fig. 4.

Each one of the document subsets depicted in Fig. 4 and Fig. 5 corresponds to one feature vector. For the construction of these vectors both the global stem frequencies over the entire document subsets and the intra-subset stem frequencies are used. The stems that are common among all the subsets are the basis for the encoding of the subsets. The intra-subset frequencies for the common stems are used to form the feature vector depicted in Fig. 5 for that particular subset (*i.e.*, the first document set). The

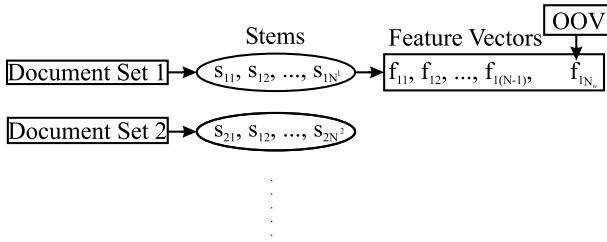


Figure 5. The formation of the feature vectors. The notion N^1, N^2, \dots correspond to the number of stems found in each document set. The notion N correspond to the cardinality of the set containing the stems that are common in all the document set plus one that correspond to the OOV term.

rest of the stems in each subsets that are not members of the intersection are grouped together to form an out-of-vocabulary (OOV) term for each one of the document subsets. This process can be seen in Fig. 5. The terms f_{11} up until $f_{1(N-1)}$ correspond to the stem frequencies of the first document set for the stems that are common in all the document subsets. The last term, f_{1N} correspond to the frequency of the OOV term found in that particular document set.

After the feature vector formation the components in each vector are rearranged in descending order according to the global frequencies. This step guarantees that each component (stem) resides in the same position inside all the feature vector. Furthermore, the Zipf distribution emerges due to the descending rearrangement of the global stem frequencies [2].

These vectors are presented an adequate number of times to each one of the NNs build using either the Euclidean or the Zipf metric in an effort to construct clusters containing semantically related document subsets. This process yields the so-called *document map* (DM) [9]. The DM corresponding to the Reuters-21578 corpus using the Zipf variant can be seen in Fig. 6. Each hexagon on the DM corresponds to one document category and the levels of grey correspond to different document densities. Hexagons with grey levels near 255 imply that fewer documents subsets have been assigned to these neurons and grey levels near 0 imply higher document subset densities.

To evaluate the performance of the algorithms, with respect to their DO capabilities, document-subset-queries are used. The documents used for queries are drawn from the test set of the Reuters-21578. Each query-subset undergoes the same steps like the training documents and is encoded by a feature vector. Following, the algorithms identify the winning neurons on the computed DMs and retrieve the set of document subsets of the training corpus associated with the winners. Subsequently, the retrieved documents are ranked according to their distance from the queries using either the Euclidean or the Zipf distance. Finally, the retrieved document subsets are labeled as either relevant or not to the document-subset-query, with respect to the annotation category they bear.

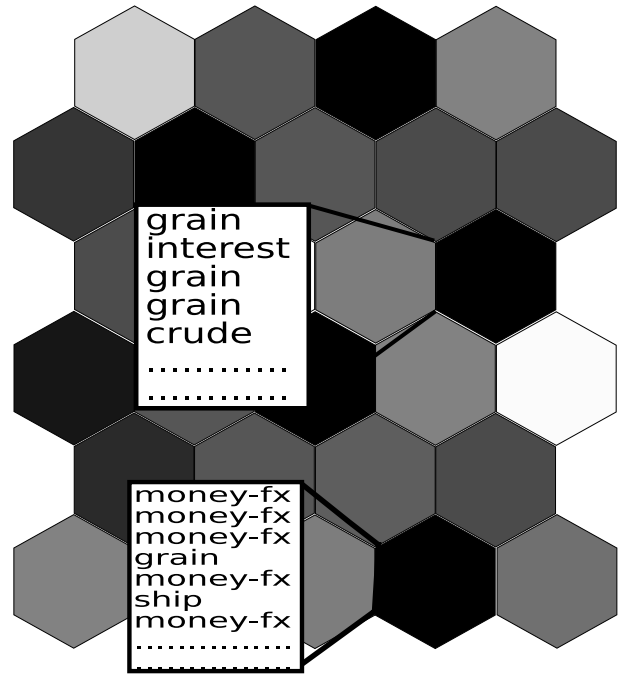


Figure 6. The document map constructed for the Reuters-21578 for a 4×6 neural network using the Zipf SOM variant. The highlighted neurons correspond to document clusters related to “gain”(top right and “financial” (bottom right).

The relevance between the retrieved document-subsets and the query-subsets leads to the partitioning of the training corpus into two sets, one containing the relevant document-subsets and another with the non-relevant document-subsets. The effectiveness of the algorithms is assessed using the *average recall-precision* curve [10]. Figure 7a depicts the average recall-precision curves for the standard SOM algorithm and the Zipf variant for the “Mergers and Acquisitions (ACQ)” category. The performance of the Zipf variant is higher to the standard SOM algorithm in small recall volumes by 2%, which is extremely important given the fact that an average user is interested in high precisions ratios even from the beginning of the list of returned relevant documents. Figure 7b depicts the average recall-precision curves for the standard SOM algorithm and the Zipf variant for the “Earnings and Earnings Forecasts (EARN)” category. At the beginning the performance of the SOM algorithm is slightly better than the proposed variant but it degrades rapidly as the volume of the retrieved document grows.

Figure 8 depicts the ANOVA analysis of the resulted recall-precision curves. The F -value for the ACQ curve is 7.42 which implies significant statistical difference between the proposed method and the standard algorithm whereas for the EARN curve the value is 1.22. In the later curve the statistical significance is inferior than the former case but still its noticeable.

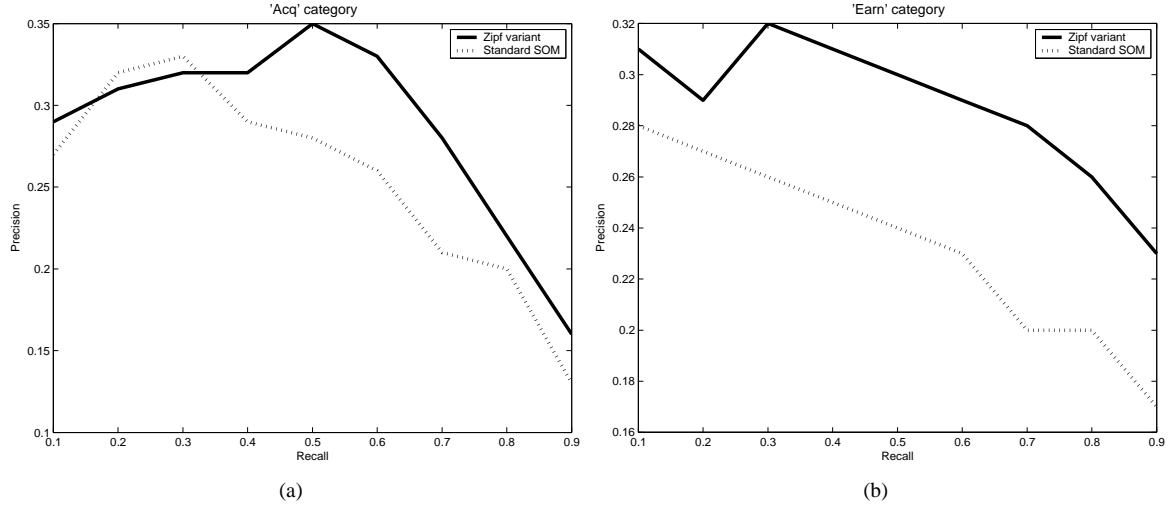


Figure 7. (a) The recall-precision curve for the standard SOM and the Zipf variant for the “Mergers and Acquisitions (ACQ)” category. (b) The recall-precision curves for both algorithms for the “Earnings and Earnings Forecasts (EARN)” category.

6. CONCLUSIONS

The present paper provides a mathematical analysis on a novel metric 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. The proposed metric was incorporated into the standard SOM algorithm. The recall and precision performance of the proposed variant is proven to be improved over the standard SOM. Further investigation will be performed towards the direction of the biasness of the introduced metric (investigate whether the proposed metric is biased or not).

7. APPENDIX A

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

Positiveness: Since $f_i(k)$ and $f_j(k)$ for $k = 1, 2, \dots, n$ contains the total probability mass of the i th and j th RV the following stems out:

$$\left. \begin{aligned}
 0 \leq x_{ik} < 1 \text{ and } \sum_{k=1}^n x_{ik} = 1 \\
 0 \leq x_{jk} < 1 \text{ and } \sum_{k=1}^n x_{jk} = 1
 \end{aligned} \right\} \Rightarrow$$

$$\begin{aligned}
 0 \leq x_{ik}x_{jk} &\leq 1 \Rightarrow \\
 0 \leq \sqrt{x_{ik}x_{jk}} &\leq 1 \Rightarrow \quad (23) \\
 0 \leq \sum_{k=1}^n \sqrt{x_{ik}x_{jk}} &\leq 1 \Rightarrow \\
 0 \leq 2L_{ij} &\leq 2 \Rightarrow \\
 2 \leq 2 + 2L_{ij} &\leq 4 \Rightarrow \\
 2 \leq D_{ij} &\leq 4
 \end{aligned}$$

In case where $i = j$ we have $L_{ii} = \sum_{k=1}^n \sqrt{x_{ik}x_{jk}} = \sum_{k=1}^n x_{ik} = 1 \Rightarrow D_{ii} = 2 + 2L_{ii} = 4.$

Symmetry: Since $x_{ik}x_{jk} = x_{jk}x_{ik} \Rightarrow D_{ij} = D_{ji}.$

Triangular inequality: In order to prove the triangular inequality we have to prove that:

$$\begin{aligned}
 D_{ij} + D_{jk} &\geq D_{ik} \Rightarrow \\
 2 + 2L_{ij} + 2 + L_{jk} &\geq 2 + L_{ik} \Rightarrow \\
 1 + L_{ij} + L_{jk} &\geq L_{ik} \quad (24)
 \end{aligned}$$

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

8. REFERENCES

- [1] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.
- [2] D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999.
- [3] “References on zipf’s law,” <http://linkage-rockefeller.edu/wli/zipf>.
- [4] L. A. Adamic, “Zipf, Power-laws, and Pareto - a ranking tutorial,” <http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html>.
- [5] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1984.
- [6] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Pubns., 10th edition, 1974.
- [7] D. D. Lewis, “Reuters-21578 text categorization test collection, distribution 1.0,” 1997, <http://kdd.ics.uci.edu/databases/reuters21578/-reuters21578.html>.

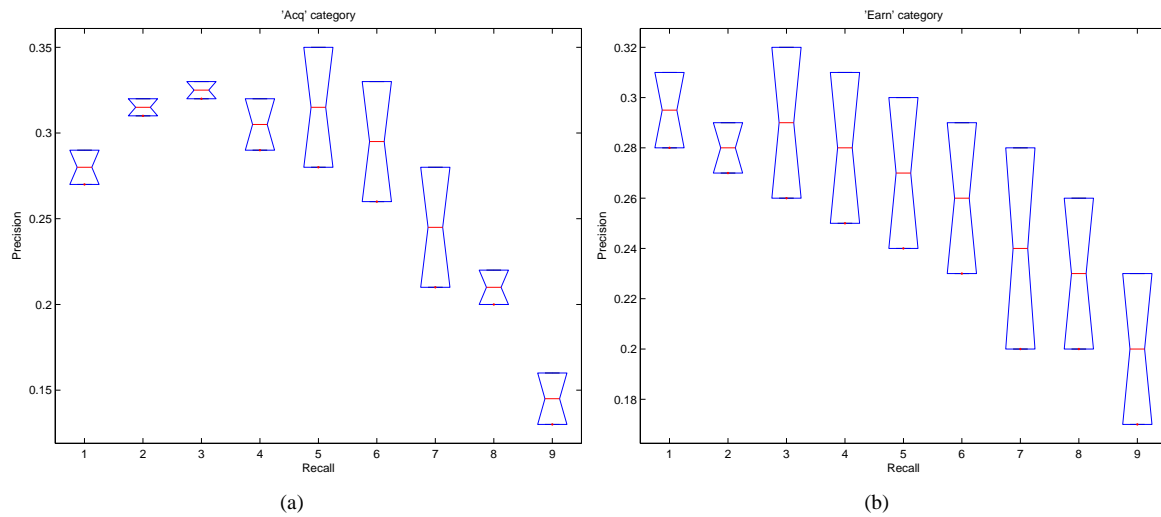


Figure 8. Analysis of variance for: (a) The recall-precision curve for the standard SOM and the Zipf variant for the “Mergers and Acquisitions (ACQ)” category and (b) The recall-precision curves for both algorithms for the “Earnings and Earnings Forecasts (EARN)” category.

- [8] M.F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [9] T. Kohonen, *Self Organizing Maps*, Springer-Verlag, 3rd edition, 2001.
- [10] R. R. Korfhage, *Information Storage and Retrieval*, New York: J. Wiley, 1997.